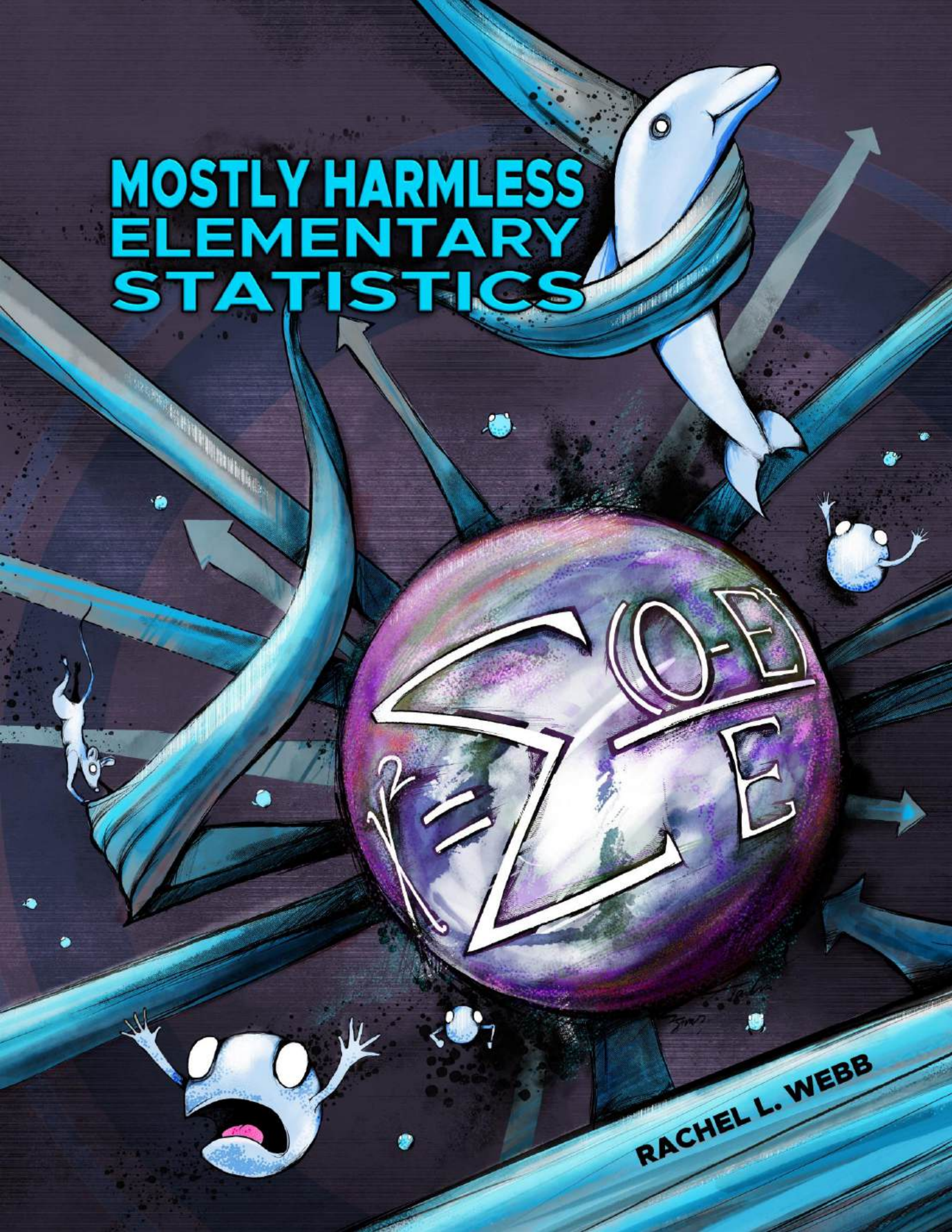


MOSTLY HARMLESS ELEMENTARY STATISTICS



RACHEL L. WEBB

Mostly Harmless Elementary Statistics

1st Edition

Rachel L. Webb
Portland State University

© 2023 Rachel L. Webb

Cover art by James Tadlock <https://redhedron.artstation.com/>. Cover art is not to be reused without artist's permission.



This textbook, except for the cover art, is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/) (CC-BY-SA).

Creative Commons Attribution Share Alike.

This license allows reuse, remixing, and distribution (including commercial), but requires any remixes use the same license as the original. This limits where the content can be remixed into, but on the other hand ensures that no one can remix the content then put the remix under a more restrictive license.

You are free to:

- Share - copy and redistribute the material in any medium or format
- Adapt - remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms. Under the following terms:

- Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Table of Contents

Preface and Acknowledgments	9
Chapter 1 Introduction to Data	10
1.1 Introduction.....	11
1.2 Samples vs. Populations.....	12
1.3 Types of Data.....	14
1.4 Sampling Techniques and Types of Studies	17
1.5 Ethics and Social Justice in Statistics	22
Chapter 1 Exercises.....	25
Chapter 2 Organizing Data	32
2.1 Introduction to Summarized Data.....	33
2.2 Tabular Displays	33
2.3 Graphical Displays.....	41
2.3.1 Stem-and-Leaf Plot	42
2.3.2 Histogram.....	44
2.3.3 Ogive.....	50
2.3.4 Pie Chart.....	53
2.3.5 Bar Graph.....	55
2.3.6 Pareto Chart	56
2.3.7 Stacked Column Chart	58
2.3.8 Multiple or Side-by-Side Bar Graph.....	62
2.3.9 Time-Series Plot.....	64
2.3.10 Scatter Plot.....	66
2.3.11 Misleading Graphs	68
Chapter 2 Exercises.....	74
Chapter 3 Descriptive Statistics	86
3.1 Measures of Center	87
3.1.1 Mode	87
3.1.2 Mean	88
3.1.3 Weighted Mean.....	89
3.1.4 Median	90
3.1.5 Outliers.....	91
3.1.6 Distribution Shapes.....	91
3.2 Measures of Spread.....	93

3.2.1 Range	94
3.2.2 Variance & Standard Deviation	94
3.2.3 Coefficient of Variation	99
3.3 Measures of Placement	100
3.3.1 Z-Scores	100
3.3.2 Percentiles	101
3.3.3 Quartiles.....	103
3.3.4 Five Number Summary & Outliers.....	105
3.3.5 Modified Box-and-Whisker Plot.....	105
3.4 Correlation and Linear Regression	108
3.4.1 Correlation	109
3.4.2 Linear Regression	115
Chapter 3 Formulas.....	121
Chapter 3 Exercises.....	122
Chapter 4 Probability	134
4.1 Introduction to Probability.....	135
4.2 Complement Rule	140
4.3 Union & Intersection.....	142
4.4 Independent Events and Conditional Probability	148
4.5 Counting Rules.....	152
Chapter 4 Formulas.....	156
Chapter 4 Exercises.....	157
Chapter 5 Probability Distributions	162
5.1 Introduction to Probability Distributions	163
5.2 Discrete Probability Distributions.....	163
5.2.1 Mean of a Discrete Probability Distribution.....	168
5.2.2 Variance & Standard Deviation of Discrete Probability Distributions.....	171
5.3 Binomial Distribution	173
5.4 Empirical Rule	179
5.5 Normal Distribution	182
5.5.1 Standard Normal Distribution.....	183
5.5.2 Applications of the Normal Distribution	187
5.5.3 Normal Probability Plot	191
5.6 The Central Limit Theorem	192

5.7 Student's T-Distribution	198
Chapter 5 Formulas.....	201
Chapter 5 Exercises.....	202
Chapter 6 Confidence Intervals for One Population.....	211
6.1 Introduction to Interval Estimates.....	212
6.2 Confidence Interval for a Proportion	213
6.3 Confidence Interval for a Mean	217
6.4 Interpreting a Confidence Interval	221
6.5 Determining Sample Size	224
Chapter 6 Formulas.....	227
Chapter 6 Exercises.....	228
Chapter 7 Hypothesis Tests for One Population.....	233
7.1 Introduction to Hypothesis Testing.....	234
7.2 Type I and II Errors.....	242
7.3 Hypothesis Test for One Proportion	246
7.4 Hypothesis Test for One Mean	250
Chapter 7 Formulas.....	262
Chapter 7 Exercises.....	263
Chapter 8 Hypothesis Tests & Confidence Intervals for Two Populations.....	271
8.1 Introduction to Inference For Two Populations.....	272
8.2 Two Proportion Hypothesis Test & Confidence Interval	272
8.3 Two Dependent Means Hypothesis Test & Confidence Interval	277
8.4 Two Independent Means Hypothesis Test & Confidence Interval.....	283
8.4.1 Unequal Variance Method	284
8.4.2 Equal Variance Method	290
Chapter 8 Formulas.....	297
Chapter 8 Exercises.....	299
Chapter 9 Chi-Square Tests.....	309
9.1 Introduction to the Chi-Square Distribution	310
9.2 Goodness of Fit Test	312
9.3 Test for Independence.....	315
Chapter 9 Formulas.....	321
Chapter 9 Exercises.....	322
Chapter 10 Analysis of Variance	329

10.1 Introduction to the F-Distribution.....	330
10.2 One-Way ANOVA.....	332
10.3 Pairwise Comparisons of Means (Post-Hoc Tests).....	339
Chapter 10 Formulas.....	345
Chapter 10 Exercises.....	346
Chapter 11 Linear Regression Analysis.....	353
11.1 Correlation	354
11.2 Hypothesis Test for a Correlation	356
11.3 Hypothesis Testing for Linear Regression.....	359
11.4 Coefficient of Determination	366
11.5 Residual Analysis.....	368
11.6 Outliers.....	374
11.7 Prediction Interval.....	376
11.8 Linear Regression Analysis	378
Chapter 11 Formulas.....	382
Chapter 11 Exercises.....	383
Answers to Odd Numbered Exercises.....	394
Data Sources and References	404
Glossary	406
Symbols	410
Greek Letter Pronunciations	411
Mostly Harmless Elementary Statistics Formula Packet.....	412

Preface and Acknowledgments

Mostly Harmless Elementary Statistics

This text is for an introductory level probability and statistics course with an intermediate algebra prerequisite. The focus of the text follows the American Statistical Association's Guidelines for Assessment and Instruction in Statistics Education ([GAISE](#)). Software examples provided for Microsoft Excel, TI-84 & TI-89 calculators.

Students new to statistics are sure to benefit from this ADA accessible and relevant textbook. The examples are current and resonate with everyday life. The casual narrative style, has a conversational tone to provide an inclusive and easy to read format for students.

This text is a pared down version of the Mostly Harmless Statistics, 2nd Edition to accommodate the 2023 Oregon Higher Education Coordinating Committee's (HECC) transfer policy for Oregon colleges and universities. This edition has over 120 new homework questions, plus updated content. This textbook, except for the cover art, is licensed under a Creative Commons Attribution-Share Alike 4.0 International License ([CC-BY-SA 4.0](#)). Cover art copyright by James Tadlock <https://www.artstation.com/redhedron>.

I am filled with profound gratitude as I extend my deepest appreciation to my life partner, Matthew Mendoza for his unwavering support and encouragement throughout the process of writing this textbook. His love, patience, and understanding have been invaluable in helping me achieve my goal of sharing my knowledge and expertise with others. His belief in me and my abilities has been a constant source of motivation, and I am truly grateful for his unwavering commitment to our shared goals.

Thank you, Gene Enneking, for hiring me and giving me a career that I love. Thank you to Paul Latiolais, my first mathematics teacher to make the subject exciting enough to switch my major to mathematics and statistics. Thank you to Ashley Lowe for the title idea and our admiration for Douglas Adams. Thank you for the edits and content help from Whitney Cave, Naghmeh Daneshi, and Jennifer Ward. A huge thank you to all my students that have given feedback and found typos, let me know if you find any more.

For students, you can download a student solution manual with solutions to all the odd exercises at <http://MostlyHarmlessStatistics.com/>. For instructors, there is a complete solution manual and a problem bank in [MyOpenMath](#) divided by chapter. Email Rachel Webb for more information.

“Remember not to panic and don't forget your towel!”
(Adams, 2002)

Rachel L. Webb, M.S. Statistics
Portland State University
mostlyharmlessstatistics@gmail.com
<http://MostlyHarmlessStatistics.com/>
<https://www.facebook.com/MostlyHarmlessStatistics>

“Ford didn't comment. He was listening to something. He passed the Guide over to Arthur and pointed at the screen. The active entry read ‘Earth. Mostly harmless.’”

...Earth, a world whose entire entry in the Hitchhiker's Guide to the Galaxy comprised the two words ‘Mostly harmless.’”
(Adams, 2002)

Chapter 1

Introduction to Data



- 1.1 Introduction
- 1.2 Samples vs. Populations
- 1.3 Types of Data
- 1.4 Sampling Techniques & Types of Studies
- 1.5 Ethics and Social Justice in Statistics

1.1 Introduction

Welcome to this introductory course on probability and statistics. Probability and statistics are essential topics in the world of data science, providing a foundation for understanding and making sense of data. Whether you are an aspiring data analyst, data scientist, or simply interested in learning about the basics of probability and statistics, this course is for you.

Throughout this course, you will learn about the fundamentals of probability theory and its applications, including the concept of randomness, probability distributions, and the laws of probability. You will also delve into the world of statistics, learning about confidence intervals, hypothesis testing, correlation and linear regression. In addition, you will gain hands-on experience working with data and using statistical software to analyze and interpret data.

The goal of this course is not only to provide you with the necessary mathematical tools for analyzing data but also to develop your critical thinking skills in approaching real-world problems. Probability and statistics play a crucial role in many fields, from healthcare and finance to engineering and social sciences. By understanding the fundamentals of probability and statistics, you will be better equipped to make informed decisions and contribute to the advancement of your chosen field.

I hope you find this course rewarding, and I wish you the best of luck on your journey into the world of probability and statistics.

What are Statistics?

Scientists seek to answer questions using rigorous methods and careful observations. These observations - collected from the likes of field notes, surveys, and experiments - form the backbone of a statistical investigation and are called data. Statistics is the study of how best to collect, analyze, and draw conclusions from data. It is helpful to put statistics in the context of a general process of investigation:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Statistics as a subject focuses on making stages 2-4 objective, rigorous, and efficient. That is, statistics has three primary components: How best can we collect data? How should the data be analyzed? What can we infer from the analysis?

The topics scientists investigate are as diverse as the questions they ask. However, many of these investigations can be addressed with a small number of data collection techniques, analytic tools, and fundamental concepts in statistical inference.

You are exposed to statistics regularly. If you are a sports fan, then you have the statistics for your favorite player. If you are interested in politics, then you look at the polls to see how people feel about certain issues or candidates. If you are an environmentalist, then you research arsenic levels in the water of a town or analyze the global temperatures. If you are in the business profession, then you may track the monthly sales of a store or use quality control processes to monitor the number of defective parts manufactured. If you are in the health profession, then you may look at how successful a procedure is or the percentage of people infected with a disease. There are many other examples from other areas.

“There are of course many problems connected with life, of which some of the most popular are: Why are people born? Why do they die? Why do they want to spend so much time wearing digital watches?”
(Adams, 2002)

To understand how to collect and analyze data, you need to understand what the field of statistics is. Many of the words defined throughout this course have common definitions that are also used in non-statistical terminology. In

statistics, some of these terms have slightly different definitions. It is important that you notice the difference and utilize the statistical definitions.

Statistics is the study of how to collect, organize, analyze, and interpret data collected from a group.

There are two main branches of statistics. One is called **descriptive statistics**, which is where you collect, organize and describe data. The other branch is called **inferential statistics**, which is where you interpret data. First, you need to look at descriptive statistics since you will use the descriptive statistics when making inferences. In order to use inferential statistics, we will briefly touch on a completely new topic called **Probability**. Once we get some background in probability combined with your knowledge of descriptive statistics, we will move into Inferential Statistics.

1.2 Samples vs. Populations

The first thing to decide in a statistical study is whom you want to measure and what you want to measure. You always want to make sure that you can answer the question of whom you measured and what you measured. The “who” is known as the individual and the “what” is known as the variable.

Individual – a person, case or object that you are interested in finding out information about.

Variable (also known as a random variable) – the measurement or observation of the individual.

Population – is the total set of all the observations that are the subject of a study.

Notice, the population answers “who” you want to measure and the variable answers “what” you want to measure. Make sure that you always answer both of these questions or you have not given the audience reading your study the entire picture. As an example, if you just say that you are going to collect data from the senators in the United States Congress, you have not told your reader what you are going to collect. Do you want to know their income, their highest degree earned, their voting record, their age, their political party, their gender, their marital status, or how they feel about a particular issue? Without telling “what” you want to measure, your reader has no idea what your study is actually about.

Sometimes the population is very easy to collect. If you are interested in finding the average age of all of the current senators in the United States Congress, there are only 100 senators. This would not be hard to find. However, if instead you were interested in knowing the average age that a senator in the United States Congress first took office for all senators that ever served in the United States Congress, then this would be a bit more work. It is still doable, but it would take a bit of time to collect. However, what if you are interested in finding the average diameter at breast height of all Ponderosa Pine trees in the Coconino National Forest? This data would be impossible to collect. What do you do in these cases? Instead of collecting the entire population, you take a smaller group of the population, a snapshot of the population. This smaller group, called a sample, is a subset of the population, see Figure 1-1.

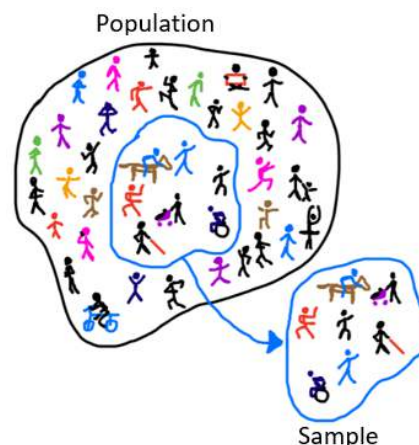


Figure 1-1

Sample – a subset from the population.

Consider the following three research questions:

1. What is the average mercury content in albacore tuna in the Pacific Ocean?
2. Over the last 5 years, what is the average time to complete a degree for Portland State University undergraduate students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target population. In the first question, the target population is all albacore tuna in the Pacific Ocean, and each fish represents a case.

A sample represents a subset of the cases and is often a small fraction of the population. For instance, 60 albacore tunas in the population might be selected and the mercury level is measured in each fish. The sample average of the 60 fish may then be used to provide an estimate of the population average of all the fish and answer the research question.

The last ever dolphin message was misinterpreted as a surprisingly sophisticated attempt to do a double-backwards – somersault through a hoop whilst whistling the "Star Sprangled Banner," but in fact the message was this: So long and thanks for all the fish.
(Adams, 2002)

We use the lower-case n to represent the number of cases in the sample and the upper-case N to represent the number of cases in the population.

n = sample size.
 N = population size.

How the sample is collected can determine the accuracy of the results of your study. There are many ways to collect samples. No sampling method is perfect, but some methods are better than other methods. Sampling techniques will be discussed in more detail later.

For now, realize that every time you take a sample you will find different data values. The sample is a snapshot of the population, and there is more information than is in this small picture. The idea is to try to collect a sample that gives you an accurate picture, but you will never know for sure if your picture is the correct picture. Unlike previous mathematics classes, where there was always one right answer, in statistics there can be many answers, and you do not know which are right.

The sample average in this case is the statistic, and the population average is the parameter. We use sample statistics to make inferences, educated guesses made by observation, about the population parameter.

Once you have your data, either from a population or from a sample, you need to know how you want to summarize the data.

As an example, suppose you are interested in finding the proportion of people who like a candidate, the average height a plant grows to using a new fertilizer, or the variability of the test scores. Understanding how you want to summarize the data helps to determine the type of data you want to collect. Since the population is what we are interested in, then you want to calculate a number from the population. This is known as a parameter.

Parameter – An unknown quantity from the population. Usually denoted with a Greek letter (for example μ “mu”). This number is a fixed, unknown number that we want to estimate.

As mentioned already, it is hard to collect the entire population. Even though this is the number you are interested in, you cannot really calculate it. Instead, you use the number calculated from the sample, called a statistic, to estimate the parameter.

Statistic – a number calculated from the sample. Usually denoted with a $\hat{}$ (called a hat, for example \hat{p} “p-hat”) or a $\bar{}$ (called a bar, for example \bar{x} “x-bar”) above the letter.

Since most samples are not exactly the same, the statistic values are going to be different from sample to sample. Statistics estimate the value of the parameter, but again, you do not know for sure if your statistic is correctly estimating the parameter.

1.3 Types of Data

There are various types of variables and data types that play a fundamental role in understanding and analyzing data. These concepts provide a framework for organizing and interpreting information, allowing us to draw meaningful conclusions and make informed decisions based on statistical analysis.

Qualitative vs. Quantitative

One important distinction in statistics is between categorical variables and numerical variables. Categorical variables, also known as **qualitative** variables, represent characteristics or attributes to describe a quality of an individual that can be sorted into distinct categories or groups. On the other hand, numerical variables, also called **quantitative** variables, represent measurable quantities or numerical values that count or measure an individual. Qualitative variables may sometimes have numeric responses that represent a category or word.

Qualitative or categorical variable – answer is a word or name that describes a quality of the individual.

Quantitative or numerical variable – answer is a number (quantity), something that can be counted or measured from the individual.

Each type of variable has different graphs, parameters and statistics that you find. Quantitative variables usually have a number line associated with graphical displays. Qualitative variables usually have a category name associated with graphical displays.

Examples of quantitative variables are number of people per household, age, height, weight, time (usually things we can count or measure). Examples of qualitative variables are eye color, gender, sports team, yes/no (usually things that we can name).

When setting up survey questions it is important to know what statistical questions you would like the data to answer. For example, a company is trying to target the best age group to market a new game. They put out a survey with the ordinal age groupings: baby, toddler, adolescent, teenager, adult, and elderly. We could narrow down a range of ages for, say, teenagers to 13-19, although many 19-year-olds may record their response as an adult. The company wants to run an ad for the new game on television and they realize that 13-year-olds do not watch the same shows nor in the same time slots as 19-year-olds. To narrow down the age range the survey question could have just asked the person’s age. Then the company could look at a graph or average to decide more specifically that 17-year-olds would be the best target audience.

Types of Measurement Scales

There are four types of data measurement scales: nominal, ordinal, interval and ratio.

Ordinal data and nominal data are two types of categorical or qualitative data.

Nominal data are data that have no inherent order or ranking. Some examples of nominal data include:

- Eye color (blue, brown, green, etc.)
- Gender (female, male, non-binary, etc.)
- Types of fruit (apple, banana, orange, etc.)

Ordinal data are data that can be ordered or ranked, but the distance between the categories is not necessarily equal. In other words, the difference between the categories is not meaningful, but the order or ranking of the categories is. Some examples of ordinal data include:

- Letter grades (A, B, C, etc.)
- Levels of education (elementary, middle school, high school, etc.)
- Customer satisfaction levels (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied)

In summary, the key difference between ordinal data and nominal data is that ordinal data can be ranked or ordered, while nominal data cannot.

Interval data and ratio data are two types of numeric or quantitative data.

Interval data are numeric values where the difference between any two values is meaningful and can be compared, but there is no true zero point. This means that you can perform operations like addition and subtraction, but you cannot meaningfully perform multiplication or division. Some examples of interval data include:

- Temperature measured in Celsius or Fahrenheit
- Time of day measured in 24-hour clock
- IQ scores

Ratio data, on the other hand, are numeric values where there is a true zero point, and the ratio of two values is meaningful. This means that you can perform all mathematical operations on these values. Some examples of ratio data include:

- Height or weight
- Length or distance
- Age
- Income
- Number of siblings

In summary, the key difference between interval data and ratio data is the presence or absence of a true zero point. Ratio data have a true zero point, while interval data do not.

Nominal and ordinal data are qualitative, while interval and ratio data are quantitative.

A **Likert scale** is a numeric scale that indicates the extent to which they agree or disagree with a series of statements. Likert scales are frequently misused as quantitative data. For example, the following is a 5-point Likert Scale:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Likert scales are ordinal in that one can easily see that the larger number corresponds to a higher level of agreeableness. Some people argue that since there is a one-unit difference between the numeric values Likert scales should be interval data. However, the number 1 is just a placeholder for someone that strongly disagrees. There is no way to quantify a one-unit difference between two different subjects that answered 1 or 2 on the scale. For example, one person's response for strongly disagree could stem from the exact same reasoning behind another person's response of disagree. People view subjects at different intensities that is not quantifiable.

Nominal data is categorical data that has no order or rank, for example the color of your car, ethnicity, race, or gender.

Ordinal data is categorical data that has a natural order to it, for example, year in school (freshman, sophomore, junior, senior), a letter grade (A, B, C, D, F), the size of a soft drink (small, medium, large) or Likert scales.

Interval data is numeric where there is a known difference between values, but zero does not mean “nothing.” Interval data is ordinal, but you can now subtract one value from another and that subtraction makes sense. You can do arithmetic on this data. For example, Fahrenheit temperature, 0° is cold but it does not mean that no temperature exists. Time, dates and IQ scores are other examples.

Ratio data is numeric data that has a true zero, meaning when the variable is zero nothing is there. Most measurement data are ratio data. Some examples are height, weight, age, distance, or time running a race.

Here are some ways to help you decide if the data are nominal, ordinal, interval, or ratio. First, if the variable is words instead of numbers, then it is either nominal or ordinal data. Now ask yourself if you can put the data in a particular order. If you can order the names then this is ordinal data. Otherwise, it is nominal data. If the variable is numbers (not including words coded as numbers like Yes = 1 and No = 0), then it is either interval or ratio data. For ratio data, a value of zero means there is no measurement. This is known as the absolute zero. If there is an absolute zero in the data, then it means it is ratio. If there is no absolute zero, then the data are interval. An example of an absolute zero is if you have \$0 in your bank account, then you are without money. The amount of money in your bank account is ratio data. Word of caution, sometimes ordinal data is displayed using numbers, such as 5 being strongly agree, and 1 being strongly disagree. These numbers are not really numbers. Instead, they are used to assign numerical values to ordinal data. In reality, you should not perform any computations on this data, though many people do. If there are numbers, make sure the numbers are inherent numbers, and not numbers that were randomly assigned.

Discrete vs. Continuous

Quantitative variables can be either discrete or continuous. This difference will be important later on when we are working with probability. Discrete variables have gaps between points that are countable, usually integers like the number of cars in a parking garage or how many people per household. A continuous variable can take on any value and is measurable, like height, time running a race, distance between two buildings. Usually, just asking yourself if you can count the variable then it is discrete and if you can measure the variable then it is continuous. If you can actually count the number of outcomes (even if you are counting to infinity), then the variable is discrete.

Discrete variables can only take on particular values like integers.
Discrete variables have outcomes you can count.

Continuous variables can take on any value.
Continuous variables have outcomes you can measure.

For example, think of someone’s age. They may report in a survey an integer value like 28 years-old. The person is not exactly 28 years-old though. From the time of their birth to the point in time that the survey respondent recorded, their age is a measurable number in some unit of time. A person’s true age has a decimal place that can keep going as far as the best clock can measure time. It is more convenient to round our age to an integer rather than 28 years 5 months, 8 days, 14 hours, 12 minutes, 27 seconds, 5 milliseconds or as a decimal 28.440206335775. Therefore, age is continuous.

When a survey question takes a continuous variable and chunks it into discrete categories, especially categories with different widths, you limit what type of statistics you can do on that data.

A continuous variable like age could be broken into discrete bins or categories. For example, instead of the question asking for a numeric response for a person's age they could have had discrete age ranges where the survey respondent just selects an age category instead of reporting their numeric age.

1. Under 18
2. 18-24
3. 25-35
4. 36-45
5. 46-62
6. Over 62

When age groups are used to summarize data, it can pose challenges in accurately calculating statistics for the sample. Therefore, it is crucial to consider the specific statistical requirements when formulating survey questions.

Figure 1-2 is a breakdown of the different variable and data types.

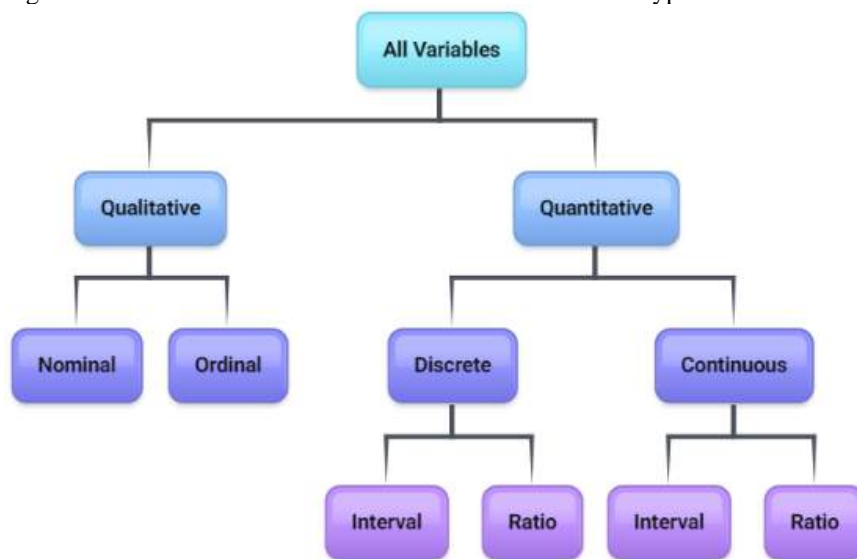


Figure 1-2

Understanding the different types of variables and data types is crucial for selecting appropriate statistical techniques and interpreting the results correctly. It provides a foundation for exploring and analyzing data in a meaningful way, enabling us to draw valid conclusions and insights from statistical analyses.

1.4 Sampling Techniques and Types of Studies

If you want to know something about a population, it is often impossible or impractical to examine the entire population. It might be too expensive in terms of time or money to survey the population. It might be impractical: you cannot test all batteries for their length of lifetime because there would not be any batteries left to sell.

When you choose a sample, you want it to be as similar to the population as possible. If you want to test a new painkiller for adults, you would want the sample to include people of different weights, age, etc. so that the sample would represent all the demographics of the population that would potentially take the painkiller. The more similar the sample is to the population, the better our statistical estimates will be in predicting the population parameters.

There are many ways to collect a sample. No sampling technique is perfect, and there is no guarantee that you will collect a representative sample. That is unfortunately the limitation of sampling. However, several techniques can result in samples that give you a semi-accurate picture of the population. Just remember to be aware that the sample

may not be representative of the whole population. As an example, you can take a random sample of a group of people that are equally distributed across all income groups, yet by chance, everyone you choose is only in the high-income group. If this happens, it may be a good idea to collect a new sample if you have the time and money.

When setting up a study there are different ways to sample the population of interest. The five main sampling techniques are:

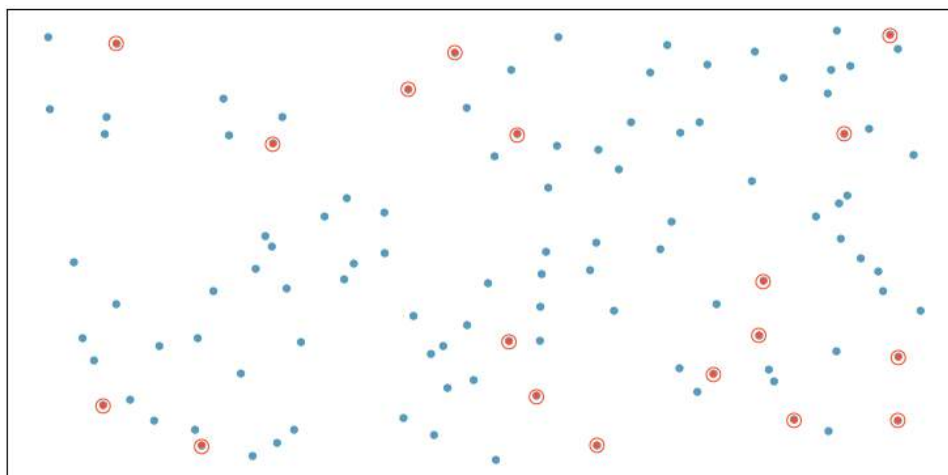
1. Simple Random Sample
2. Systematic Sample
3. Stratified Sample
4. Cluster Sample
5. Convenience Sample

A **simple random sample** (SRS) means selecting a sample size of n objects from the population so that every sample of the same size n has equal probability of being selected as every other possible sample of the same size from that population.

A random sample is a subset of a population chosen randomly so that each individual in the population has an equal chance of being selected. This means that any member of the population can be chosen, but the process of selection may not be entirely random. On the other hand, a simple random sample is a specific type of random sample that is chosen strictly randomly, without any bias or systematic approach. In a simple random sample, each individual in the population has an equal chance of being selected, and every possible combination of individuals has an equal chance of being selected as well.

In other words, a simple random sample is a subset of a population chosen by a random process where every member of the population has an equal chance of being selected and every subset of a given size has an equal chance of being selected. Simple random samples are often used in statistical studies because they minimize bias and provide a representative sample of the population.

For example, we have a database of all Portland State University (PSU) student data and we use a random number generator to randomly select 18 students to receive a questionnaire on the type of transportation they use to get to school. See Figure 1-3. Simple random sampling was used to randomly select the 18 cases.

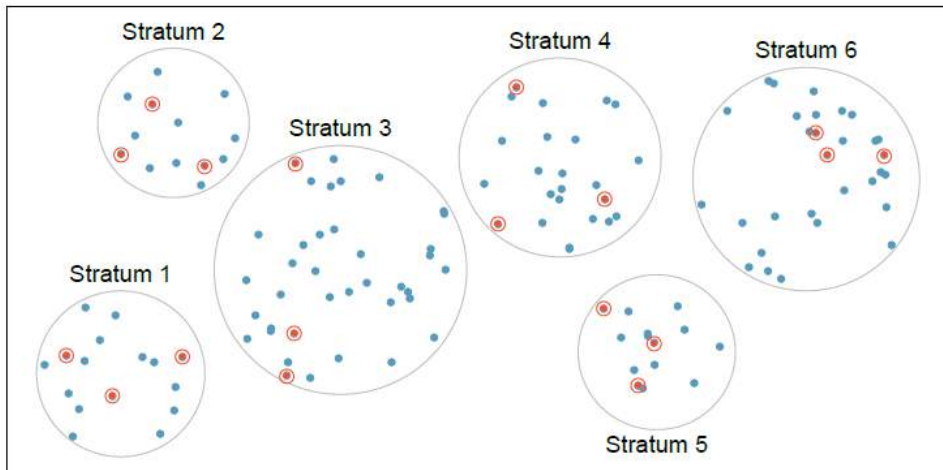


Retrieved from [OpenIntroStatistics](https://openintrostatistics.com/).

Figure 1-3

A **stratified sample** is where the population is split into groups called strata, then a random sample is taken from each stratum. For instance, we divide Portland by Zone Improvement Plan (ZIP) code and then randomly select n registered

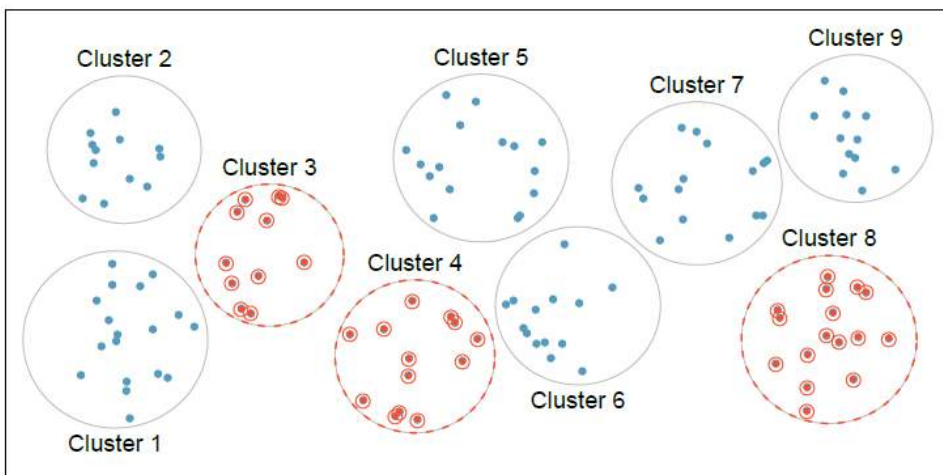
voters out of each ZIP code. See Figure 1-4. Cases were grouped into strata, then simple random sampling was employed within each stratum.



Retrieved from [OpenIntroStatistics](https://openintrostatistics.com).

Figure 1-4

A **cluster sample** is where the population is split up into groups called clusters, then one or more clusters are randomly selected and all individuals in the chosen clusters are sampled. Similar to the previous example, we split Portland up by ZIP code, randomly pick 5 ZIP codes and then sample every registered voter in those 5 ZIP codes. See Figure 1-5. Data were binned into nine clusters, three of these clusters were sampled, and all observations within these three clusters were included in the sample.



Retrieved from [OpenIntroStatistics](https://openintrostatistics.com).

Figure 1-5

Both stratified sampling and cluster sampling methods are easy to get mixed up. Stratified sampling takes a random sample from each stratum. Cluster sampling, on the other hand, involves dividing the population into clusters and then randomly selects a few of these clusters and collects data from all members of the chosen clusters.

A **systematic sample** is where we list the entire population, then randomly pick a starting point at the n^{th} object, and then take every n^{th} value until the sample size is reached. For example, we alphabetize every PSU student, randomly choose the number 7. We would sample the 7th, 14th, 21st, 28th, 35th, etc. student.

A **convenience sample** is picking a sample that is conveniently at hand. For example, asking other students in your statistics course or using social media to take your survey. Most convenience samples will give biased views and are not encouraged.

There are many more types of sampling, snowball, multistage, voluntary, purposive, and quota sampling to name some of the ways to sample from a population. We can also combine the different sampling methods. For example, we could stratify by rural, suburban and urban school districts, then take 3rd grade classrooms as clusters.

Guidelines for planning a statistical study

- Identify the individuals that you are interested in studying. Realize that you can only make conclusions for these individuals. As an example, if you use a fertilizer on a certain genus of plant, you cannot say how the fertilizer will work on any other types of plants. However, if you diversify too much, then you may not be able to tell if there really is an improvement since you have too many factors to consider.
- Specify the variable. You want to make sure the variable is something that you can measure, and make sure that you control for all other factors too. For example, if you are trying to determine if a fertilizer works by measuring the height of the plants on a particular day, you need to make sure you can control how much fertilizer you put on the plants (which is what we call a treatment), and make sure that all the plants receive the same amount of sunlight, water, and temperature.
- Specify the population. This is important in order for you to know for whom and what conclusions you can make.
- Specify the method for taking measurements or making observations.
- Determine if you are taking a census or sample. If taking a sample, decide on the sampling method.
- Collect the data.
- Use appropriate descriptive statistics methods and make decisions using appropriate inferential statistics methods.
- Note any concerns you might have about your data collection methods and list any recommendations for future.

Observational vs. Experimental Studies

The section is an introduction to experimental design. This is a brief introduction on how to design an experiment or a survey so that they are statistically sound. Experimental design is a very involved process, so this is just a small overview.

There are two types of studies:

1. An **observational study** is when the investigator collects data by observing, measuring, counting, watching or asking questions. The investigator does not change anything.
2. An **experiment** is when the investigator changes a variable or imposes a treatment to determine its effect.

For instance, if you were to poll students to see if they favor increasing tuition, this would be an observational study since you are asking a question and getting data. Give a patient a medication that lowers their blood pressure. This is an experiment since you are giving the treatment and then getting the data.

Many observational studies involve surveys. A **survey** uses questions to collect the data and needs to be written so that there is no bias. Bias is the tendency of a statistic to incorrectly estimate a parameter. There are many ways bias can seep into statistics. Sometimes we don't ask the correct question, give enough options for answers, survey the wrong people, misinterpret data, sampling or measurement errors, or unrepresentative samples.

There are different time-periods for data collection to consider for observational studies:

- **Cross-sectional study:** observational data collected at a single point in time.
- **Retrospective study:** observational data collected from the past using records, interviews, and other similar artifacts.
- **Prospective (or longitudinal or cohort) study:** Subjects are measured from a starting point over time for the occurrence of the condition of interest.

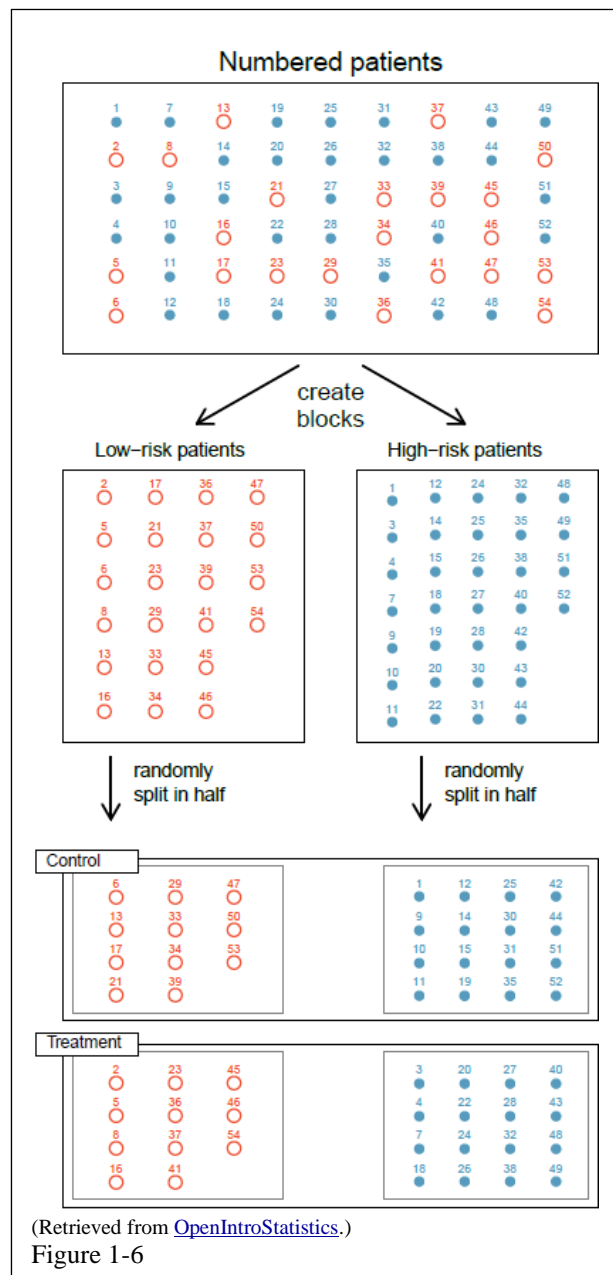
In an experiment, there are different options to assign treatments.

- **Completely Randomized Experiment:** In this experiment, the individuals are randomly placed into two or more groups. One group gets either no treatment or a placebo (a fake treatment); this group is called the control group. The groups getting the treatment are called the treatment groups. The idea of the placebo is that a person thinks they are receiving a treatment, but in reality, they are receiving a sugar pill or fake treatment. Doing this helps to account for the placebo effect, which is where a person's mind makes their body respond to a treatment because they think they are taking the treatment when they are not really taking the treatment. Note, not every experiment needs a placebo, such as when using animals or plants. In addition, you cannot always use a placebo or no treatment. For example, if you are testing a new Ebola vaccination you cannot give a person with the disease a placebo or no treatment because of ethical reasons.

- **Randomized Block Design:** A block is a group of subjects that are considered similar or the same subject measured multiple times, but the blocks differ from each other. Then randomly assign treatments to subjects inside each block. For instance, a company has several new stitching methods for a soccer ball and would like to pick the ball that travels the fastest. We would expect variation in different soccer player's abilities which we do not want affect our results. We randomly choose players to kick each of the new types of balls where the order of the ball design is also randomized.

Figure 1-6 shows blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

- **Matched Pairs Design:** This is a subset of the randomized block design where the treatments are given to two groups that can be matched up with each



other in some way. One example would be to measure the effectiveness of a muscle relaxer cream on the right arm and the left arm of individuals, and then for each individual you can match up their right arm measurement with their left arm. Another example of this would be before and after experiments, such as weight of a person before and weight after a diet.

- **Factorial Design:** This design has two or more independent categorical variables called factors. Each factor has two or more different treatment levels. The factorial design allows the researcher to test the effect of the different factors simultaneously on the dependent variable. For example, an educator believes that both the time of day (morning, afternoon, evening) and the way an exam is delivered (multiple-choice paper, short answer paper, multiple-choice electronic, short answer electronic) affects a student's grade on their exam.

No matter which experiment type you conduct, you should also consider the following:

Replication: repetition of an experiment on more than one subject so you can make sure that the sample is large enough to distinguish true effects from random effects. It is also the ability for someone else to duplicate the results of the experiment.

Blind study is where the individual does not know which treatment they are getting or if they are getting the treatment or a placebo.

Double-blind study is where neither the individual nor the researcher knows who is getting the treatment and who is getting the placebo. This is important so that there can be no bias in the results created by either the individual or the researcher.

1.5 Ethics and Social Justice in Statistics

Statistical ethics involves ensuring the integrity, accuracy, and transparency of data analysis, as well as addressing ethical considerations related to data collection, privacy, confidentiality, and the appropriate use and interpretation of statistical results. Ethics play a critical role in statistical analysis. As statistics become increasingly important in decision-making, it is essential that researchers and analysts understand the ethical considerations that must be taken into account when conducting statistical analysis.

One of the primary reasons ethics are important for statistical analysis is that statistical results can have significant consequences. Statistics are often used to inform decisions in fields such as medicine, finance, and public policy. If the statistical analysis is conducted unethically or with bias, the results can have harmful consequences. For example, an unethical study that falsely concludes that a certain drug is safe and effective could lead to serious harm to patients who rely on that drug. Therefore, it is crucial that statistical analysis is conducted ethically to ensure that the results are reliable and trustworthy. Fraudulent statistics are difficult to identify. The American Statistical Association, clearly define expectations for researchers.

Another reason ethics are important for statistical analysis is that statistical methods can be easily misinterpreted or manipulated. It is essential that researchers and analysts are transparent about their methods and findings to ensure that others can verify the results. This requires ethical behavior such as being honest about the limitations of the data and acknowledging potential sources of bias. Additionally, statistical analysis often involves making assumptions about the data, and it is crucial that these assumptions are based on sound ethical principles to ensure that the results are not biased or misleading.

Furthermore, ethical considerations are particularly important when it comes to data privacy and confidentiality. Statistical analysis often involves working with sensitive data, such as personal health information or financial data. Analysts must take steps to ensure that the data is protected and that the privacy of individuals is not compromised. This may involve obtaining informed consent from study participants or using techniques such as de-identification to protect personal information. Many institutions establish oversight committees known as Institutional Review Boards (IRB) that pre-approve human subject studies.

The use of research data in the United States is regulated by a variety of laws, including federal statutes and regulations. The following is a brief overview of some of the key federal laws that govern the use of research data:

- **The Common Rule:** The Common Rule is a federal regulation that applies to research involving human subjects that is conducted or supported by federal agencies. The Common Rule establishes ethical principles and regulatory requirements for the protection of human subjects in research, including requirements for informed consent, IRB review, and data security.
- **HIPAA:** The Health Insurance Portability and Accountability Act (HIPAA) is a federal law that regulates the use and disclosure of individually identifiable health information. HIPAA applies to research that involves protected health information (PHI) and requires covered entities to obtain authorization from individuals before using or disclosing their PHI for research purposes.
- **FERPA:** The Family Educational Rights and Privacy Act (FERPA) is a federal law that governs the privacy of student education records. FERPA generally prohibits the disclosure of personally identifiable information from education records without the student's consent, except in certain specified circumstances, such as for research purposes.
- **The National Research Act:** The National Research Act is a federal law that emphasizes informed consent, mandates IRB oversight, and ensures the protection of human subjects and promoting ethical practices in research studies.
- **The Privacy Act:** The Privacy Act is a federal law that regulates the collection, use, and dissemination of personal information by federal agencies. The Privacy Act applies to research that involves the use of personal information that is maintained by federal agencies.
- **The Paperwork Reduction Act:** The Paperwork Reduction Act (PRA) is a federal law that requires federal agencies to obtain clearance from the Office of Management and Budget (OMB) before conducting certain types of information collection activities, including research studies that involve the collection of information from the public.

In addition to these federal laws, there are numerous other federal regulations and guidance documents that apply to the use of research data, including regulations governing the protection of human subjects in research conducted by the Department of Defense, the Department of Energy, and the Environmental Protection Agency, among others. It is important for researchers to be familiar with the applicable federal laws and regulations that govern the use of research data in order to ensure that their research is conducted in compliance with applicable legal and ethical standards.

Finally, ethical considerations are essential for ensuring that statistical analysis is conducted in a fair and just manner. This includes issues such as avoiding discrimination and ensuring that statistical methods do not perpetuate systemic biases or injustices. For example, statistical analysis can be used to identify disparities in healthcare access and outcomes between different demographic groups. It is essential that researchers and analysts approach this issue ethically, ensuring that the data is accurate and that any interventions to address the disparities are fair and effective.

Unfortunately, fraud is prevalent in statistical studies. The following website <http://www.retractionwatch.com> has a catalogue of fraudulent and retracted studies and papers. Learning more about probability and statistics will allow you to look at studies with a critical eye and not be fooled by misleading data.

Ethics are essential for statistical analysis because statistical results can have significant consequences, statistical methods can be easily misinterpreted or manipulated, and ethical considerations are particularly important when it comes to data privacy and confidentiality. Additionally, ethical considerations are essential for ensuring that statistical analysis is conducted in a fair and just manner. Researchers and analysts must take steps to ensure that they conduct statistical analysis ethically, including being transparent about their methods and findings, protecting data privacy and confidentiality, and avoiding discrimination and systemic biases. By conducting statistical analysis ethically, we can ensure that statistical results are reliable and trustworthy, and that they contribute to the betterment of society.

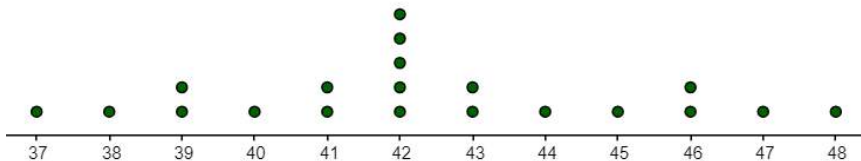
Using statistics for social justice involves leveraging statistical methods and data analysis to address social inequalities, promote fairness, and advocate for marginalized communities. Here are a few examples of how statistics can be applied in the context of social justice:

- **Data-driven advocacy:** Statistics can empower social justice advocates by providing evidence-based insights and compelling narratives. By analyzing and presenting data on social issues, such as income inequality, racial disparities, or environmental injustice, advocates can raise awareness, shape public opinion, and influence policy decisions.
- **Disparity analysis:** Statistics can be used to identify and measure disparities or inequities in various social domains, such as education, healthcare, employment, and criminal justice. By quantifying and documenting these disparities, statistical analysis can help shed light on systemic biases and inform policies and interventions aimed at reducing inequalities.
- **Evaluating interventions and programs:** Statistical methods enable the evaluation of social justice interventions and programs to assess their effectiveness and impact. By collecting and analyzing data on key outcomes, such as educational attainment, employment rates, or access to healthcare, researchers and policymakers can determine whether interventions are achieving their desired goals and identify areas for improvement.
- **Fairness in algorithmic decision-making:** As programming algorithms and automated decision-making systems play an increasingly significant role in various domains, including criminal justice, lending, and hiring, statistical techniques can be employed to examine and address biases and discrimination in these systems. By analyzing data and models used in algorithmic decision-making, statisticians can detect and mitigate potential biases to ensure fair and equitable outcomes.
- **Intersectional analysis:** Intersectionality recognizes that individuals may experience multiple forms of oppression or discrimination based on their intersecting identities, such as race, gender, sexuality, or disability. Statistics can help analyze and understand the unique experiences and challenges faced by underrepresented groups at the intersections of these identities, providing a more comprehensive understanding of social inequities and informing targeted interventions.

These examples highlight how statistics can be a powerful tool in promoting social justice by uncovering inequalities, evaluating interventions, discovering biases, and advocating for underrepresented communities. By applying rigorous statistical methods and ethical considerations, statisticians and researchers can contribute to creating a more equitable and just society.

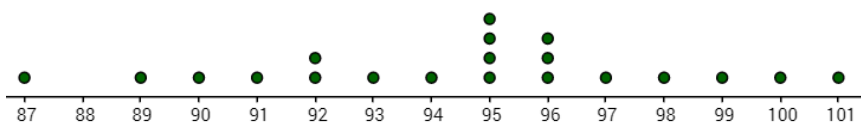
Chapter 1 Exercises

1. The dotplot shows the height of some 5-year-old children measured in inches. Use the distribution of heights to find the approximate answer to the question, “How many inches tall are 5-year-olds?”



“‘Alright,’ he said, ‘but where do we start? How should I know? They say the Ultimate Answer or whatever is Forty-two, how am I supposed to know what the question is? It could be anything. I mean, what’s six times seven?’ Zaphod looked at him hard for a moment. Then his eyes blazed with excitement. ‘Forty-two!’ he cried.”
(Adams, 2002)

2. The dotplot shows the height of some 3-year-old children measured in cm. You are asked, “How many cm tall are 3-year-olds?”



- Is this a statistical question?
 - Use the distribution of heights to approximate the answer for the question, “How many cm tall are 3-year-olds?”
3. What are statistics?
- A question with a variety of answers.
 - A way to measure the entire population.
 - The science of collecting, organizing, analyzing and interpreting data.
 - A question from a survey.
4. What is a statistical question?
- A question where you expect to get a variety of answers and you are interested in the distribution and tendency of those answers.
 - A question using reported statistics.
 - A question on a survey.
 - A question on a census.
5. Which of the following are statistical questions? Select all that apply.
- How old are you?
 - What is the weight of a mouse?
 - How tall are all 3-year-olds?
 - How tall are you?
 - What is the average blood pressure of adult men?
6. In 2010, the Pew Research Center questioned 1,500 adults in the United States to estimate the proportion of the population favoring marijuana use for medical purposes. It was found that 73% are in favor of using marijuana for medical purposes. Identify the individual, variable, population, sample, parameter and statistic.
- Percentage who favors marijuana for medical purposes calculated from sample.
 - Set of 1,500 responses of United States adults who are questioned.
 - All adults in the United States.
 - Percentage who favors marijuana for medical purposes calculated from population.
 - The response to the question “should marijuana be used for medical purposes?”
 - An adult in the United States.

7. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1,000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic, or a population parameter.
 - a) Percentage of all videos on YouTube that are cat videos.
 - b) A video in your sample.
 - c) 2%
 - d) Whether a video is a cat video.

8. A doctor wants to see if a new treatment for cancer extends the life expectancy of a patient versus the old treatment. She gives one group of 25 cancer patients the new treatment and another group of 25 the old treatment. She then measures the life expectancy of each of the patients. Identify the individual, variable, population, sample, parameter and statistic.
 - a) Cancer patient given the new treatment and cancer patient given the old treatment.
 - b) The two groups of 25 cancer patients given the old and new treatments.
 - c) Average life expectancy of 25 cancer patients given the old treatment and average life expectancy of 25 cancer patients given the new treatment.
 - d) Average life expectancy of all cancer patients given the old and new treatment.
 - e) All cancer patients.
 - f) Life expectancy of the cancer patients.

9. The 2010 General Social Survey asked the question, “After an average workday, about how many hours do you have to relax or pursue activities that you enjoy?” to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an individual, a variable, a sample statistic, or a population parameter.
 - a) Average number of hours all Americans spend relaxing after an average workday.
 - b) 1.65
 - c) An American in the sample.
 - d) Number of hours spent relaxing after an average workday.

10. In a study, the sample is chosen by dividing the population by gender, and choosing 30 people of each gender. Which sampling method is used?

11. In a study, the sample is chosen by separating all cars by size, and selecting 10 of each size grouping. What is the sampling method?

12. In a study, the sample is chosen by writing everyone’s name on a playing card, shuffling the deck, then choosing the top 20 cards. What is the sampling method?

13. In a study, the sample is chosen by asking people on the street. What is the sampling method?

14. In a study, the sample is chosen by selecting a room of the house, and appraising all items in that room. What is the sampling method?

15. In a study, the sample is chosen by surveying every 3rd driver coming through a tollbooth. What is the sampling method?

16. Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide (CO) were recorded in parts per million, nitrogen dioxide (NO₂) and ozone (O₃) in parts per hundred million, and coarse particulate matter (PM₁₀) in μg/m³. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM₁₀ and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births. [B. Ritz et al. “Effect of air pollution on preterm birth

among children born in Southern California between 1989 and 1993.” In: *Epidemiology* 11.5 (2000), pp. 502–511.] In this study, identify the variables. Select all that apply.

- a) Ozone
- b) Carbon Monoxide
- c) PM₁₀
- d) Preterm Births in California
- e) Length of Gestation
- f) 143,196 Births
- g) 1989-1993
- h) Nitrogen Dioxide

17. State whether each study is observational or experimental.

- a) You want to determine if cinnamon reduces a person’s insulin sensitivity. You give patients who are insulin sensitive a certain amount of cinnamon and then measure their glucose levels.
- b) A researcher wants to evaluate whether countries with lower fertility rates have a higher life expectancy. They collect the fertility rates and the life expectancies of countries around the world.
- c) A researcher wants to determine if diet and exercise together helps people lose weight over just exercising. The researcher solicits volunteers to be part of the study, and then randomly assigns the volunteers to be in the diet and exercise group or the exercise only group.
- d) You collect the weights of tagged fish in a tank. You then put an extra protein fish food in water for the fish and then measure their weight a month later.

18. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life. [McGowan. “Health Education: Does the Buteyko Institute Method make a difference?” In: *Thorax* 58 (2003).] Which of the following is the main research question?

- a) The Buteyko method causes shallow breathing.
- b) The Buteyko method can reduce asthma symptoms and an improvement in quality of life.
- c) Effectiveness of the Buteyko method.
- d) The patients score on quality of life, activity, asthma symptoms and medication reduction.

19. Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. In the no instruction group, the probability of cheating was found to be uniform across groups based on child’s characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating did not vary by age for boys, it decreased with age for girls. [Alessandro Bucciol and Marco Piovesan. “Luck or cheating? A field experiment on honesty with children.” In: *Journal of Economic Psychology* 32.1 (2011), pp. 73–78.] In this study, identify the variables. Select all that apply.

- a) Age
- b) Sex
- c) Paper Sheet
- d) Cheated or Not
- e) Reward for White Side of Coin
- f) White or Black Side of Coin
- g) Only Child or Not

20. Select the measurement scale Nominal, Ordinal, Interval or Ratio for each scenario.
- A person's age.
 - A person's race.
 - Age groupings (baby, toddler, adolescent, teenager, adult, elderly).
 - Clothing brand.
 - A person's IQ score.
 - Temperature in degrees Celsius.
 - The amount of mercury in a tuna fish.
21. Select the measurement scale Nominal, Ordinal, Interval or Ratio for each scenario.
- Temperature in degrees Kelvin.
 - Eye color.
 - Year in school (freshman, sophomore, junior, senior).
 - The weight of a hummingbird.
 - The height of a building.
 - The amount of iron in a person's blood.
 - A person's gender.
22. State which type of variable each is, qualitative or quantitative?
- A person's age.
 - A person's gender.
 - The amount of mercury in a tuna fish.
 - The weight of an elephant.
 - Temperature in degrees Fahrenheit.
23. State which type of variable each is, qualitative or quantitative?
- The height of a giraffe.
 - A person's race.
 - Hair color.
 - A person's ethnicity.
 - Year in school (freshman, sophomore, junior, senior).
24. State whether the variable is discrete or continuous.
- A person's weight.
 - The height of a building.
 - A person's age.
 - The number of floors of a skyscraper.
 - The number of clothing items available for purchase.
25. State whether the variable is discrete or continuous.
- Temperature in degrees Celsius.
 - The number of cars for sale at a car dealership.
 - The time it takes to run a marathon.
 - The amount of mercury in a tuna fish.
 - The weight of a hummingbird.
26. State whether each study is cross-sectional, retrospective or prospective.
- To see if there is a link between smoking and bladder cancer, patients with bladder cancer are asked if they currently smoke or if they smoked in the past.
 - The Nurses Health Survey was a survey where nurses were asked to record their eating habits over a period of time, and their general health was recorded.
 - A new study is underway to track the eating and exercise patterns of people at different time-periods in the future, and see who is afflicted with cancer later in life.
 - The prices of generic items are compared to the prices of the equivalent named brand items.

27. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
- The quality control officer at a manufacturing plant needs to determine what percentage of items in a batch are defective. The officer chooses every 15th batch off the line and counts the number of defective items in each chosen batch.
 - The local grocery store lets you survey customers during lunch hour on their preference for a new bottle design for laundry detergent.
 - Put all names in a hat and draw a certain number of names out.
 - The researcher randomly selects 5 hospitals in the United States then measures the cholesterol level of all the heart attack patients in each of those hospitals.
28. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
- If you want to calculate the average price of textbooks, you could divide the individuals into groups by major and then conduct simple random samples inside each group.
 - Obtain a list of patients who had surgery at a hospital. Divide the patients according to type of surgery. Draw simple random samples from each group.
 - You want to measure whether a tree in the forest is infected with bark beetles. Instead of having to walk all over the forest, you divide the forest up into sectors, and then randomly pick the sectors that you will travel to. Then record whether a tree is infected or not for every tree in that sector.
 - You select every 3rd customer that comes that orders from your website.
29. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
- In a research study, a list of all registered voters in a city is obtained, and a random sample of 500 voters is selected.
 - A company wants to survey its employees about job satisfaction. They randomly select every 10th employee from the employee list to participate in the survey.
 - A researcher wants to study the purchasing behavior of different age groups in a city. They divide the city into four age groups (18-25, 26-35, 36-45, 46 and above) and randomly select participants from each group.
 - A university wants to conduct a survey among students. The university divides the campus into several sections and randomly selects a few sections. All students within the selected sections are surveyed.
30. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
- A political advocacy group sets up a booth in the city park and invites people passing by to participate in a political opinion poll.
 - A researcher wants to study the opinions of customers who visit a specific store. They randomly select customers as they exit the store to participate in the survey.
 - A government agency wants to collect data about the unemployment rate in different cities. They randomly select a few cities and collect data from all unemployed individuals within those cities.
 - A research team wants to investigate the dietary habits of people in a country. They divide the population into groups based on geographical regions and randomly select participants from each region.
31. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
- A school district wants to assess the performance of students in different grades. They randomly select a few schools from the district and test all students within those selected schools.
 - A company wants to conduct a customer satisfaction survey. They select every 4th customer who makes a purchase at the company's online store to participate in the survey.
 - A survey is conducted at a shopping center, and shoppers passing by are asked to participate. The surveyors approach individuals who are readily available and willing to participate.
 - A survey is conducted to determine the preferences of college students regarding online learning. The research team randomly selects 300 students from a list of all enrolled students to participate in the survey.

32. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
- A researcher wants to study the reading habits of high school students. They randomly select two high schools from each district and survey all students within those selected schools.
 - A research team wants to investigate the prevalence of a disease in different age groups. They divide the population into age categories and randomly select participants from each category.
 - A company wants to study the shopping habits of its customers. They randomly select 50 customers from their entire customer database to participate in a focus group.
 - A survey is conducted to gather feedback on a new mobile application. The researchers select every 5th user who downloads the application to participate in the survey.
33. Which of the following best describes statistical ethics?
- Ensuring the correct statistical analysis is used.
 - Protecting data privacy and confidentiality.
 - Using statistical methods to mislead others.
 - Ignoring ethical considerations in data collection.
34. Why are ethics important in statistical analysis?
- They ensure accurate and reliable results.
 - They help manipulate data for personal gain.
 - They are required by law but have no practical importance.
 - They are irrelevant to statistical analysis.
35. Statistical results can have significant consequences because they:
- Are always accurate and reliable.
 - Can impact decision-making in various fields.
 - Have no practical implications in real-world applications.
 - Are unrelated to ethical considerations.
36. Ethical behavior in statistical analysis includes which of the following?
- Being transparent about methods and findings.
 - Manipulating data to support personal beliefs.
 - Ignoring potential sources of bias.
 - Disregarding privacy and confidentiality of data.
37. Which of the following is a potential negative consequence of conducting statistical analysis unethically?
- Reliable and trustworthy results.
 - Harmful effects on individuals or communities.
 - Accurate interpretations of data.
 - Improved decision-making processes.
38. What is the role of transparency in statistical analysis?
- To hide limitations of the data.
 - To mislead others about the findings.
 - To ensure others can verify the results.
 - To manipulate the data to support a specific outcome.
39. Ethical considerations in statistical analysis include:
- Protecting data privacy and confidentiality.
 - Ignoring the potential for biases.
 - Promoting discrimination and systemic biases.
 - Manipulating data to achieve desired results.

40. Which federal law regulates the use and disclosure of individually identifiable health information?
- a) The Common Rule
 - b) HIPAA
 - c) FERPA
 - d) The Privacy Act
41. What is the purpose of Institutional Review Boards (IRBs)?
- a) To oversee statistical analysis in research studies.
 - b) To protect the privacy and confidentiality of research participants.
 - c) To manipulate data to support specific conclusions.
 - d) To ignore ethical considerations in research studies.
42. Ethical considerations in statistical analysis positively contribute to which of the following?
- a) A fair and just society.
 - b) Misleading interpretations of data.
 - c) Ignoring data privacy and confidentiality.
 - d) Manipulating statistical results for personal gain.

Chapter 2

Organizing Data



- 2.1 Introduction to Summarized Data
- 2.2 Tabular Displays
- 2.3 Graphical Displays
 - 2.3.1 Stem-and-Leaf Plot
 - 2.3.2 Histogram
 - 2.3.3 Ogive
 - 2.3.4 Pie Chart
 - 2.3.5 Bar Graph
 - 2.3.6 Pareto Chart
 - 2.3.7 Stacked Column Chart
 - 2.3.8 Multiple or Side-by-Side Bar Graph
 - 2.3.9 Time-Series Plot
 - 2.3.10 Scatter Plot
 - 2.3.11 Misleading Graphs

2.1 Introduction to Summarized Data

Once a sample is collected, we can organize and present the data in tables and graphs. These tables and graphs help summarize, interpret and recognize characteristics within the data more easily than raw data. There are many types of graphical summaries. We will concentrate mostly on the ones that we can use technology to create.

A population is a collection of all the measurements from the individuals of interest. Remember, in most cases you cannot collect data on the entire population, so you have to take a sample. Now you have a large number of data values. What can you do with them? Just looking at a large set of numbers does not answer our questions. If we organize the data into a table or graph, we can see patterns in the data. Ultimately, though, you want to be able to use that table or graph to interpret the data, to describe the distribution of the data set, explore different characteristics of the data and make inferences about the original population.

Some characteristics to look for in tables and graphs:

1. Center: middle of the data set, also known as the average.
2. Variation: how spread out is the data.
3. Distribution: shape of the data.
4. Outliers: data values that are far from the majority of the data.
5. Time: changing characteristics of the data over time.

There is technology that will create most of the graphs you need, though it is important for you to understand the basics of how they are created.

Qualitative data are words describing a characteristic of the individual. Qualitative data are graphed using several different types of graphs, bar graphs, Pareto charts, and pie charts. Quantitative data are numbers that we count or measure. Quantitative data are graphed using stem-and-leaf plots, dotplots, histograms, ogives, and time series.

The bar graph for quantitative data called a **histogram** looks similar to a bar graph for qualitative data, except there are some major differences. First, in a bar graph the categories can be put in any order on the horizontal axis. There is no set order for these data values. You cannot say how the data is distributed based on the shape, since the shape can change just by putting the categories in different orders. With quantitative data, the data are in specific orders since you are dealing with numbers. With quantitative data, you can talk about a distribution; the shape changes depending on how many categories you set up. This shape of the quantitative graph is called a **frequency distribution**.

This leads to the second difference from bar graphs. In a bar graph, the categories are determined by the name of the label. In quantitative data, the categories are numerical categories, and the frequencies are determined by how many categories (or what are called classes) you choose. There can be many different classes depending on the point of view of the author and how many classes there are.

The third difference is that the categories touch with quantitative data, and there will be no gaps in the graph. The reason that bar graphs have gaps is to show that the categories do not continue on, as they do in quantitative data.

2.2 Tabular Displays

Frequency Tables for Quantitative Data

To create many of these graphs, you must first create the frequency distribution. The idea of a frequency distribution is to take the interval that the data spans and divide it into equal sized subintervals called classes. The grouped frequency distribution gives either the frequency (count) or the relative frequency (usually expressed as a percent) of individuals who fall into each class. When creating frequency distributions, it is important to note that the number of classes that are used and the value of the first class boundary will change the shape of, and hence the impression given by, the distribution. There is no one correct width to the class boundaries. It usually takes several tries to create a frequency distribution that looks just the way you want. As a reader and interpreter of such tables, you should be aware that such features are selected so that the table looks the way it does to show a particular point of view. For small samples, we usually have between four and seven classes, and as you get more data you will need more classes.

Example 2-1: We will start with an example of a random sample of 35 ages from credit card applications given below in no particular order. Organize the data in a frequency distribution table.

46	47	49	25	46	22	42	32	39
24	46	40	39	27	25	30	31	29
33	27	46	21	29	20	26	39	26
25	25	26	35	49	33	26	30	

Solution: As you can see, the data is hard to interpret in this format. If we were to peruse through the data, we could find that minimum age is 20 and the maximum age is 49. If we cut the age groups up in to 10-year intervals, we get only three classes 20-29, 30-39 and 40-49. Although this would work, if we had more classes, we can sometimes see trends within the data at a more granular level. If we split the ages up in to 5-year intervals to 20-24, 25-29, 30-34, 35-39, 40-44, and 45-49 we would get six classes.

Note that the class limits should never overlap; we call these mutually exclusive classes. For example, if your class went from 20-25, 25-30 etc., the 25-year-olds would fall within both classes. Also, make sure each of the classes have the same width. A more formal way to pick your classes uses the following process.

Steps involved in making a frequency distribution table:

1. Find the range = largest value – smallest value.
2. Pick the number of classes to use. Usually, the number of classes is between five and twenty. Five classes are used if there are a small number of data points and twenty classes if there are a large number of data points (over 1,000 data points).
3. Class width = $\frac{\text{range}}{\# \text{ of classes}}$. Always round up to the next integer (if the answer is already a whole number, go to the next integer). If you do not round up, your last class will not contain your largest data value, and you would have to add another class just for it. If you round up, then your largest data value will fall in the last class, and there are no issues.
4. Create the classes. Each class has limits that determine which values fall in each class. To find the **class limits**, set the smallest value in the data set as the lower class limit for the first class. Then add the class width to the lower class limit to get the next lower class limit. Repeat until you get all the classes. The upper class limit for a class is one less than the lower limit for the next class.
5. If your data value has decimal places, then round up the class width to the nearest value with the same number of decimal places as the original data. As an example, if your data was out to two decimal places, and you divided your range by the number of classes to get 4.8333, then the class width would round the second decimal place up and end on 4.84.
6. The frequency for a class is the number of data values that fall in the class.

For the age data let us use 6 classes. Find the range by taking $49 - 20 = 29$ and divide this by the number of classes $29/6 = 4.8333$. Round this number up to 5 and use 5 for the class width. Once you determine your class width and class limits, place each of these classes in a table and then tally up the ages that fall within each class. Count the tally marks and record the number in the frequency table.

The total of the frequency column should be the number of observations in the data. You may want to total the frequencies to make sure you did not leave any of the numbers out of the table.

Class Limits	Tally
20-24	
25-29	
30-34	
35-39	
40-44	
45-49	

Class	Frequency
20-24	4
25-29	12
30-34	6
35-39	4
40-44	2
45-49	7
Total	35

Using the frequency table, we can now see that there are more people in the 25-29 year-old class, followed by the 45-49 year-old class.

We call this most frequent category the **modal class**. There may be no mode at all or more than one mode.

Figure 2-1

Frequency Tables for Qualitative Data

A frequency distribution can also be made for qualitative data.

Example 2-2: Suppose you have the following data for which type of car students at a college drive. Make a frequency table to summarize the data.

Ford	Honda	Nissan	Chevy	Chevy
Chevy	Chevy	Toyota	Chevy	Saturn
Honda	Toyota	Nissan	Honda	Toyota
Toyota	Nissan	Ford	Toyota	Chevy
Toyota	Ford	Chevy	Chevy	Chevy
Nissan	Toyota	Toyota	Ford	Nissan
Kia	Nissan	Nissan	Nissan	Honda
Nissan	Mercedes	Honda	Toyota	Toyota
Chevy	Chevy	Porsche	Chevy	Toyota
Toyota	Ford	Hyundai	Honda	Nissan

Solution: The list of data is hard to analyze, so you need to summarize it. The classes in this case are the car brands. However, several car brands only have one car in the list. In that case, it is easier to make a category called “other” for the categories with low frequencies.

Category	Frequency
Chevy	12
Ford	5
Honda	6
Nissan	10
Toyota	12
Other	5
Total	50

Count how many of each type of cars there are: there are 12 Chevys, 5 Fords, 6 Hondas, 10 Nissans, 12 Toyotas, and 5 other brands (Hyundai, Kia, Mercedes, Porsche, and Saturn). Place the other brands into a frequency distribution table alphabetically:

For nominal data, either alphabetize the classes or arrange the classes from most frequent to least frequent, with the “other” category always at the end. For ordinal data put the classes in their order with the “other” category at the end.

Relative Frequency Tables

Frequencies by themselves are not as useful to tell other people what is going on in the data. If you want to know what percentage the category is of the total sample then we can use the relative frequency of each category. The relative frequency is just the frequency divided by the total. The relative frequency is the proportion in each category and may be given as a decimal, percentage, or fraction.

Using the car data’s frequency distribution, we will create a third column labeled relative frequency. Take each frequency and divide by the sample size, see Figure 2-2. The relative frequencies should add up to one (ignoring rounding error).

Type of Car	Frequency	Relative Frequency
Chevy	12	$12/50 = 0.24$
Ford	5	$5/50 = 0.1$
Honda	6	$6/50 = 0.12$
Nissan	10	$10/50 = 0.2$
Toyota	12	$12/50 = 0.24$
Other	5	$5/50 = 0.1$
Total	50	1

Figure 2-2

Many people understand percentages better than proportions so we may want to multiply each of these decimals by 100% to get the following relative frequency percent table shown in Figure 2-3.

Type of Car	Percent
Chevy	24%
Ford	10%
Honda	12%
Nissan	20%
Toyota	24%
Other	10%
Total	100%

Figure 2-3

We can summarize the car data and see that for college students Chevy and Toyota make up 48% of the car models.

Excel

Example 2-3: Recall the frequency table for the credit card applicants. The relative frequency table for the random sample of 35 ages from credit card applications follows.

Class	Frequency	Relative Frequency
20-24	4	$4/35 = 0.1143$
25-29	12	$12/35 = 0.3429$
30-34	6	$6/35 = 0.1714$
35-39	4	$4/35 = 0.1143$
40-44	2	$2/35 = 0.0571$
45-49	7	$7/35 = 0.2$
Total	35	1

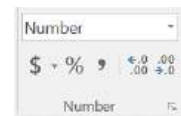
Making a relative frequency table using Excel.

Solution: In Excel, type your frequencies into a column and then in the next column type in =(cell reference number)/35. Then copy and paste the formula in cell B2 down the page. If you used the cell reference number, Excel will automatically change the copied cells to the next row down.

	A	B
1	Frequency	Relative Frequency
2	4	=A2/35
3	12	=A3/35
4	6	=A4/35
5	4	=A5/35
6	2	=A6/35
7	7	=A7/35
8	=SUM(A2:A7)	=SUM(B2:B7)

You get the following relative frequency table. The sum of the relative frequencies will be one (the sum may add to 0.9999 or 1.0001 if you are doing the calculations by hand due to rounding).

To get Excel to show the percentage instead of the proportion highlight the relative frequencies and select the percent % button on the Home tab.



Class	Frequency	Relative Frequency	Relative Frequency Percent
20-24	4	0.1143	11%
25-29	12	0.3429	34%
30-34	6	0.1714	17%
35-39	4	0.1143	11%
40-44	2	0.0571	6%
45-49	7	0.2	20%
Total	35	1	100%

The relative frequency table lets us quickly see that a little more than half $34\% + 20\% = 54\%$ of the ages of the credit card holders are between the ages of 25 – 29 and 45 – 49 years-old.

Cumulative & Cumulative Relative Frequency Tables

Another useful piece of information is how many data points fall below a particular class. As an example, a teacher may want to know how many students received below a 70%, a doctor may want to know how many adults have cholesterol above 160, or a manager may want to know how many stores gross less than \$2,000 per day. This calculation is known as a **cumulative frequency** and is used for ordinal or quantitative data. If you want to know what percent of the data falls below a certain class, then this fact would be a **cumulative relative frequency**.

To create a **cumulative frequency distribution**, count the number of data points that are below the upper class limit, starting with the first class and working up to the top class. The last upper class should have all of the data points below it.

Example 2-4: Recall the credit card applicants. Make a cumulative frequency table.

Class	Frequency	Relative Frequency
20-24	4	11%
25-29	12	34%
30-34	6	17%
35-39	4	11%
40-44	2	6%
45-49	7	20%
Total	35	100%

Solution: To find the cumulative frequency, carry over the first frequency of 4 to the first row of the cumulative frequency column. Then take this 4 and add it to the next frequency of 12 to get 16 for the second cumulative frequency value. For the third cumulative frequency, take 16 and add to the third frequency of 6 to get 22. Keep doing this additive process until you finish the column.

The cumulative frequency in the last class should have the last number as the sample size. The cumulative relative frequency is just the cumulative frequency divided by the sample size. The cumulative relative frequencies should have one as the last number.

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
20-24	4	11%	4	$4/35 = 0.1143$
25-29	12	34%	$4 + 12 = 16$	$16/35 = 0.4571$
30-34	6	17%	$16 + 6 = 22$	$22/35 = 0.6286$
35-39	4	11%	$22 + 4 = 26$	$26/35 = 0.7429$
40-44	2	6%	$26 + 2 = 28$	$28/35 = 0.8$
45-49	7	20%	$28 + 7 = 35$	$35/35 = 1$
Total	35	100%		

Use Excel to add up the values for you.

	A	B	C	D
1	Frequency	Relative Frequency	Excel Formula	Cumulative Frequency
2	4	=A2/35	4	4
3	12	=A3/35	=C2+A3	16
4	6	=A4/35	=C3+A4	22
5	4	=A5/35	=C4+A5	26
6	2	=A6/35	=C5+A6	28
7	7	=A7/35	=C6+A7	35
8	=SUM(A2:A7)	=SUM(B2:B7)		

You can also express the cumulative relative frequencies as a percentage instead of a proportion.

Class	Cumulative Frequency	Cumulative Relative Frequency
20-24	4	11%
25-29	16	46%
30-34	22	63%
35-39	26	74%
40-44	28	80%
45-49	35	100%

If a manager wanted to know how many applicants were under the age of 40, we could look across the 35-39 year-old class to see that there were 26 applicants that were under 40 (39 years old or younger), or if we used the cumulative relative frequency, 74% of the applicants were under 40 years old. If the manager wants to know how many applicants were over 44, we could subtract $100\% - 80\% = 20\%$.

Contingency Tables

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. A contingency table summarizes the frequency of two qualitative variables. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. There are other names for contingency tables. Excel calls the contingency table a pivot table. Other common names are two-way table, cross-tabulation, or cross-tab for short.

Example 2-5: A fitness center coach kept track of members over the last year. They recorded if the person stretched before they exercised, and whether they sustained an injury. The following contingency table shows their results. Find the relative frequency for each value in the table.

	Injury	No Injury
Stretched	52	270
Did Not Stretch	21	57

Solution: Each value in the table represents the number of times a particular combination of variable outcomes occurred, for example, there were 57 members that did not stretch and did not sustain an injury. It is helpful to total up the categories. The row totals provide the total counts across each row (i.e., $52 + 270 = 322$), and column totals are total counts down each column (i.e., $52 + 21 = 73$). See Figure 2-4.

	Injury	No Injury	Total
Stretched	52	270	322
Did Not Stretch	21	57	78
Total	73	327	400

Figure 2-4

We can quickly summarize the number of athletes for each category. The bottom right-hand number in Figure 2-4 is the grand total and represents the total number of 400 people. There were 322 people that stretched before exercising. There were 73 people that sustained an injury while exercising, etc.

If we find the relative frequency for each value in the table, we can find the proportion of the 400 people for each category. To find a relative frequency we divide each value in the table by the grand total. See Figure 2-5.

	Injury	No Injury	Total
Stretched	$52/400 = 0.13$	$270/400 = 0.675$	$322/400 = 0.805$ or $0.13 + 0.675 = 0.805$
Did Not Stretch	$21/400 = 0.0525$	$57/400 = 0.1425$	$78/400 = 0.195$ or $0.0525 + 0.1425 = 0.195$
Total	$73/400 = 0.1825$ or $0.13 + 0.0525 = 0.1825$	$327/400 = 0.8175$ or $0.675 + 0.1425 = 0.8175$	$400/400 = 1$ or the sum of either the row or column totals

Figure 2-5

When data is collected, it is usually presented in a spreadsheet where each row represents the responses from an individual or case.

“Of course, one never has the slightest notion what size or shape different species are going to turn out to be, but if you were to take the findings of the latest Mid-Galactic Census report as any kind of accurate Guide to statistical averages you would probably guess that the craft would hold about six people, and you would be right. You'd probably guessed that anyway. The Census report, like most such surveys, had cost an awful lot of money and didn't tell anybody anything they didn't already know - except that every single person in the Galaxy had 2.4 legs and owned a hyena. Since this was clearly not true the whole thing had eventually to be scrapped.”
(Adams, 2002)

Example 2-6: Make a pivot table using Excel. A random sample of 500 records from the 2010 United States Census were downloaded to Excel. Below is an image of just the first 20 people.

There are seven variables:

- State
- Total family income (in United States dollars)
- Age
- Biological Sex (with reported categories female and male)
- Race (with reported categories American Indian or Alaska Native, Black, Chinese, Japanese, Other Asian or Pacific Islander, Two major races, White, Other)
- Marital status (with reported categories Divorced, Married/spouse absent, Married/spouse present, Never married/single, Separated, Widowed)
- Total personal income (in United States dollars).

	A	B	C	D	E	F	G
1	stateFIPScod	totalFamilyI	age	sex	raceGeneral	maritalStatus	totalPersona
2	Alabama	5500	73	Female	Black	Divorced	3000
3	Alabama	63820	40	Male	White	Married/spouse present	47950
4	Alabama	11200	60	Female	Black	Divorced	11200
5	Alabama	34500	43	Female	Other	Married/spouse present	15300
6	Alabama	33600	7	Male	White	Never married/single	NA
7	Arizona	32500	9	Female	White	Never married/single	NA
8	Arizona	46800	53	Female	White	Married/spouse absent	30000
9	Arizona	30000	60	Female	White	Widowed	30000
10	Arizona	0	67	Female	White	Widowed	0
11	Arizona	51000	27	Male	White	Married/spouse present	14000
12	Arizona	6000	10	Male	White	Never married/single	NA
13	Arizona	45000	48	Male	Japanese	Married/spouse present	45000
14	Arkansas	10940	6	Female	Black	Never married/single	NA
15	California	52000	40	Female	White	Married/spouse present	27000
16	California	156000	65	Female	White	Married/spouse present	10000
17	California	59000	80	Male	White	Married/spouse present	52100
18	California	37500	46	Female	White	Married/spouse present	17500
19	California	67500	3	Female	White	Never married/single	NA
20	California	21800	6	Male	Other Asian	Never married/single	NA
21	California	79000	26	Female	White	Never married/single	25000

Solution:

In Excel, select the Insert tab, then select Pivot Table.

stateFIPScod	totalFamilyIncome	age	sex	raceGeneral	maritalStatus	totalPersonalIncome
Alabama	5500	73	Female	Black	Divorced	3000
Alabama	63820	40	Male	White	Married/spouse present	47950
Alabama	11200	60	Female	Black	Divorced	11200
Alabama	34500	43	Female	Other	Married/spouse present	15300
Alabama	33600	7	Male	White	Never married/single	NA
Arizona	32500	9	Female	White	Never married/single	NA
Arizona	46800	53	Female	White	Married/spouse absent	30000
Arizona	30000	60	Female	White	Widowed	30000
Arizona	0	67	Female	White	Widowed	0
Arizona	51000	27	Male	White	Married/spouse present	14000
Arizona	6000	10	Male	White	Never married/single	NA
Arizona	45000	48	Male	Japanese	Married/spouse present	45000
Arkansas	10940	6	Female	Black	Never married/single	NA
California	52000	40	Female	White	Married/spouse present	27000
California	156000	65	Female	White	Married/spouse present	10000
California	59000	80	Male	White	Married/spouse present	52100
California	37500	46	Female	White	Married/spouse present	17500
California	67500	3	Female	White	Never married/single	NA
California	21800	6	Male	Other Asian	Never married/single	NA
California	79000	26	Female	White	Never married/single	25000

Excel should automatically select all 500 rows in the Table/Range cell, if not then use your mouse and highlight all the data including the labels. Then select OK.

Each version of Excel may look different at this point. One common area, though, is the bottom right-hand drag and drop area of the Pivot Table dialog box.

PivotTable Fields

Choose fields to add to report:

- stateFIPScod
- totalFamilyIncome
- age
- sex
- raceGeneral
- maritalStatus

Drag fields between areas below:

FILTERS	COLUMNS
ROWS	VALUES

Defer Layout Update... **UPDATE**

PivotTable Fields

Choose fields to add to report:

- age
- sex
- raceGeneral
- maritalStatus
- totalPersonalIncome

Drag fields between areas below:

FILTERS	COLUMNS
ROWS	VALUES

sex

maritalSt... Count of ...

Defer Layout Update... **UPDATE**

Drag the sex variable to the COLUMNS box, and marital status variable to the ROWS box.

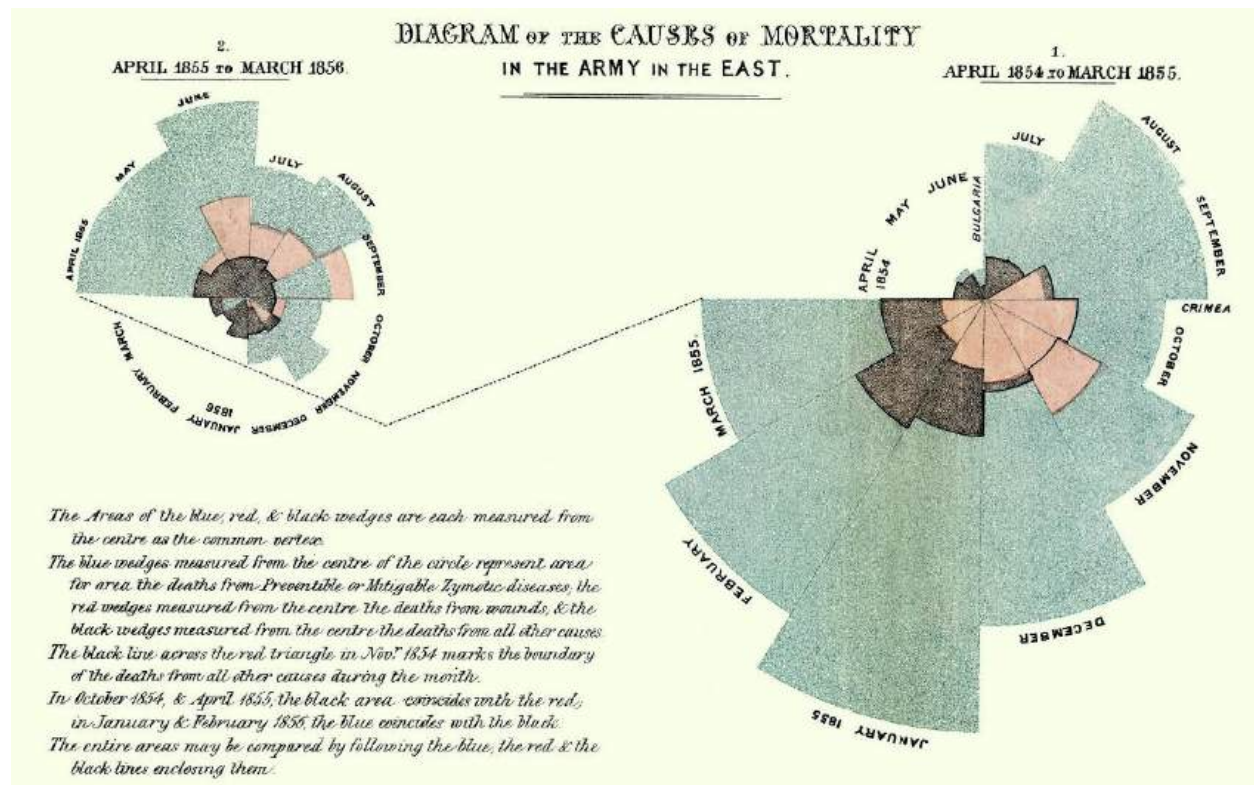
You will see the contingency table column and row headers appear as you drop the variables in the boxes.

To get the counts to appear, drag and drop marital status into the Values box and the default will usually say, “Count of maritalStatus,” if not then change it to count in the drop-down menu. A contingency table should appear on your spreadsheet as you fill in the pivot table dialogue box. Contingency tables go by many names including pivot tables, two-way tables, cross tabulations, or cross-tabs.

Count of Marital Status	Column Labels		
Row Labels	Female	Male	Grand Total
Divorced	21	17	38
Married/spouse absent	5	9	14
Married/spouse present	92	100	192
Never married/single	93	129	222
Separated	1	2	3
Widowed	20	11	31
Grand Total	232	268	500

2.3 Graphical Displays

Statistical graphs are useful in getting the audience’s attention in a publication or presentation. Data presented graphically is easier to summarize at a glance compared to frequency distributions or numerical summaries. Graphs are useful to reinforce a critical point, summarize a data set, or discover patterns or trends over a period of time. Florence Nightingale (1820-1910) was one of the first people to use graphical representations to present data. Nightingale was a nurse in the Crimean War and used a type of graph that she called polar area diagram, or coxcombs to display mortality figures for contagious diseases such as cholera and typhus.



Nightingale-mortality.jpg. (2021, May 18). Wikimedia Commons, the free media repository. Retrieved July 2021 from <https://commons.wikimedia.org/w/index.php?title=File:Nightingale-mortality.jpg&oldid=561529217>.

It is hard to provide a complete overview of the most recent developments in data visualization with the onset of technology. The development of a variety of highly interactive software has accelerated the pace and variety of graphical displays across a wide range of disciplines.

2.3.1 Stem-and-Leaf Plot

Stem-and-leaf plots (or stemplots) are a useful way of getting a quick picture of the shape of a distribution by hand. Turn the graph sideways and you can see the shape of your data. You can now easily identify outliers. Each observation is divided into two pieces; the stem and the leaf. If the number is just two digits, then the stem would be the tens digit and the leaf would be the ones digit. When a number is more than two digits then the cut point should split the data into enough classes that is useful to see the shape of the data.

To create a stem-and-leaf plot:

1. Separate each observation into a stem and a leaf.
2. Write the stems in a vertical column in ascending order (from smallest to largest). Fill in missing numbers even if there are gaps in the data. Draw a vertical line to the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order.

Example 2-7: Create a stem-and-leaf plot for the sample of 35 ages.

46	47	49	25	46	22	42
24	46	40	39	27	25	30
33	27	46	21	29	20	26
25	25	26	35	49	33	26
32	31	39	30	39	29	26

Solution: Divide each number so that the tens digit is the stem and the ones digit is the leaf. The smallest observation is 20. The stem = 2 and the leaf = 0. The next value is 21 and the stem = 2 and the leaf = 1, up to the last value of 49 which would have a stem = 4 and a leaf = 9. If we use the tens categories, we have the stems 2, 3 and 4. Line up the stems without skipping a number even if there are no values in that stem. In other words, the stems should have equal spacing (for example, count by ones, tens, hundreds, thousands, etc.). Then place a vertical line to the right of the stems. In each row put the leaves with a space between each leaf. Sort each row from smallest to largest. In Figure 2-6 the $2 | 0 = 20$.

```

2 | 0 1 2 4 5 5 5 5 6 6 6 6 7 7 9 9
3 | 0 0 1 2 3 3 5 9 9 9
4 | 0 2 6 6 6 6 7 9 9

```

Figure 2-6

It is hard to see the shape with so few classes and so many leaves in each class.

We can break each stem in half, putting leaves 0-4 in the first row and 5-9 in the second row, as in Figure 2-7.

```

2 | 0 1 2 4
2 | 5 5 5 5 6 6 6 6 7 7 9 9
3 | 0 0 1 2 3 3
3 | 5 9 9 9
4 | 0 2
4 | 6 6 6 6 7 9 9

```

Figure 2-7

Now, add labels and make sure the leaves are in ascending order. Be careful to line the leaves up in columns. You need to be able to compare the lengths of the rows when you interpret the graph.

Imagine lines around the leaves and turn the graph 90 degrees to the left. You can now see in Figure 2-8 the shape of the distribution. Note that Excel uses the upper class limit for the axis label.

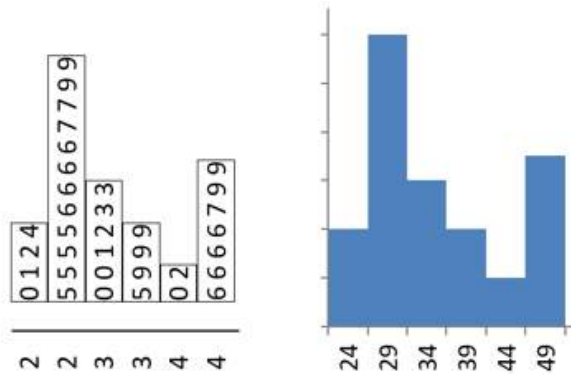


Figure 2-8

If a leaf takes on more than the ones category then supply a footnote at the bottom of the plot with the units.

Example 2-8: A small sample of house prices in thousands of dollars was collected: 375, 189, 432, 225, 305, 275. Make a stem-and-leaf plot.

Solution: If we were to split the stem and leaf between the ones and tens place, then we would need stems going from 18 up to 43. Twenty-six stems for only six data points are too many. The next break then for a stem would be between the tens and hundreds. This would give stems from 1 to 4. Then each leaf will be the ones and tens. For example, then number 375 would have a stem = 3 and a leaf = 75.

1	89
2	25 75
3	05 75
4	32

Leaf = \$1,000

Example 2-9: A small sample of coffee prices: 3.75, 1.89, 4.32, 2.25, 3.05, 2.75 was collected. Make a stem-and-leaf plot.

Solution:

1	89
2	25 75
3	05 75
4	32

Leaf = \$0.01

Note that the last two stem-and-leaf plots look identical except for the footnote. It is important to include units to tell people what the stems and leaves mean by inserting a key or legend.

Back-to-back stem-and-leaf plots let us compare two data sets on the same number line. The two samples share the same set of stems. The sample on the right is written backward from largest leaf to smallest leaf, and the sample on the left has leaves from smallest to largest.

Example 2-10: Use the following back-to-back stem-and-leaf plot to compare pulse rates before and after exercise.

Before	Pulse Rates	After
9 8 8 7 6 5 2	6	
9 8 8 8 6 5 5 5 1 1 0 0	7	
8 8 7 5 4 2	8	5 6 6 7 8 9
4 0	9	0 1 1 2 3 4 5 5 6 8
4	10	0 1 4 6 7
	11	6 7
	12	4 5 8

Solution: The group on the left has leaves going in descending order and represent the pulse rates before exercise. The stems are in the middle column. The group on the right has leaves going in ascending order and represent the pulse rates after exercise. The first row has pulse rates of 62, 65, 66, 67, 68, 68 and 69. The last row of pulse rates are 124, 125, and 128.

2.3.2 Histogram

A **histogram** is a graph for quantitative data (we call these bar graphs for qualitative data). The data is divided into a number of classes. The class limits become the horizontal axis demarcated with a number line and the vertical axis is either the frequency or the relative frequency of each class. Figure 2-9 is an example of a histogram.

The histogram for quantitative data looks similar to a bar graph, except there are some major differences.

First, in a bar graph the categories can be put in any order on the horizontal axis. There is no set order for these nominal data. You cannot say how the data is distributed based on the shape, since the shape can change just by putting the categories in different orders. With quantitative data, the data are in a specific order, since you are dealing with numbers. With quantitative data, you can talk about a distribution shape.

This leads to the second difference from bar graphs. In a bar graph, the categories that you made in the frequency table were the words used for the category name. In quantitative data, the categories are numerical categories, and the numbers are determined by how many classes you choose. If two people have the same number of categories, then they will have the same frequency distribution. Whereas in qualitative data, there can be many different categories depending on the point of view of the author.

The third difference is that the bars touch with quantitative data, and there will be no gaps in the graph. The reason that bar graphs have gaps is to show that the categories do not continue on, as they do in quantitative data. Since the graph for quantitative data is different from qualitative data, it is given a different name of histogram.

Some key features of a histogram:

- Equal spacing on each axis
- Bars are the same width
- Label each axis and title the graph
- Show the scale on the frequency axis
- Label the categories on the category axis
- The bars should touch at the class boundaries.

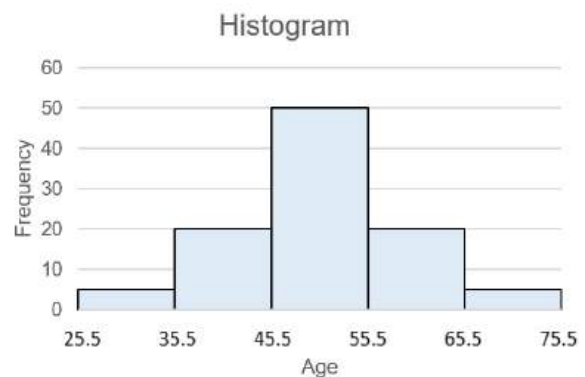


Figure 2-9

To create a histogram, you must first create a frequency distribution. Software and calculators can create histograms easily when a large amount of sample data is being analyzed.

Excel

To create a histogram in Excel you will need to first install the Data Analysis tool.

If your Data Analysis is not showing in the Data tab, follow the directions for installing the free add-in here: <https://support.office.com/en-us/article/Load-the-Analysis-ToolPak-in-Excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4>.

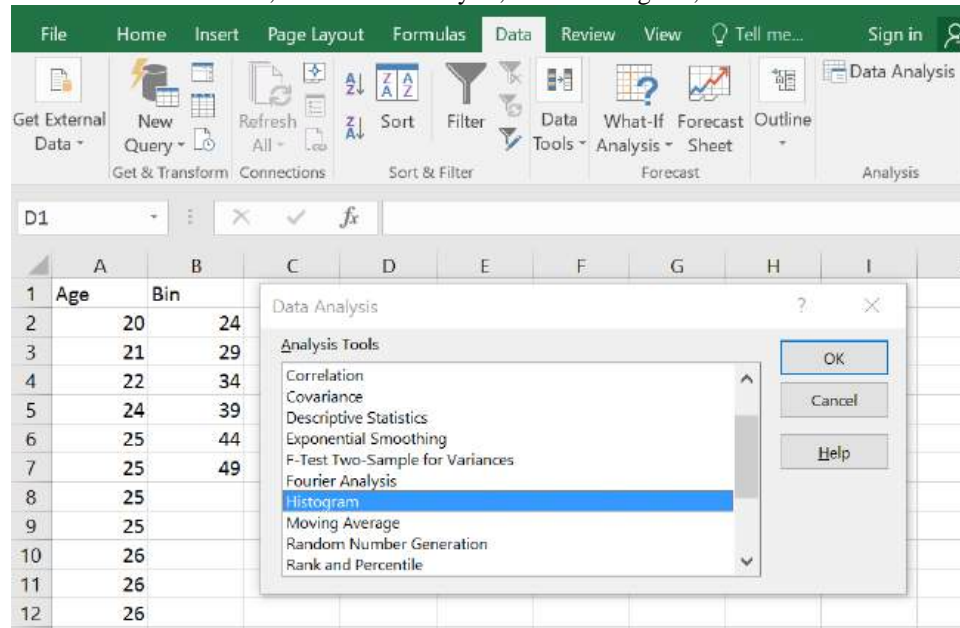
Type in the data into one blank column in any order. If you want to have class widths other than Excel's default setting, type in a new column the endpoints of each class found in your frequency distribution, these are called the bins in Excel.

Example 2-11: Using the sample of 35 ages, make a histogram using Excel.

46	47	49	25	46	22	42
24	46	40	39	27	25	30
33	27	46	21	29	20	26
25	25	26	35	49	33	26
32	31	39	30	39	29	26

Solution: Type the data in any order into column A and the bins in order in column B as shown below.

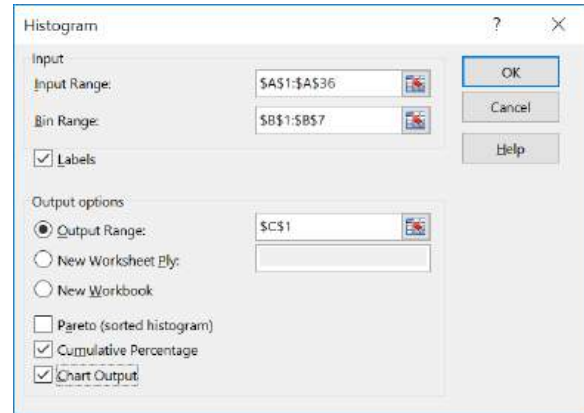
Then select the Data tab, select Data Analysis, select Histogram, then select OK.



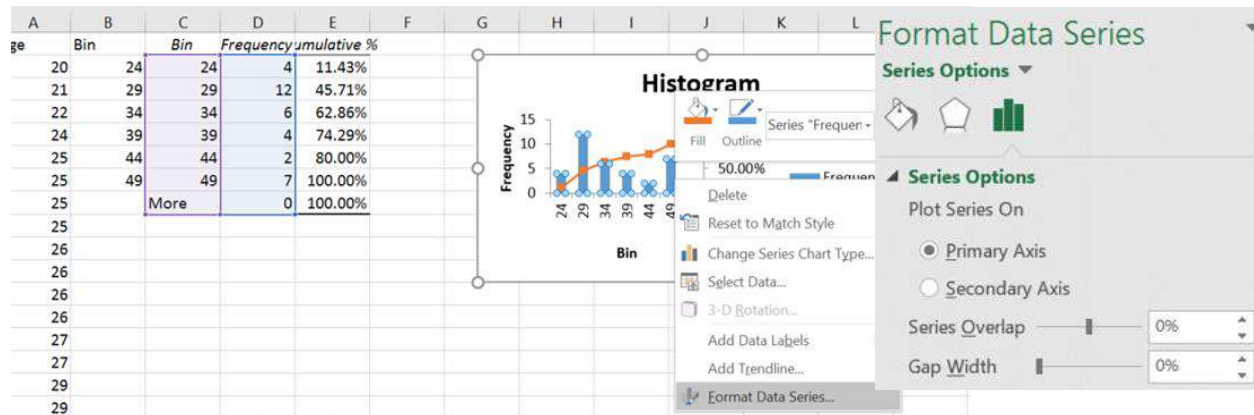
In the dialogue box, click into the Input Range box, then use your mouse and highlight the ages including the label.

Then click into the Bin Range box and use your mouse to highlight the bins including the label.

Select the box for Labels only if you included the labels in your ranges. You can have your output default to a new worksheet, or select the circle to the left of Output Range, click into the box to the right of Output Range and then select one blank cell on your spreadsheet where you want the top left-hand corner of your table and graph to start. Then check the boxes next to Cumulative Percentage and Chart Output. Then select OK, and see below.



A histogram needs to have bars that touch, which is not the default in Excel. To get the bars to touch, right-click on one of the blue bars and select Format Data Series and slide the Gap Width to 0%.



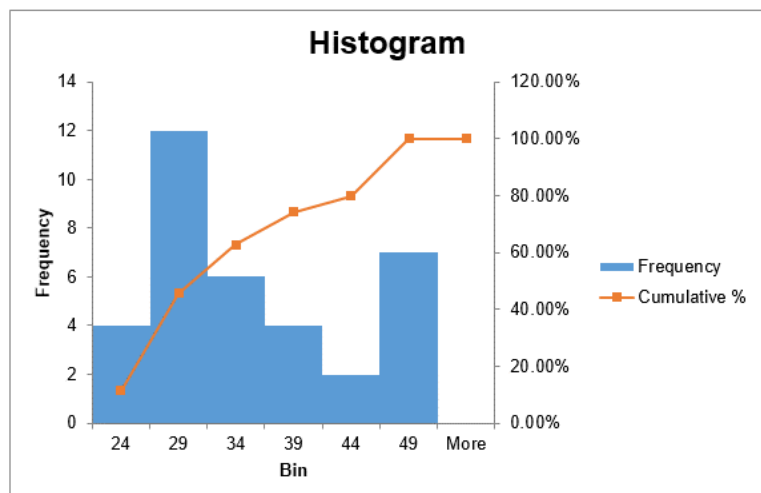
Excel produces both a frequency table and a histogram. The table has the frequencies and the cumulative relative frequencies.

Bin	Frequency	Cumulative %
24	4	11.43%
29	12	45.71%
34	6	62.86%
39	4	74.29%
44	2	80.00%
49	7	100.00%
More	0	100.00%

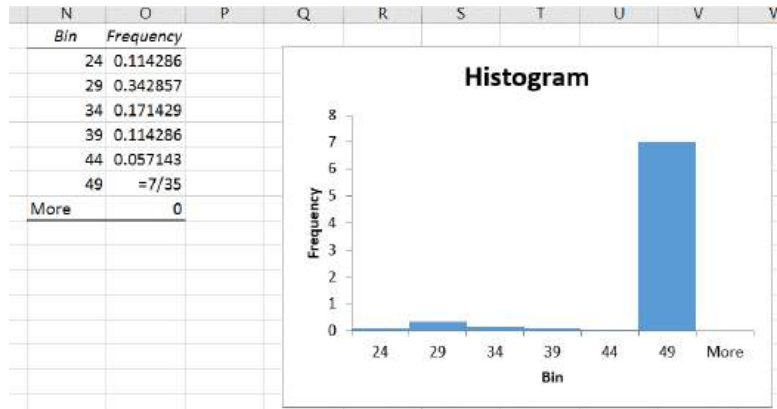
The histogram has bars for the height of each frequency and then makes a line graph of the cumulative relative frequencies over the bars. This red line is a line graph of the cumulative relative frequencies, also called an ogive and is discussed in a later section.

It is important to note that the number of classes that are used and the value of the first class boundary will change the shape of the histogram.

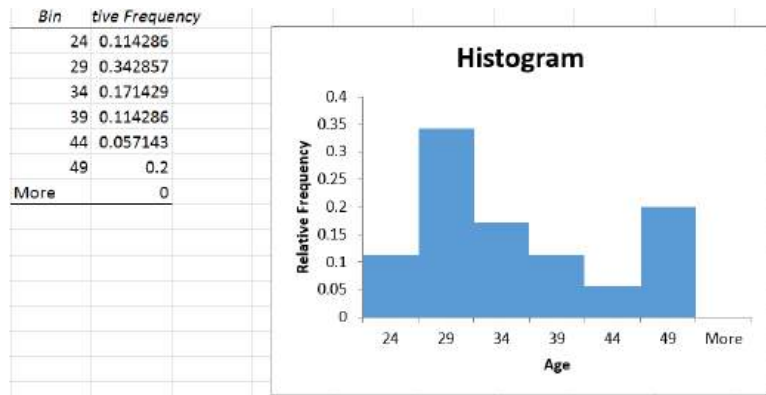
A relative frequency histogram is when the relative frequencies are used for the vertical axis instead of the frequencies and the y-axis will represent a percent instead of the number of people.



In Excel, after you create your histogram, you can manually change the frequency column to the relative frequency values by dividing each number by the sample size. Here is a screen shot just as the last number was changed, note as soon as you press enter the bars will shrink and adjust.

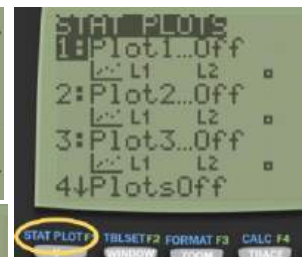


After the last value =7/35 was entered and the label changed to Relative Frequency you get the following graph.

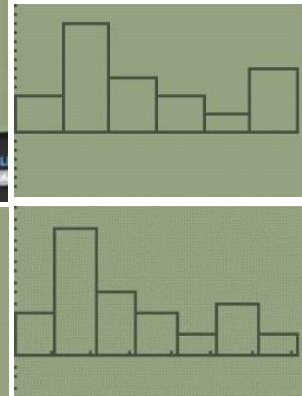


The shape of the histogram will be the same for the relative frequency distribution and the frequency distribution; the height, though, is the proportion instead of frequency.

TI-84: To make a histogram, enter the data by pressing [STAT]. The first option is already highlighted (1:Edit) so you can either press [ENTER] or [1]. Make sure the cursor is in the list, not on the list name and type the desired values pressing [ENTER] after each one.

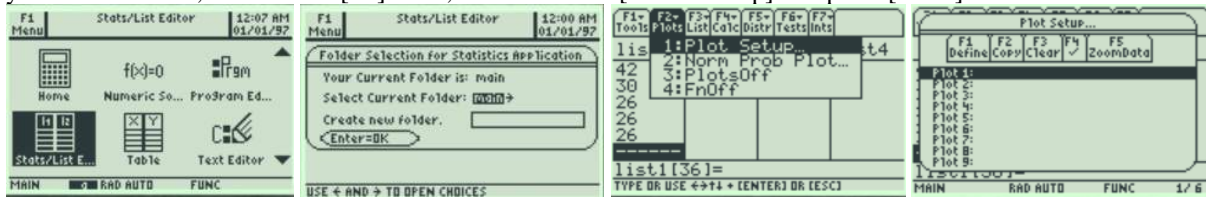


Press [2nd] [QUIT] to return to the home screen. To clear a previously stored list of data values, arrow up to the list name you want to clear, press [CLEAR], and then press enter. An alternative way is press [STAT], press 4 for 4:ClrList, press [2nd], then press the number key corresponding to the data list you wish to clear, for example, [2nd] [1] will clear L₁, then press [ENTER]. After you enter the data, press [2nd] [STAT PLOT]. Select the first plot by hitting [Enter] or the number [1:Plot 1]. Turn the plot [On] by moving the cursor to On and selecting Enter. Select the Histogram option using the right arrow keys. Select [Zoom], then [ZoomStat].

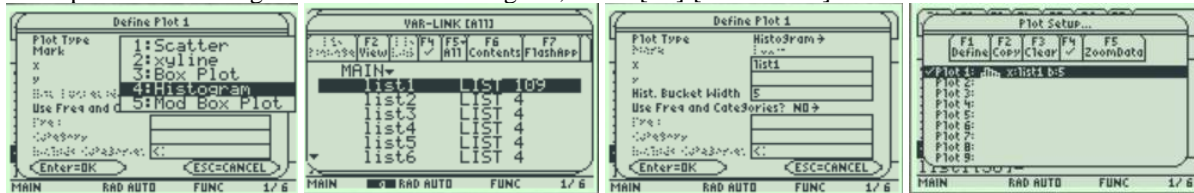


You can see and change the class width by selecting [Window], then change the minimum x value $X_{min}=20$, the maximum x value $X_{max}=50$, the x -scale to $X_{scl}=5$ and the minimum y value $Y_{min}=-6.5$ and the maximum y value to $Y_{max}=14$. Select the [GRAPH] button. We get a similar looking Histogram compared to the stem-and-leaf plot and Excel histogram. Select the [TRACE] button to see the height of each bar and the classes.

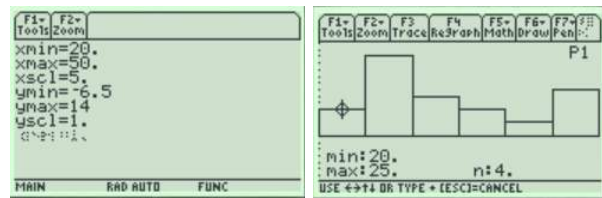
TI-89: First, enter the data into the Stat/List editor under list 1. Press [APP] then scroll down to Stat/List Editor, on the older style TI-89 calculators, go into the Flash/App menu, and then scroll down the list. Make sure the cursor is in the list, not on the list name, and type the desired values pressing [ENTER] after each one. To clear a previously stored list of data values, arrow up to the list name you want to clear, press [CLEAR], and then press enter. After you enter the data, select Press [F2] Plots, scroll down to [1: Plot Setup] and press [Enter].



Select [F1] Define. Use your arrow keys to select Histogram for Type, and then scroll down to the x-variable box. Press [2nd] [Var-Link] this key is above the [+] sign. Then arrow down until you find your List1 name under the Main file folder. Then press [Enter] and this will bring the name List1 back to the menu. You will now see that Plot1 has a small picture of a histogram. To view the histogram, select [F5] [Zoom Data].



The histogram looks a little different from Excel; you can change the settings for the bucket to match your table. Press [♦] [F2:Window]. Change the minimum x value $x_{min}=20$, the maximum x value $x_{max}=50$, the x -scale to $x_{sc1}=5$ and the minimum y value $y_{min}=-6.5$ and the maximum y value to $y_{max}=14$. Then press the [♦] [F3:GRAPH] button. Select [F3:Trace] to see the frequency for each bar. Then use your left and right arrow keys to move to the other bars.



Example 2-12: Make a histogram for the following random sample of student rent prices using Excel.

1500	1350	350	1200	850	900
1500	1150	1500	900	1400	1100
1250	600	610	960	890	1325
900	800	2550	495	1200	690

Solution: Start by making a relative frequency distribution table with 7 classes.

1. Find the range: largest value – smallest value = $2550 - 350 = 2200$, range = \$2,200.
2. Find the class width: $\text{width} = \frac{\text{range}}{7} = \frac{2200}{7} \approx 314.286$. Round up to 315. Always round up to the next integer even if the width is already an integer.
3. Find the class limits: Start at the smallest observation. This is the lower class limit for the first class. Add the class width to get the lower limit of the next class. Keep adding the class width to get all the lower limits, $350 + 315 = 665$, $665 + 315 = 980$, $980 + 315 = 1295$, etc. The upper limit is one unit less than the next lower limit: so, for the first class the upper class limit would be $665 - 1 = 664$. When you have all 7 classes, make sure the last number, in this case the 2550, is at least as large as the largest value in the data. If not, you made a mistake somewhere.

Using Excel: Type the raw data in Excel in column A, the right-hand class endpoints for the bins in column B. Select Data, Data Analysis, Histogram.

Select the Input Range, Bin Range, Labels (if you selected them), output option, Chart Output, then OK. See Figure 2-10.

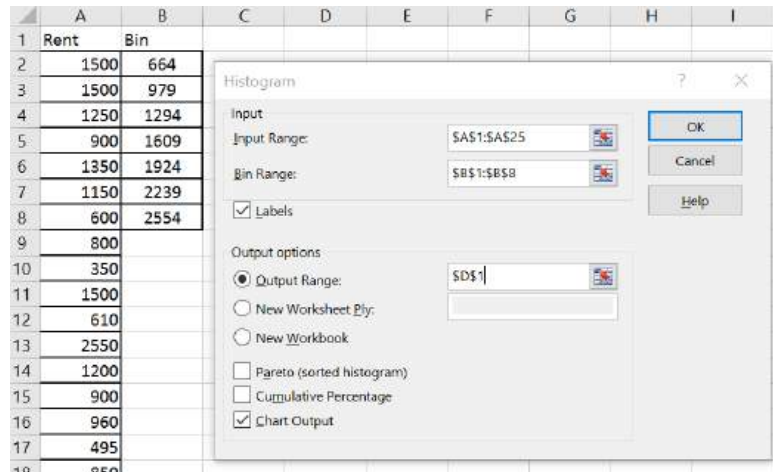


Figure 2-10

See finished histogram below in Figure 2-13.

By hand: Tally and find the frequencies and relative frequencies of the data as shown in Figure 2-11.

Frequency Distribution for Monthly Rent

Class Limits	Tally	Frequency	Relative Frequency
350 – 664		4	0.1667
665 – 979		8	0.3333
980 – 1294		5	0.2083
1295 – 1609		6	0.25
1610 – 1924		0	0
1925 – 2239		0	0
2240 – 2554		1	0.0417
Total		24	1

Figure 2-11

Make sure the total of the frequencies is the same as the number of data points and the total of the relative frequency is one. Since we want the bars on the histogram to touch, the number line needs to use the class boundaries that are half way between the endpoints of the class limits. Start by finding the distance between the class endpoints and divide by two: $(665-664)/2 = 0.5$. Then subtract 0.5 from the left-hand side of each class limit and this will give you the points to use on the x-axis: 349.5, 664.5, 979.5, 1294.5, 1609.5, 1924.5, 2239.5, and 2554.5. Then draw your graph as shown in Figure 2-12. You can use frequencies or relative frequencies for the y-axis as shown in Figure 2-13.

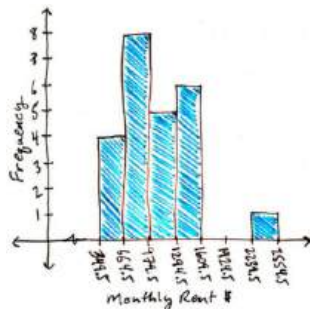


Figure 2-12

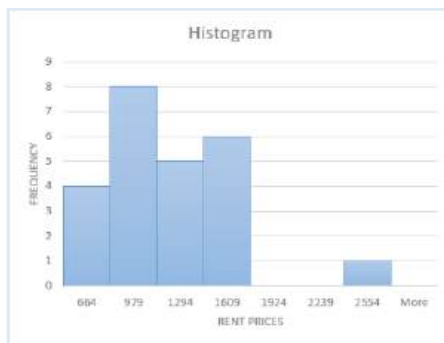
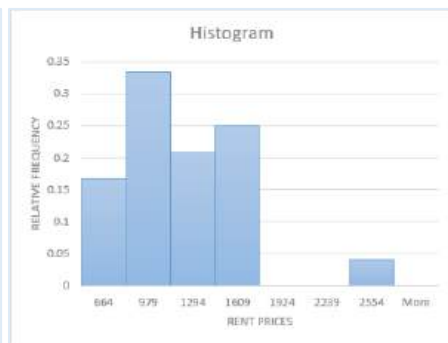


Figure 2-13



Reviewing the graphs in Figure 2-13, you can see that most of the students pay around \$750 per month for rent, with about \$1,500 being the other common value. Most students pay between \$600 and \$1,600 per month for rent. Of course, these values are just estimates pulled from the graph.

There is a large gap between the \$1,500 class and the highest data value. This seems to say that one student is paying a great deal more than everyone else is. This value may be an outlier.

An **outlier** is a data value that is far from the rest of the values. It may be an unusual value or a mistake. It is a data value that should be investigated. In this case, the student lives in a very expensive part of town, thus the value is not a mistake, and is just very unusual. There are other aspects that can be discussed, but first some other concepts need to be introduced.

2.3.3 Ogive

The line graph for the cumulative or cumulative relative frequency is called an **ogive** ([oh-jyve](#)). To create an ogive, first create a scale on both the horizontal and vertical axes that will fit the data. Then plot the points of the upper class boundary versus the cumulative (or cumulative relative) frequency. Make sure you include the point with the lowest class and the zero cumulative frequency. Then just connect the dots.

The steeper the line the more accumulation occurs across the corresponding class. If the line is flat then the frequency for that class is zero. The ogive graph will always be going uphill from left to right and should never dip below the previous point. Figure 2-14 is an example of an ogive.

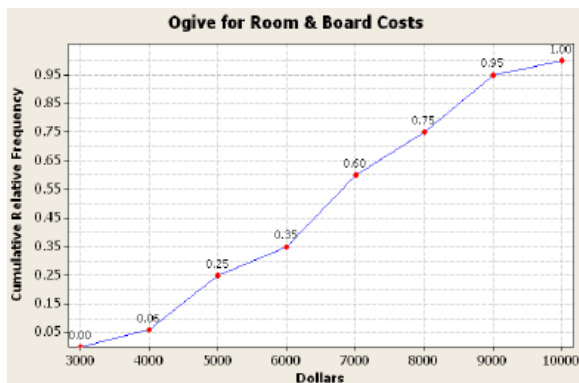


Figure 2-14



Example 2-13: Make an ogive for the following random sample of rent prices students pay with the corresponding cumulative frequency distribution table.

1500	1350	350	1200	850	900	1250	600	610	960	890	1325
1500	1150	1500	900	1400	1100	900	800	2550	495	1200	690

Class Limits	Frequency	Cumulative Frequency
350 – 664	4	4
665 – 979	8	12
980 – 1294	5	17
1295 – 1609	6	23
1610 – 1924	0	23
1925 – 2239	0	23
2240 – 2554	1	24

Solution:

Find the class boundaries, 349.5, 664.5 ... use these for the tick mark labels on the horizontal x-axis, the same as what was used for the histogram. The y-axis uses the cumulative frequencies. The largest cumulative frequency is 24. Every third number is marked on the y-axis units. See Figure 2-15 and Figure 2-16.

By hand:

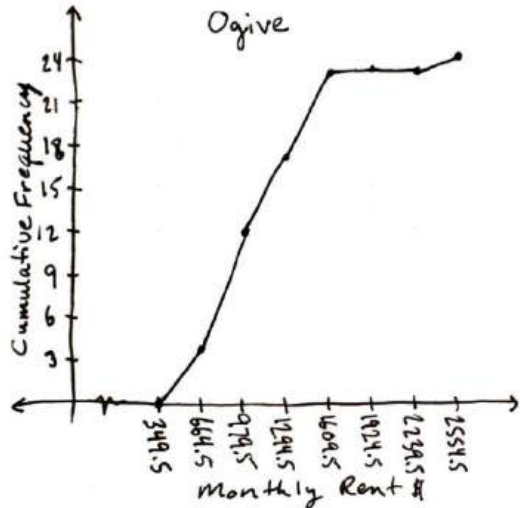


Figure 2-15

Using software:

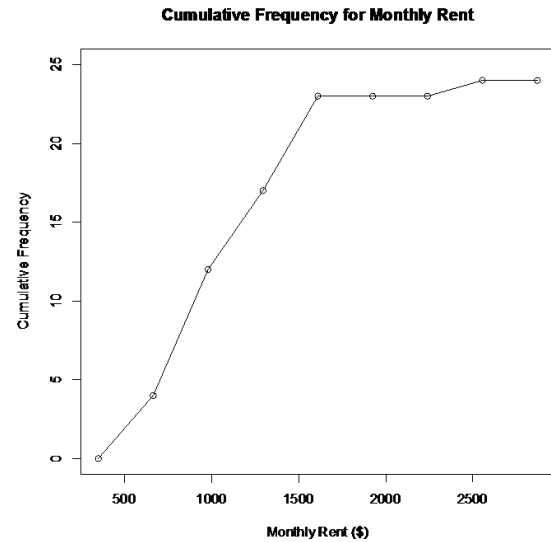


Figure 2-16

The usefulness of an ogive is to allow the reader to find out how many students pay less than a certain value, and what amount of monthly rent a certain number of students pay.

For instance, if you want to know how many students pay less than \$1,500 a month in rent, then you can go up from the \$1,500 until you hit the line and then you go left to the cumulative frequency axis to see what cumulative frequency corresponds to \$1,500. It appears that around 21 students pay less than \$1,500. See Figure 2-17.

If you want to know the cost of rent that 15 students pay less than, then you start at 15 on the vertical axis and then go right to the line and down to the horizontal axis to the monthly rent of about \$1,200. You can see that about 15 students pay less than about \$1,200 a month. See Figure 2-18.

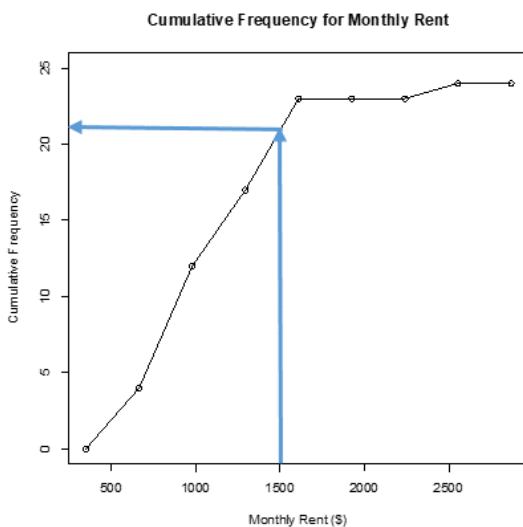


Figure 2-17

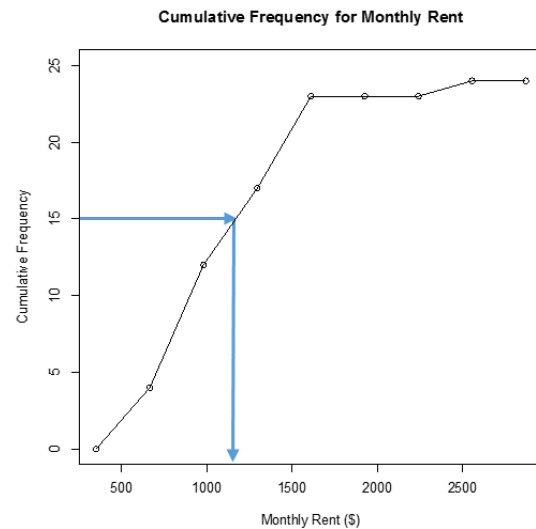


Figure 2-18

If you graph the cumulative relative frequency then you can find out what percentage is below a certain number instead of just the number of people below a certain value.

Example 2-14: Using the sample of 35 ages, make an ogive.

46	47	49	25	46	22	42
24	46	40	39	27	25	30
33	27	46	21	29	20	26
25	25	26	35	49	33	26
32	31	39	30	39	29	26

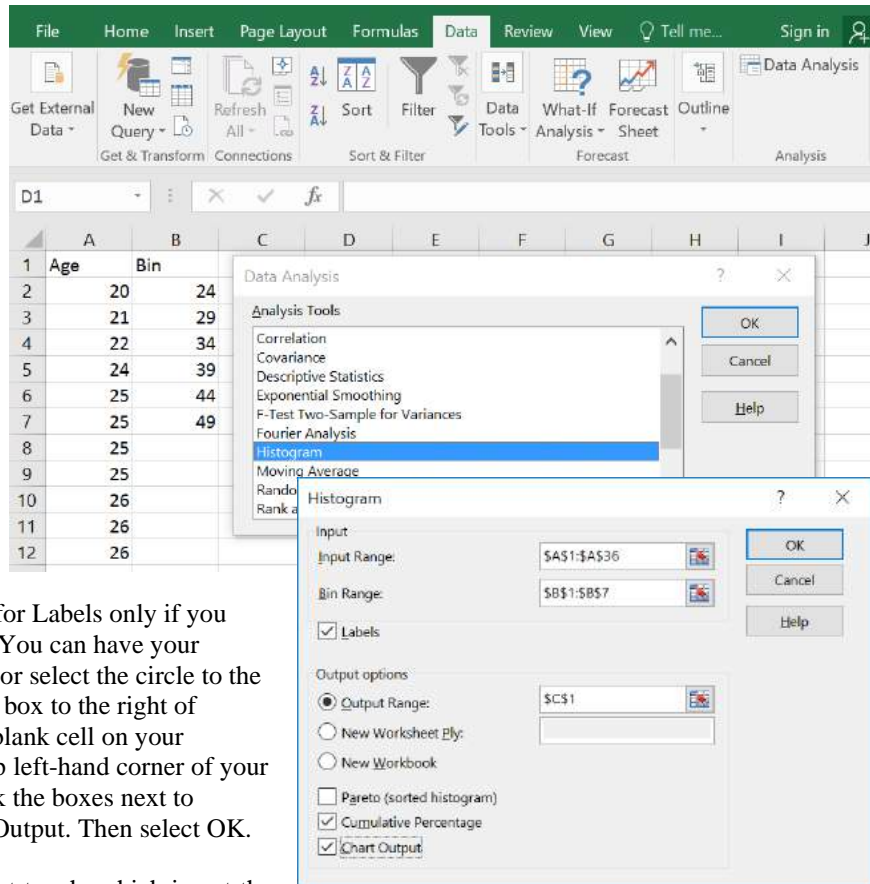
Solution:

Excel

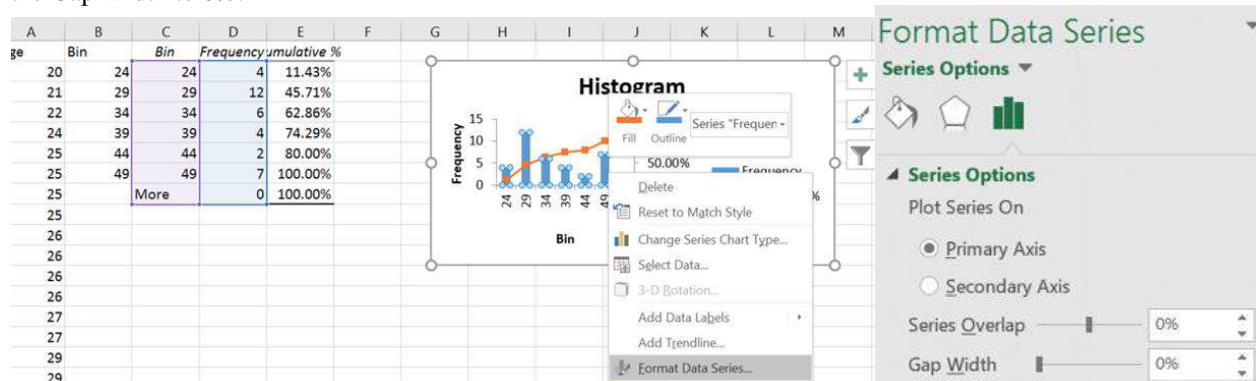
Excel will plot an ogive over a histogram as one of its options, but the scale is harder to read.

Type the data in any order into column A and the bins in order in column B as shown below. Then select the Data tab, select Data Analysis, select Histogram, then select OK, see below.

In the dialogue box, click into the Input Range box, then use your mouse and highlight the ages including the label. Then click into the Bin Range box and use your mouse to highlight the bins including the label. Select the box for Labels only if you included the labels in your ranges. You can have your output default to a new worksheet, or select the circle to the left of Output Range, click into the box to the right of Output Range and then select one blank cell on your spreadsheet where you want the top left-hand corner of your table and graph to start. Then check the boxes next to Cumulative Percentage and Chart Output. Then select OK.



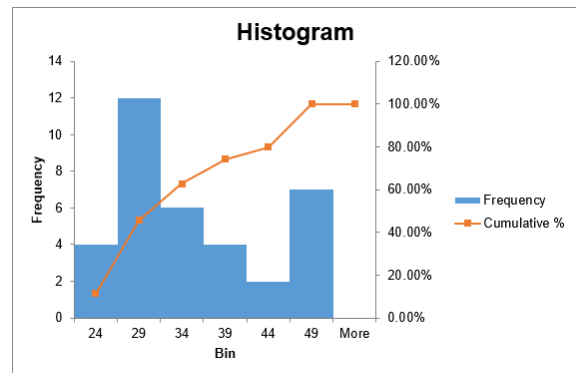
A histogram needs to have bars that touch, which is not the default in Excel. To get the bars to touch, right-click on one of the blue bars and select Format Data Series and slide the Gap Width to 0%.



Excel produces both a frequency table and a histogram. The table has the frequencies and the cumulative relative frequencies.

The orange line is the ogive and the vertical axis is on the right side.

Bin	Frequency	Cumulative %
24	4	11.43%
29	12	45.71%
34	6	62.86%
39	4	74.29%
44	2	80.00%
49	7	100.00%
More	0	100.00%



2.3.4 Pie Chart

You cannot make stem-and-leaf plots, histograms, ogives or time series graphs for qualitative data. Instead, we use bar or pie charts for a qualitative variable, which lists the categories and gives either the frequency (count) or the relative frequency (percent) of individual items that fall into each category.

A **pie chart** or pie graph is a very common and easy-to-construct graph for qualitative data. A pie chart takes a circle and divides the circle into pie shaped wedges that are proportional to the size of the relative frequency. There are 360 degrees in a full circle. Relative frequency is just the percentage as a decimal. To find the angle for each pie wedge, multiply the relative frequency for each category by 360 degrees. Figure 2-19 is an example of a pie chart.

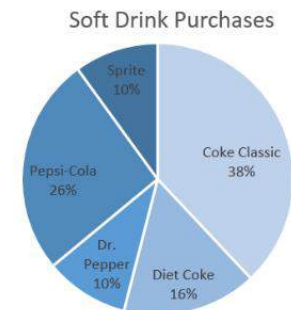


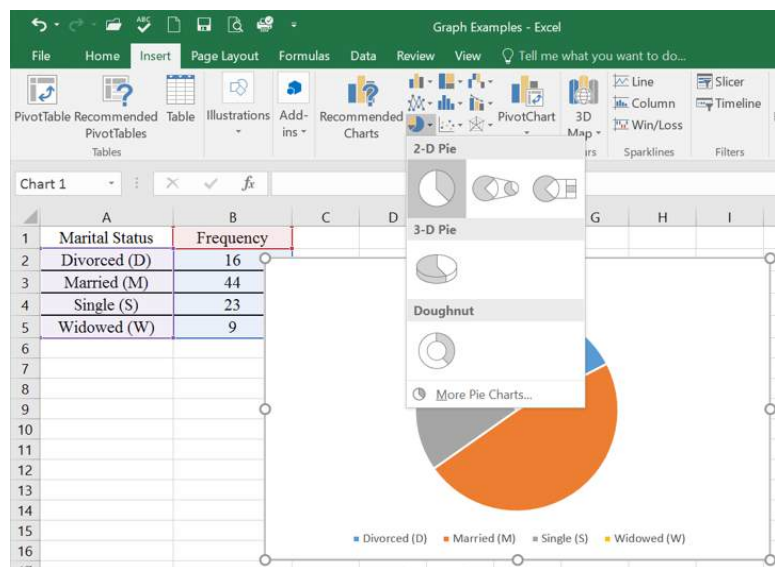
Figure 2-19

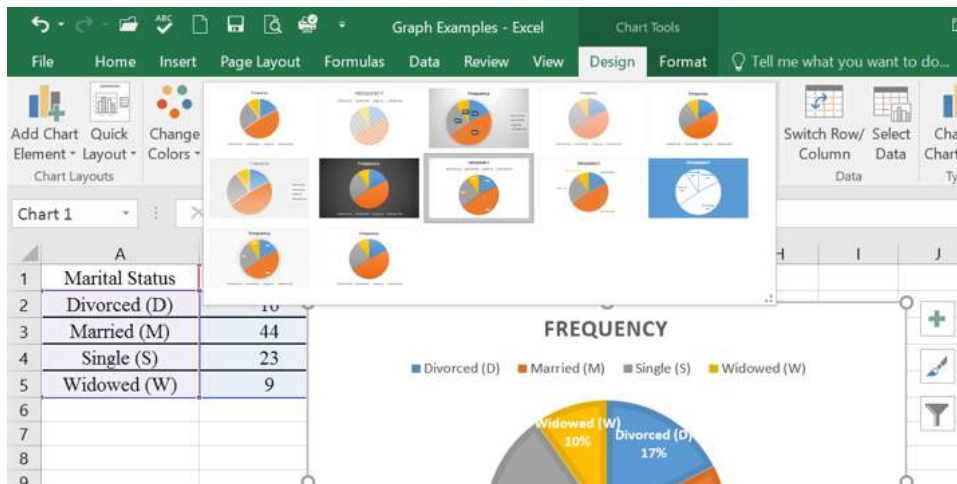
Example 2-15: Use Excel to make a pie chart for the following frequency distribution of marital status.

Marital Status	Frequency
Divorced (D)	16
Married (M)	44
Single (S)	23
Widowed (W)	9

Solution: In Excel, type in the table as it appears, then use your mouse and highlight the entire table. Select the Insert tab, then select the pie graph icon, then select the first option under the 2-D Pie.

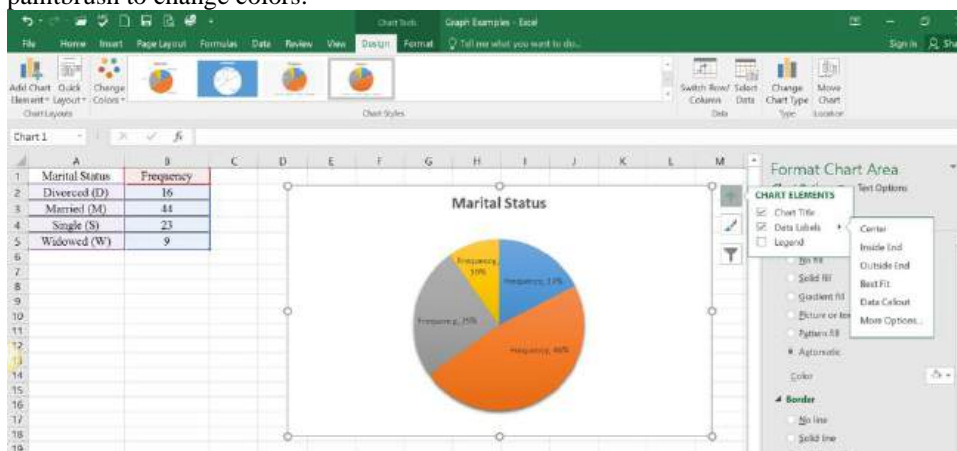
Once you have the pie chart you can select the Design window to get a graph to your liking.



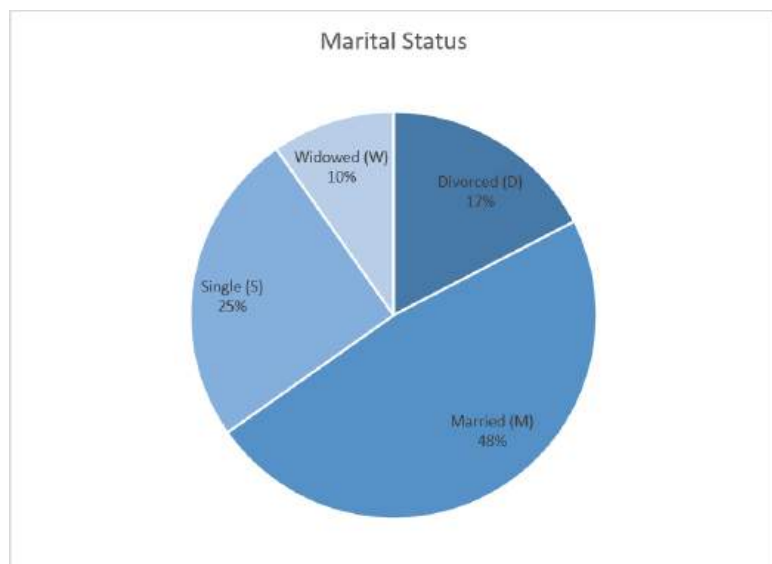


It is good practice to include the class label and the percent. The percent should add up to 100%, although with rounding sometimes the sum can be off by 1%.

You can also click on the green plus sign to the right of the graph and add different formatting options, or the paintbrush to change colors.



Here is the finished pie graph.



2.3.5 Bar Graph

A bar graph (column graph or bar chart) is another graph of a distribution for qualitative data. **Bar graphs or charts** consist of frequencies on one axis and categories on the other axis. Then you draw rectangles for each category with a height (if frequency is on the vertical axis) or length (if frequency is on the horizontal axis) that is equal to the frequency. All of the rectangles should be the same width, and there should be equally wide gaps between each bar. Figure 2-20 is an example of a bar chart.

Some key features of a bar graph:

- Equal spacing on each axis
- Bars are the same width
- Label each axis and title the graph
- Show the scale on the frequency axis
- Label the categories on the category axis
- The bars do not touch.

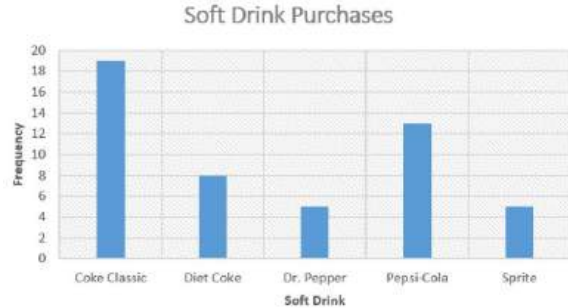


Figure 2-20

You can draw a bar graph with frequency or relative frequency on the vertical axis. The relative frequency is useful when you want to compare two samples with different sample sizes. The relative frequency graph and the frequency graph should look the same, except for the scaling on the frequency axis.

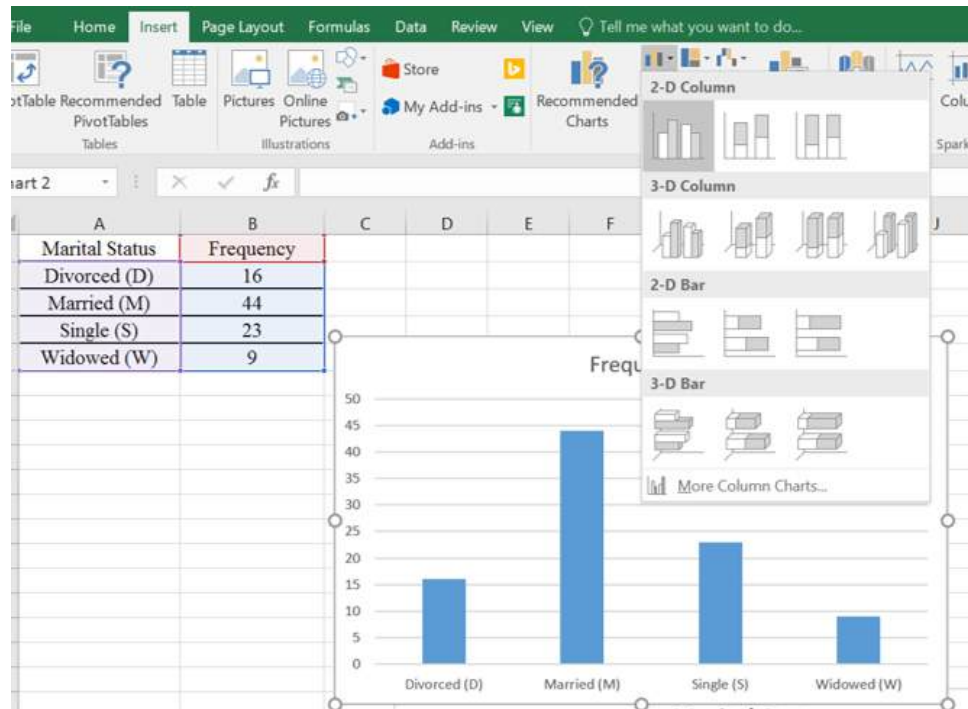
Example 2-16: Use Excel to make a bar chart for the following frequency distribution of marital status.

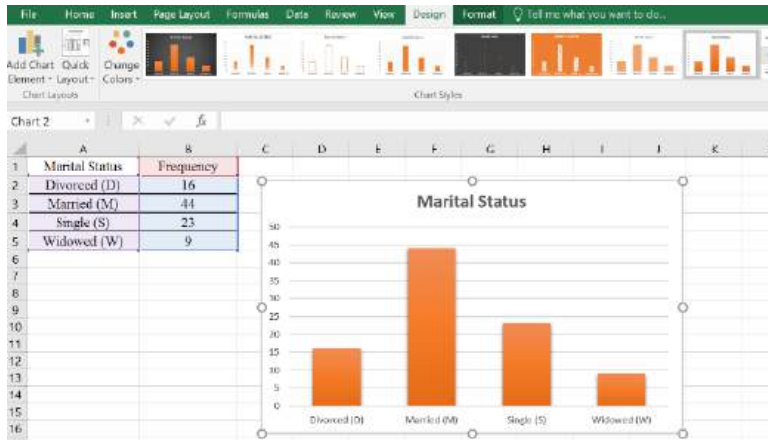
Marital Status	Frequency
Divorced (D)	16
Married (M)	44
Single (S)	23
Widowed (W)	9

Solution: In Excel, type in the table as it appears, then use your mouse and highlight the entire table.

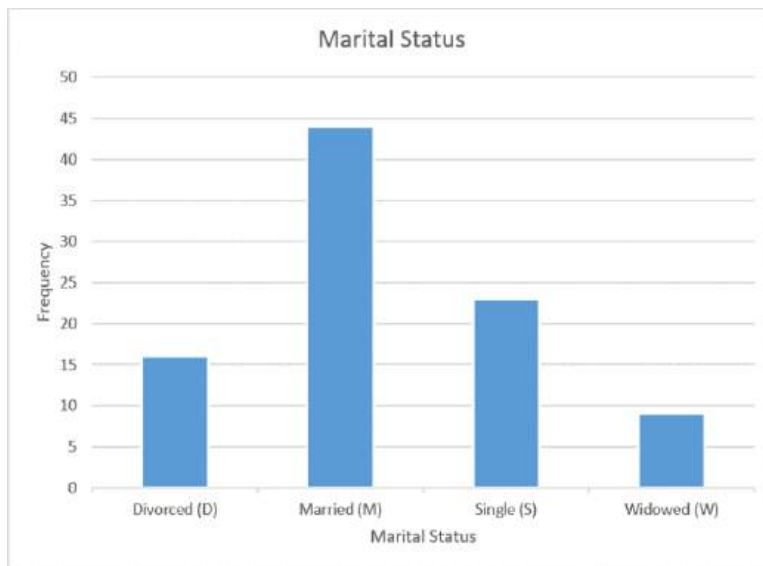
Similar steps as the pie chart, but this time choose the column graph option we get the following bar graph for marital status.

Then format the graph as needed.





The completed bar graph is below.



Pie charts are useful for comparing sizes of categories. Bar charts show similar information. It really is a personal preference and what information you are trying to address. However, pie charts are best when you only have a few categories and the data can be expressed as a percentage.

The data does not have to be percentages to draw the pie chart, but if a data value can fit into multiple categories, you cannot use a pie chart to display the data. As an example, if you are asking people which is their favorite national park and you ask them to pick their top three choices, then the total number of answers can add up to more than 100% of the people surveyed. Therefore, you cannot use a pie chart to display the favorite national park, but a bar chart would be appropriate.

2.3.6 Pareto Chart

A **Pareto** (pronounced pə-**RAY**-toh) **chart** is a bar graph that starts from the most frequent class to the least frequent class. The advantage of Pareto charts is that you can visually see the more popular answer to the least popular. This is especially useful in business applications, where you want to know what services your customers like the most, what processes result in more injuries, which issues employees find more important, and other type of questions where you are interested in comparing frequency. Figure 2-21 is an example of a Pareto chart.

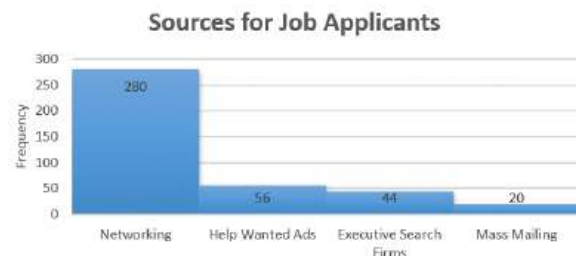
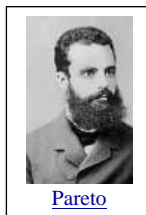
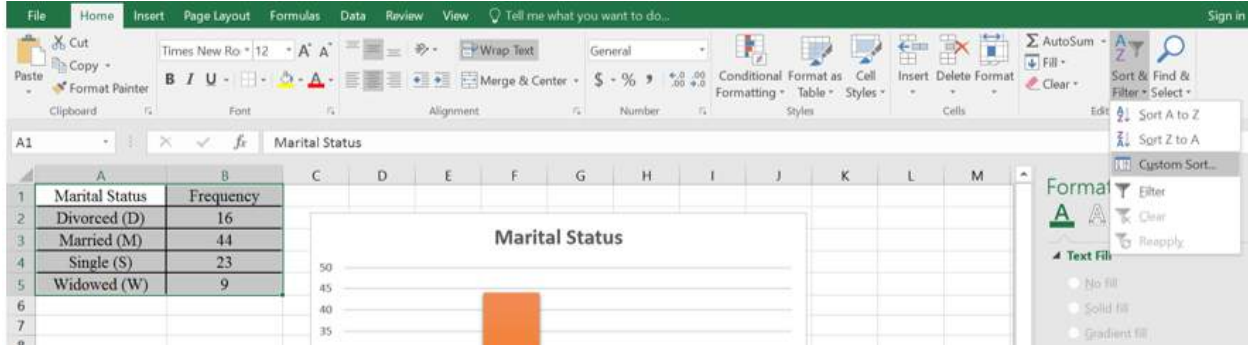


Figure 2-21

Example 2-17: Use Excel to make a Pareto chart for the following frequency distribution of marital status.

Marital Status	Frequency
Divorced (D)	16
Married (M)	44
Single (S)	23
Widowed (W)	9

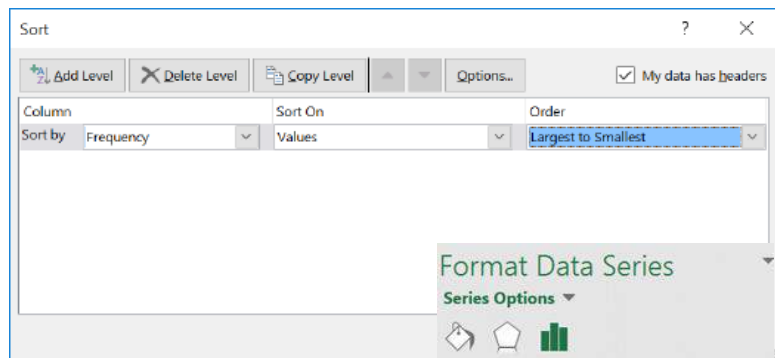
Solution: In Excel, type in the table as it appears, then use your mouse and highlight the entire table. Highlight the table, then select the Home tab, then select Sort & Filter, then select Custom Sort.



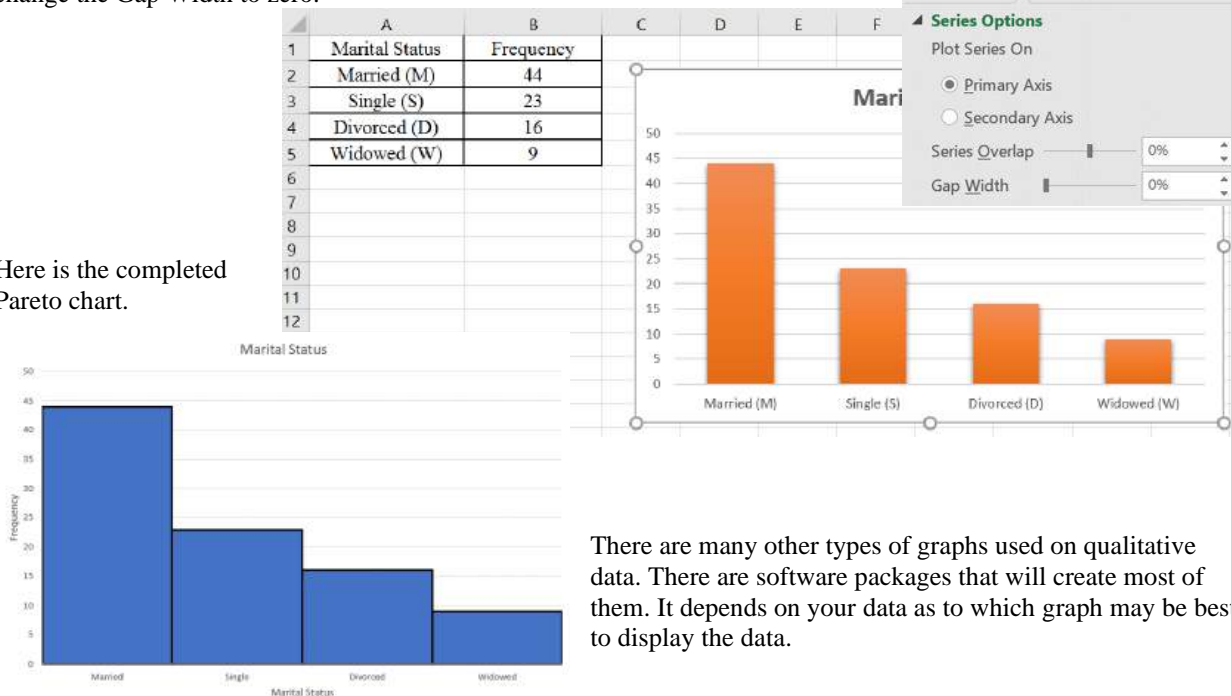
Change the Sort by to Frequency and the Order to Largest to Smallest and click OK.

This will automatically arrange the bars in your bar chart from largest to smallest.

Many Pareto charts will have the bars touching. You can right click on the bars, choose format data series, and then change the Gap Width to zero.



Here is the completed Pareto chart.



There are many other types of graphs used on qualitative data. There are software packages that will create most of them. It depends on your data as to which graph may be best to display the data.

2.3.7 Stacked Column Chart

Both stacked column charts and 100% stacked column charts are types of bar charts that are commonly used to display and compare categorical data. However, there is a significant difference in how the data is represented in these two types of charts. A stacked column chart displays the data as a series of vertical bars, each representing a category, with sub-bars stacked on top of each other representing different subcategories. The height of each bar represents the total value of the subcategories, and the different sub-bars within each category represent the contribution of each subcategory to that total value.

The next example illustrates one of these types known as a stacked column chart. Stacked column (bar) charts are used when we need to show the ratio between a total and its parts. Each color shows the different series as a part of the same single bar, where the entire bar is used as a total.

Example 2-18: In the Wii Fit game, you can do four different types of exercises: yoga, strength, aerobic, and balance. The Wii system keeps track of how many minutes you spend on each of the exercises every day. The following graph is the data for Niko over one-week time-period. Discuss any interpretations you can infer from the graph.

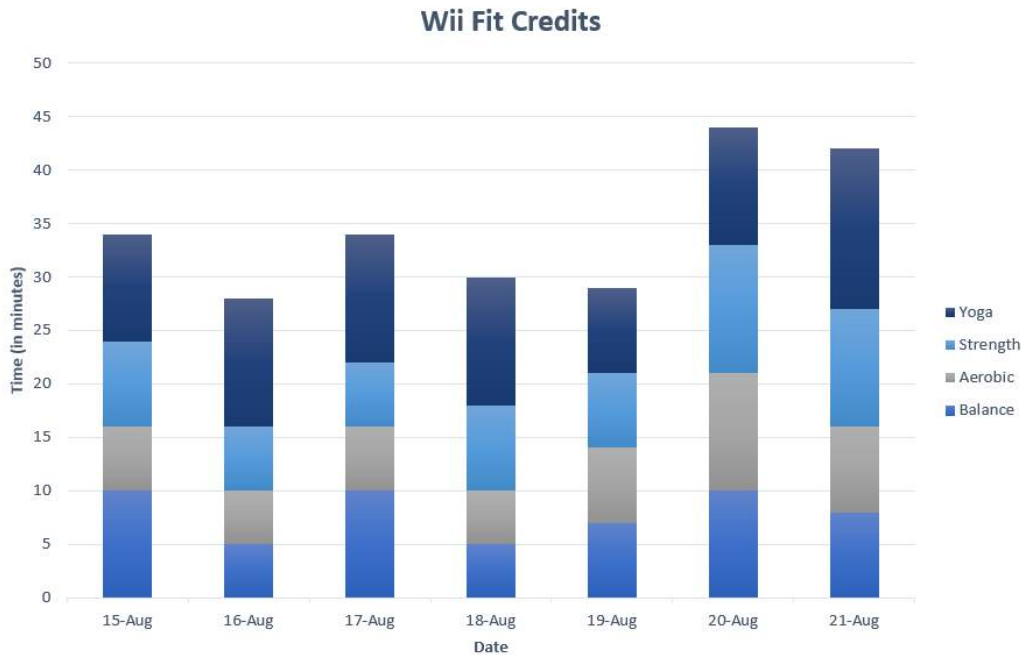


Figure 2-22

Solution: It appears that Niko spends more time on yoga than on any other exercises on any given day. He seems to spend less time on aerobic exercises on a given day. There are several days when the amount of exercise in the different categories is almost equal. The usefulness of a stacked column chart is the ability to compare several different categories over another variable, in this case time. This allows a person to interpret the data with a little more ease.

100% Stacked Column Chart

A 100% stacked column chart also displays the data as a series of vertical bars, with sub-bars stacked on top of each other representing different subcategories. However, in this type of chart, the height of each bar is standardized to be 100%, and each sub-bar within the bar represents the percentage contribution of that subcategory to the total value of the category.

Example 2-19: The data used to make the graph in Figure 2-22 is shown below. Construct a 100% stacked frequency bar graph with the Wii Fit game exercise data.

	15-Aug	16-Aug	17-Aug	18-Aug	19-Aug	20-Aug	21-Aug
Balance	10	5	10	5	7	10	8
Aerobic	6	5	6	5	7	11	8
Strength	8	6	6	8	7	12	11
Yoga	10	12	12	12	8	11	15

Solution: A 100% stacked bar graph shows the percentage of each category relative to its total. The bars are all the same height and each exercise will get its own color. Each day of the week will have a bar with a height of 100% divided by the percentage for each type of exercise. For example, if we take the first day under Balance and divide by the total of each day's total, we would get the relative frequency $10/34 = 0.2941$ or 29.41%. Then do the same for the next 3 exercises. Aerobic is $6/34 = 0.1765$ or 17.65%. Strength is $8/34 = 0.2353$ or 23.53%. Yoga is $10/34 = 0.2941$ or 29.41%. Then the bar for August 15 would be split into 4 sections with different colors for the 29.41%, 17.65%, 23.53%, and 29.41%. Repeat for each day.

	15-Aug	16-Aug	17-Aug	18-Aug	19-Aug	20-Aug	21-Aug
Balance	10	5	10	5	7	10	8
Aerobic	6	5	6	5	7	11	8
Strength	8	6	6	8	7	12	11
Yoga	10	12	12	12	8	11	15
Total	34	28	34	30	29	44	42

Divide each cell by its corresponding column total to get the relative frequency.

	15-Aug	16-Aug	17-Aug	18-Aug	19-Aug	20-Aug	21-Aug
Balance	$10/34 = 0.2941$	$5/28 = 0.1786$	$10/34 = 0.2941$	$5/30 = 0.1667$	$7/29 = 0.2414$	$10/44 = 0.2273$	$8/42 = 0.1905$
Aerobic	$6/34 = 0.1765$	$5/28 = 0.1786$	$6/34 = 0.1765$	$5/30 = 0.1667$	$7/29 = 0.2414$	$11/44 = 0.25$	$8/42 = 0.1905$
Strength	$8/34 = 0.2353$	$6/28 = 0.2143$	$6/34 = 0.1765$	$8/30 = 0.2667$	$7/29 = 0.2414$	$12/44 = 0.2727$	$11/42 = 0.2619$
Yoga	$10/34 = 0.2941$	$12/28 = 0.4286$	$12/34 = 0.3529$	$12/30 = 0.4$	$8/29 = 0.2759$	$11/44 = 0.25$	$15/42 = 0.3571$
Total	$34/34 = 1$	$28/28 = 1$	$34/34 = 1$	$30/30 = 1$	$29/29 = 1$	$44/44 = 1$	$42/42 = 1$

Convert the relative frequencies to percentages.

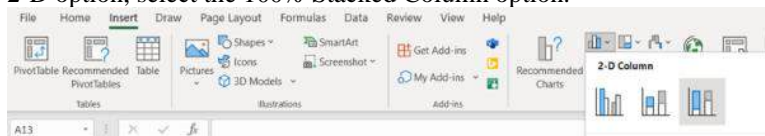
	15-Aug	16-Aug	17-Aug	18-Aug	19-Aug	20-Aug	21-Aug
Balance	29.41%	17.86%	29.41%	16.67%	24.14%	22.73%	19.05%
Aerobic	17.65%	17.86%	17.65%	16.67%	24.14%	25.00%	19.05%
Strength	23.53%	21.43%	17.65%	26.67%	24.14%	27.27%	26.19%
Yoga	29.41%	42.86%	35.29%	40.00%	27.59%	25.00%	35.71%
Total	100%	100%	100%	100%	100%	100%	100%

Make a bar graph, where each day has a bar with a height of 100%. Within each bar, use a different color for each type of exercise and divide the bar up into their corresponding percentages. See Figure 2-23 for the finished 100% stacked column graph. The height of each color in Figure 2-23 represents the percentage of time spent each day for each of the four exercises. The height of the dark blue yoga section is generally longest throughout the week.



Figure 2-23

This was done in Excel by highlighting the original frequency table. Select the Insert tab, then under the bar graph, 2-D option, select the 100% Stacked Column option.

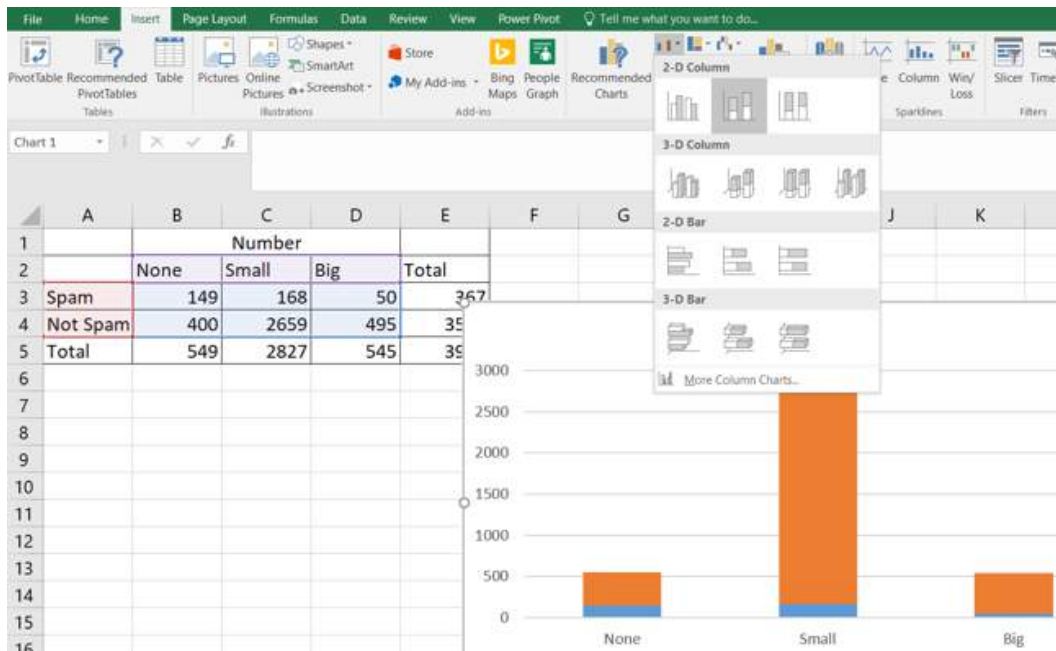


Example 2-20: Data scientists write programming using statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One of those characteristics is whether the email contains no numbers, small numbers, or big numbers. Make a stacked column chart with the data in the table. Which type of email is more likely to be spam?

	Number			Total
	None	Small	Big	
Spam	149	168	50	367
Not Spam	400	2659	495	3554
Total	549	2827	545	3921

Example from OpenIntroStatistics.com.

Solution: Type the summarized table into Excel. Highlight just the inside of the table from the row label, column label and data (do not include the totals or Number label). Select the Insert tab, and then select the 2nd option under the column chart. Add a legend, labels and change colors for clarity.



The completed stacked bar graph is shown in Figure 2-24.

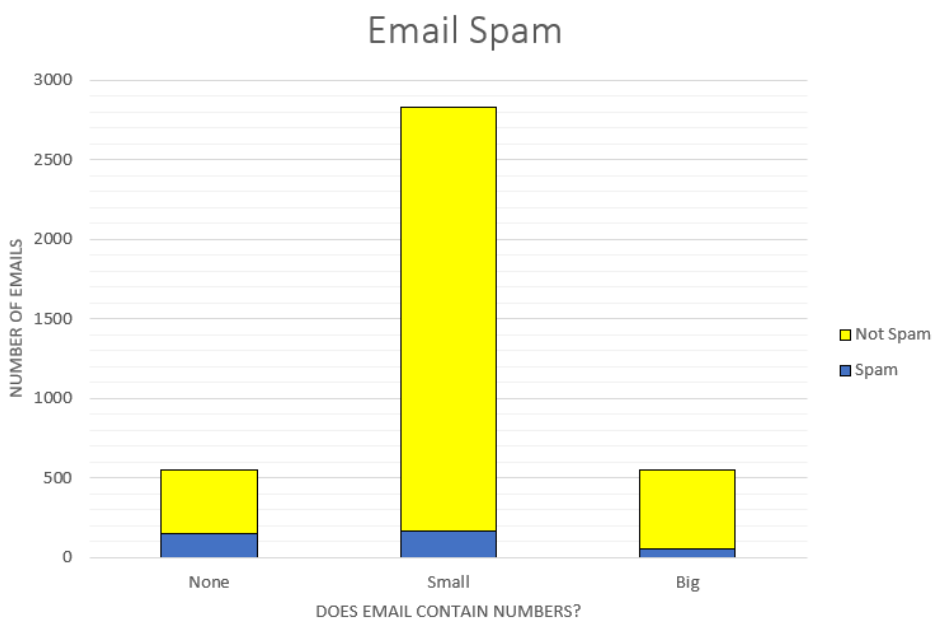


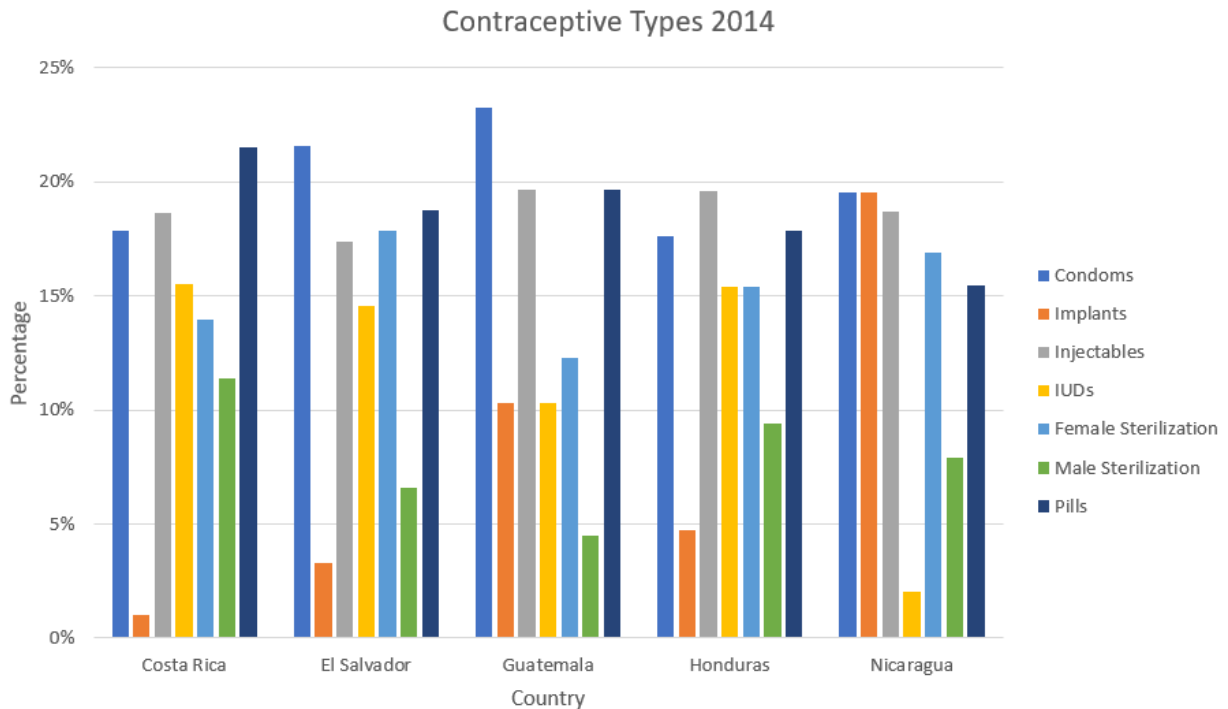
Figure 2-24

Emails with no numbers have a relatively high rate of spam email ($149/549 = 0.271$) about 27%. On the other hand, less than 10% of email with small numbers ($168/2827 = 0.059$) or big numbers ($50/545 = 0.092$) are spam.

2.3.8 Multiple or Side-by-Side Bar Graph

A multiple bar graph, also called a side-by-side bar graph, allows comparisons of several different categories over another variable.

Example 2-21: The percentages of people who use certain contraceptives in Central American countries are displayed in the graph below. Use the graph in Figure 2-25 to find the type of contraceptive that is most used in Costa Rica and El Salvador.



(9/21/2020) Retrieved from <https://public.tableau.com/profile/prbdata#!/vizhome/AccessstoContraceptiveMethods/AccessstoContraceptiveMethods>
Figure 2-25

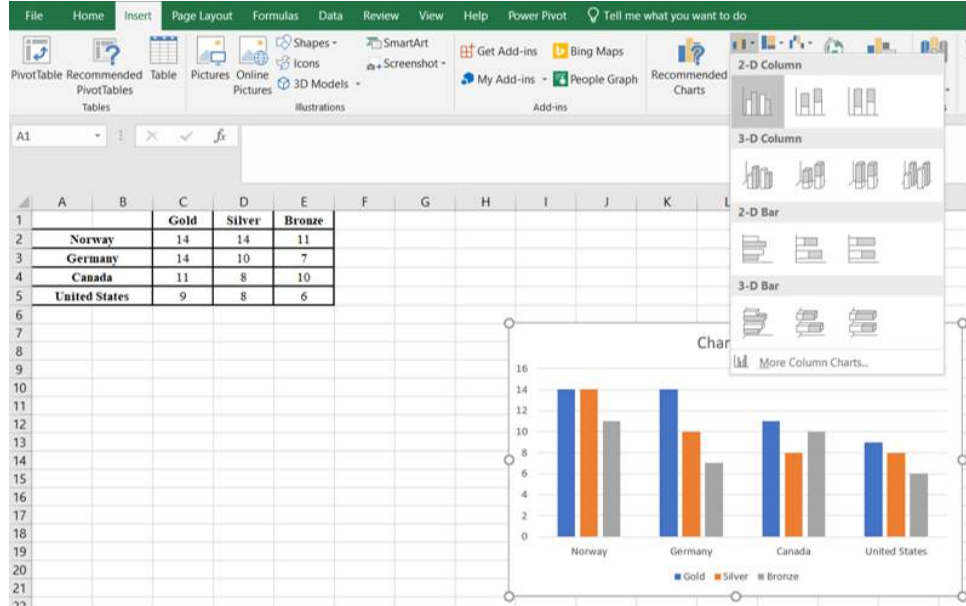
Solution: This side-by-side bar graph allows you to quickly see the differences between the countries. For instance, the birth control pill is used most often in Costa Rica, while condoms are most used in El Salvador.

Example 2-22: Make a side-by-side bar graph for the following medal count for the 2018 Olympics.

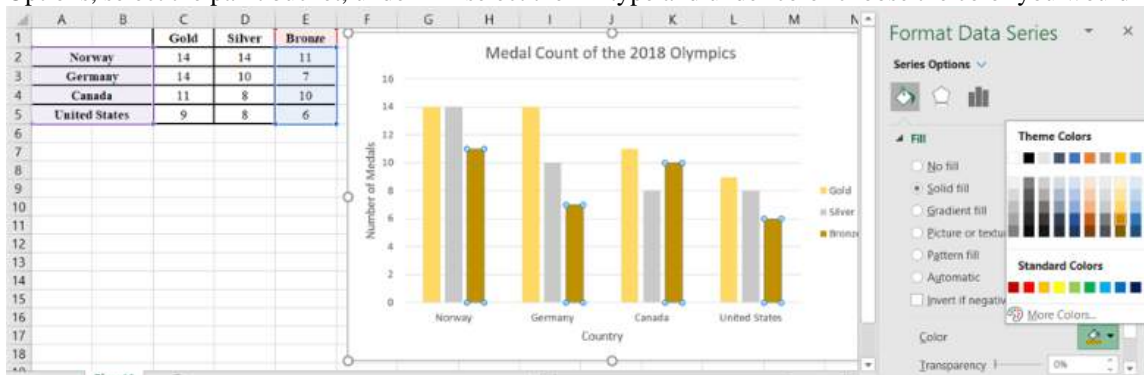
	Gold	Silver	Bronze
Norway	14	14	11
Germany	14	10	7
Canada	11	8	10
United States	9	8	6

Solution: Copy the table over to Excel. Highlight the entire table, then use similar steps as the regular bar graph.

Add labels and change the color.



If you want to customize the colors of the bars, select the graph, then click on the bar you want to change the color for. All three bars should highlight, if not select a different bar again. Select Format Data Series. Under the Series Options, select the paint bucket, under Fill select the fill type and under color choose the color you would like.



Each of the different colors in Figure 2-26 has a legend so that you know which color goes with each medal type.

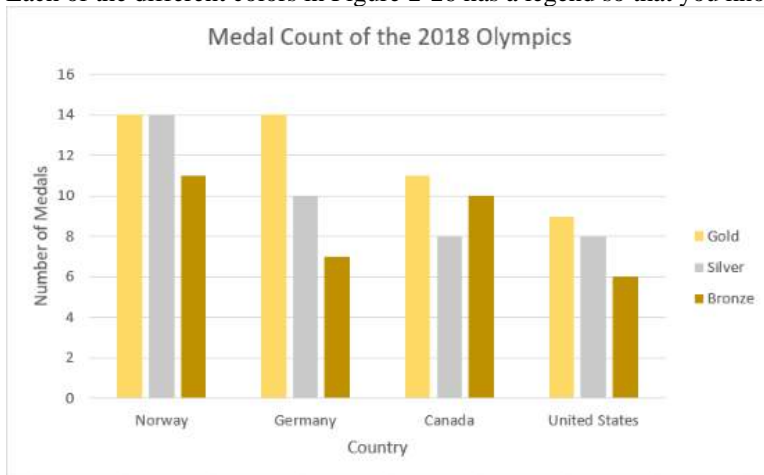


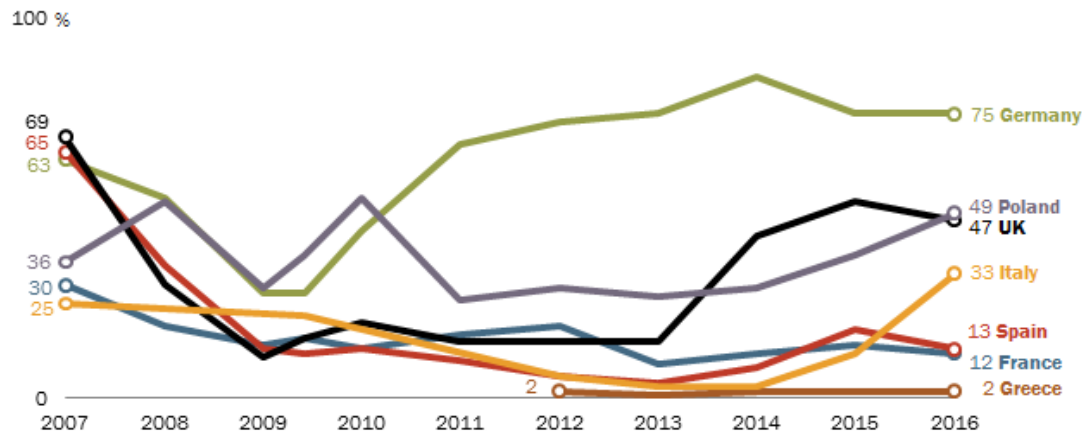
Figure 2-26

2.3.9 Time-Series Plot

A **time-series** plot is a graph showing the data measurements in chronological order, where the data is quantitative data. For example, in Figure 2-27, a time-series plot is used to show the proportion of survey respondents by country that responded that the current economic situation in their country was good over a 9 year period. To create a time-series plot, time always goes on the horizontal axis, and the frequency or relative frequency goes on the vertical axis. Then plot the ordered pairs and connect the dots. A time series allows you to see trends over time. Caution: You must realize that the trend may not continue. Just because you see an increase does not mean the increase will continue forever. As an example, prior to 2007, many people noticed that housing prices were increasing. The belief at the time was that housing prices would continue to increase. However, the housing bubble burst in 2007, and many houses lost value during the recession.

Some European publics view economy on the rebound, but others remain negative

The current economic situation in our country is good



Source: Spring 2016 Global Attitudes Survey, Q3.

PEW RESEARCH CENTER

Figure 2-27

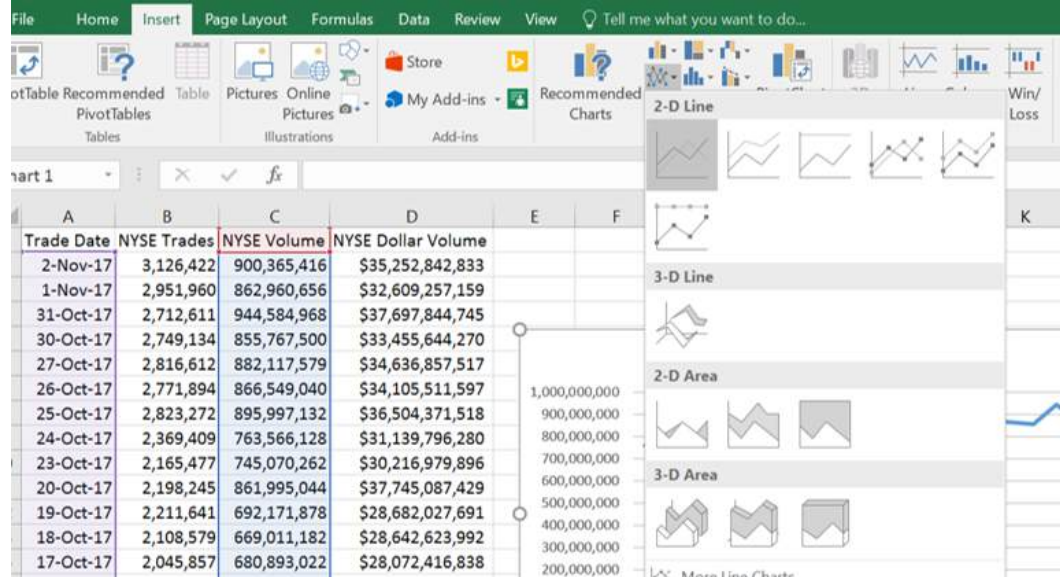
Example 2-23: The New York Stock Exchange (NYSE) has a website where you can download information on the stock market. Use technology to make a time-series plot.

The daily trading volume for two weeks was downloaded at <http://www.nyxdata.com/Data-Products/NYSE-Volume-Summary#summaries>.

Trade Date	NYSE Trades	NYSE Volume	NYSE Dollar Volume
2-Nov-17	3,126,422	900,365,416	\$35,252,842,833
1-Nov-17	2,951,960	862,960,656	\$32,609,257,159
31-Oct-17	2,712,611	944,584,968	\$37,697,844,745
30-Oct-17	2,749,134	855,767,500	\$33,455,644,270
27-Oct-17	2,816,612	882,117,579	\$34,636,857,517
26-Oct-17	2,771,894	866,549,040	\$34,105,511,597
25-Oct-17	2,823,272	895,997,132	\$36,504,371,518
24-Oct-17	2,369,409	763,566,128	\$31,139,796,280
23-Oct-17	2,165,477	745,070,262	\$30,216,979,896
20-Oct-17	2,198,245	861,995,044	\$37,745,087,429
19-Oct-17	2,211,641	692,171,878	\$28,682,027,691
18-Oct-17	2,108,579	669,011,182	\$28,642,623,992
17-Oct-17	2,045,857	680,893,022	\$28,072,416,838
16-Oct-17	2,078,792	685,406,032	\$27,524,409,199
13-Oct-17	2,151,643	757,155,836	\$30,624,653,749

Solution:

Using Excel, we will make a time series plot for NYSE daily trading volume. Using the Ctrl key highlight just the date column and the NYSE Volume, then select the Insert tab and the first 2-D line graph option.



You can then select different designs.



The time-series graph shows the behavior of one variable over time and does not reflect other variables that are influencing the trading volume. One can use time-series plots to see when they want to cash out or buy a stock. Figure 2-28 shows the completed time series graph. The jagged blue line shows the loss and gain in trade volume over time.

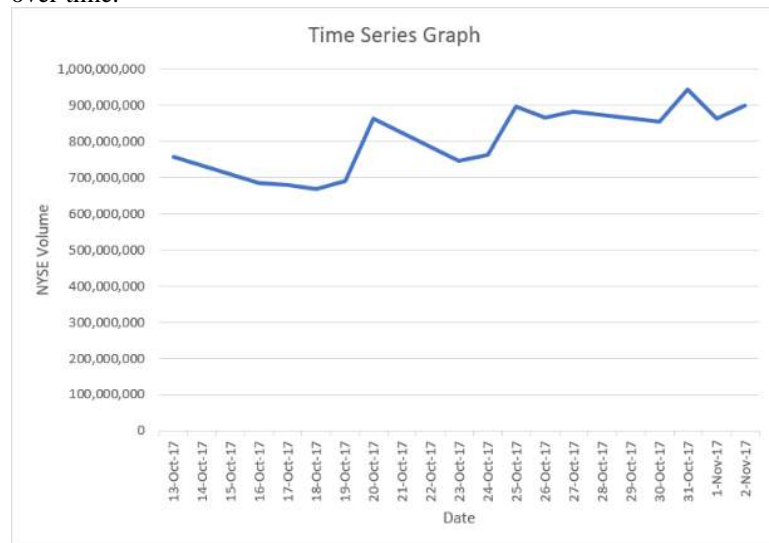


Figure 2-28

2.3.10 Scatter Plot

Sometimes you have two quantitative variables and you want to see if they are related in any way. A scatter plot helps you to see what the relationship may look like. A scatter plot is just a plotting of the ordered pairs.

- When you see the dots increasing from left to right then there is a positive relationship between the two quantitative variables.
- If the dots are decreasing from left to right then there is a negative relationship.
- If there is no apparent pattern going up or down, then we say there is no relationship between the two variables.

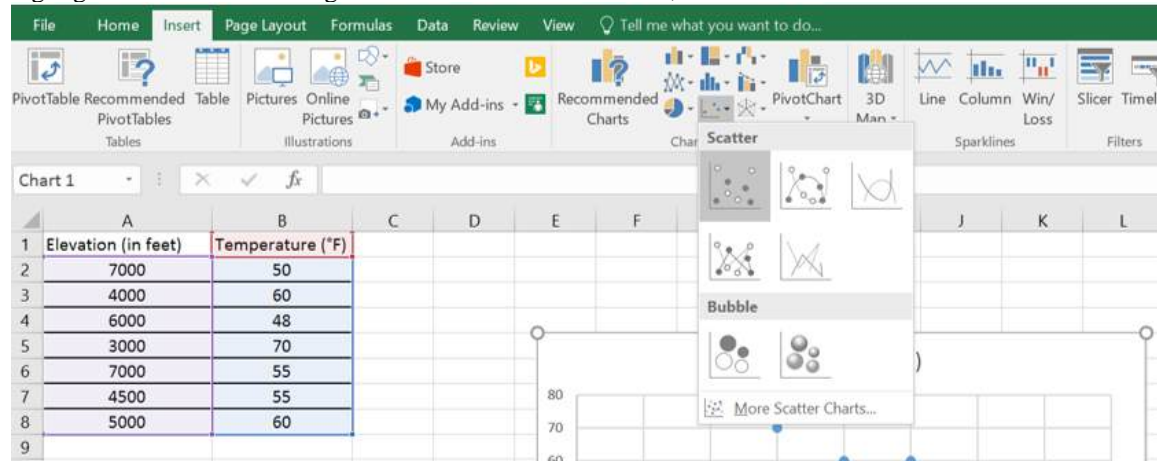
Is there any relationship between elevation and high temperature on a given day? The following data are the high temperatures at various cities on a single day and the elevation of the city.

Example 2-24: Make a scatterplot to see what type of relationship exists.

Elevation (in feet)	7000	4000	6000	3000	7000	4500	5000
Temperature (°F)	50	60	48	70	55	55	60

Solution:

Excel: Type the data into two columns next to each other. It is important not to have a blank column between the points or Excel may give you an error message. Once you type your data into columns A and B, use your mouse and highlight all the data including the labels. Select the Insert tab, and then select the first box under Scatter.



Add appropriate labels. The completed scatter plot is shown below in Figure 2-29.

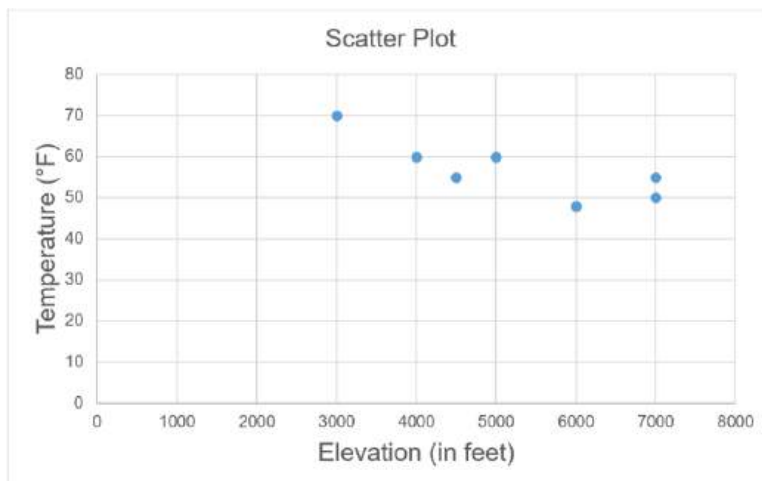
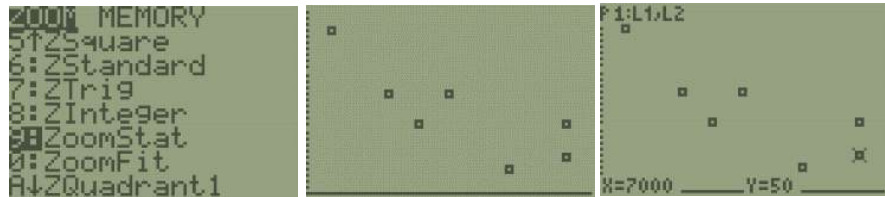


Figure 2-29

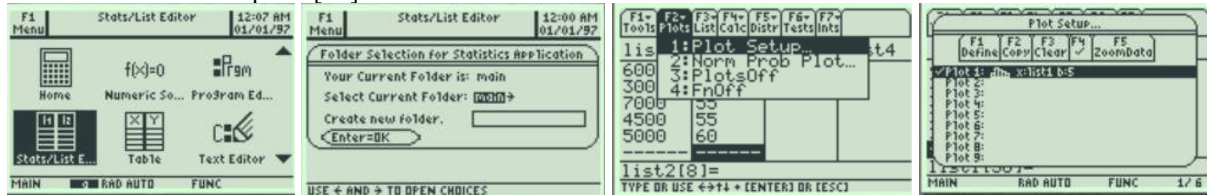
TI-84: First, enter the data into lists 1 and 2. Press [STAT] the first option is already highlighted (1:Edit) so you can either press [ENTER] or 1.Type in the data pressing [ENTER] after each one. For x-y data pairs, enter all x-values in one list. Enter all corresponding y-values in a second list. Press [2nd] [QUIT] to return to the home screen. Make sure you turn off other stat plots or graphs in the y= menu. Press [2nd] then the [y=] button. Select the first plot. Highlight On and press [Enter] so that On is highlighted. Arrow down to Type and highlight the first option that looks like a scatter plot. Make sure your x and y lists are using L₁ and L₂.



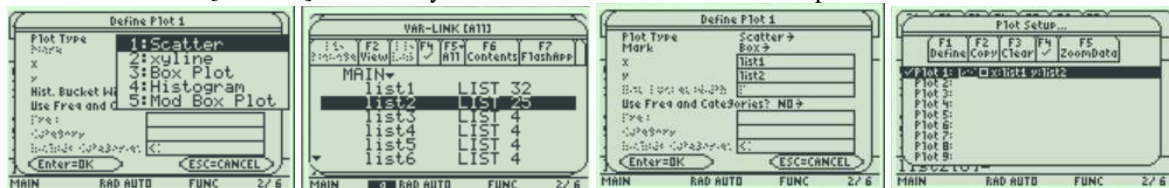
Select Zoom and arrow down to ZoomStat and press [Enter]. You will get the following scatterplot. Select Trace and use your arrow keys to see the values at different points.



TI-89: Press [♦] then [F1] (to get Y=) and clear any equations that are in the y-editor. Open the Stats/List Editor. Press [APPS], select **FlashApps** then press [ENTER]. Highlight **Stats/List Editor** then press [ENTER]. Press [ENTER] again to select the main folder. Type in the data pressing [ENTER] after each one. Enter all x-values in one list. Enter all corresponding y-values in a second list. In the Stats/List Editor, select [F2] for the Plots menu. Use cursor keys to highlight 1:Plot Setup. Make sure that the other graphs are turned off by pressing [F4] button to remove the check marks. Under “Plot 1” press [F1] Define.



In the “Plot Type” menu, select “Scatter.” Move the cursor to the “x” space press [2nd] Var-Link, scroll down to list1, and then press [Enter]. This will put the list name in the dialogue box. Do the same for the y values, but this time choose list2. Press [ENTER] twice and you will be returned to the Plot Setup menu.



Press F5 ZoomData to display the graph.

Press F3 Trace and use the arrow keys to scroll along the different points.



Interpreting the Scatter Plot

The graph in Figure 2-30 indicates a linear relationship between temperature and elevation. If you were to hold a pencil up to cover the dots, note that you would see that the dots roughly follow a fat line downhill.

The scatterplot in Figure 2-30 also appears to be a negative relationship, thus as elevation increases, the temperature decreases.

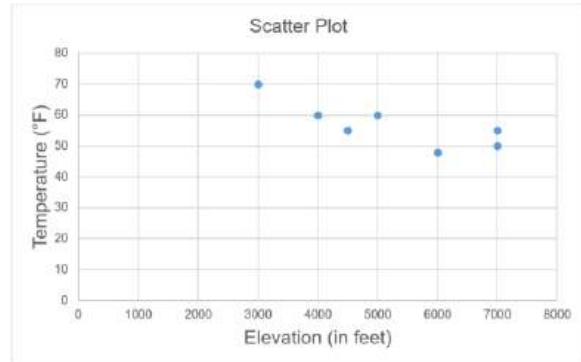


Figure 2-30

Be careful with the vertical axis of both time-series and scatter plots. If the axis does not start at zero the slope of the line can be exaggerated to show more or less of increase than there really is. This is done in politics and advertising to manipulate the data.

For example, if we change the vertical axis of temperature to go between 45°F and 75°F we get the following scatter plot in Figure 2-31.

We have the same arrangements of dots, but the slope looks much steeper over the 30° range.

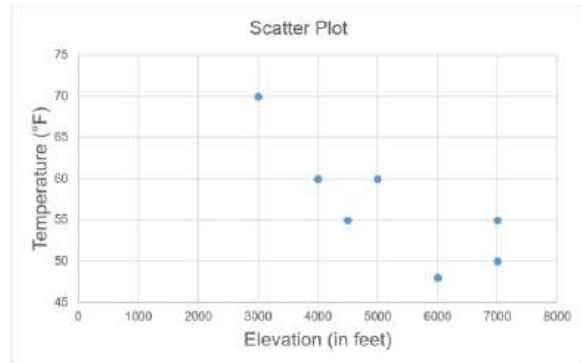


Figure 2-31

2.3.11 Misleading Graphs

One thing to be aware of as a consumer, data in the media may be represented in misleading graphs. Misleading graphs not only misrepresent the data, they can lead the reader to false conclusions. There are many ways that graphs can be misleading. One way to mislead is to use picture graphs or 3D graphs that exaggerate differences and should be used with caution. Leaving off units and labels can result in a misleading graph. Another more common example is to rescale or reverse the vertical axis to try to show a large difference between categories. Not starting the vertical axes at zero will show a more dramatic rate of change. Other ways that graphs can be misleading is to change the horizontal axis labels so that they are out of time sequence, using inappropriate graphs, not showing the base population.

Here are some key factors to consider when examining graphical displays to avoid being misled by misleading graphs.

- **Misleading Bar or Column Chart:** A bar chart can be misleading if it distorts the scale or selectively presents data to create a false impression. This can include using truncated or extended axes, non-zero baselines, inappropriate labeling, or using pictures for bars that distorts the relative proportions of the bars.
- **Misleading Pie Chart:** A pie chart can be misleading if it misrepresents the proportions of the categories by manipulating the size of the slices or using inadequate labeling, or having the total of all category percentages not add up to 100%.
- **Misleading Line Graph:** The line graph may be misleading if it doesn't represent the data accurately, such as using inconsistent scales, omitting data points, or altering the axes to exaggerate or downplay trends.
- **Misleading Scatter Plot:** A scatter plot can be misleading if the data points are manipulated or misrepresented to create a false perception of correlation or lack thereof. This can involve selectively choosing data points, altering scales, inadequate labeling, omitting relevant information, or altering the axes to exaggerate or downplay trends.

Example 2-25: An ad for a new diet pill shows the following time-series plot for someone that has lost weight over a 5-month period shown in Figure 2-32. What is misleading about the graph?

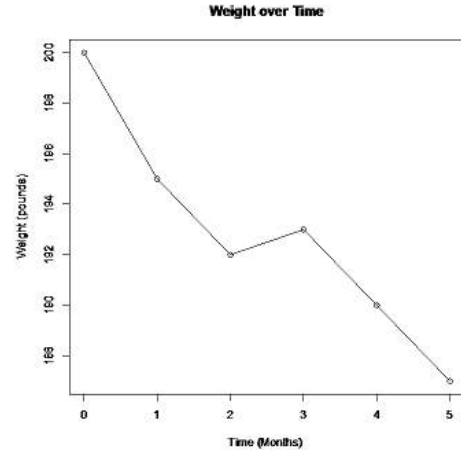


Figure 2-32

Solution:

If you do not start the vertical axis at zero, then a change can look much more dramatic than it really is. Notice the decrease in weight looks much larger in Figure 2-33. The graph in Figure 2-34 has the vertical axis starting at zero. Notice that over the 5 months, the weight appears to be decreasing, however, it does not look like there is a large decrease.

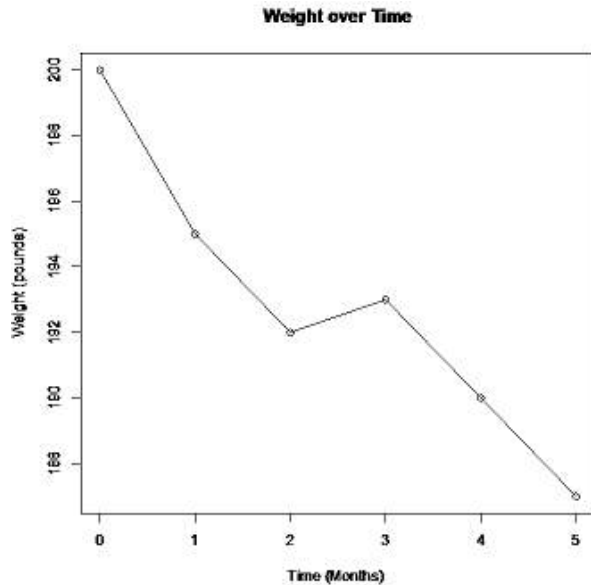


Figure 2-33

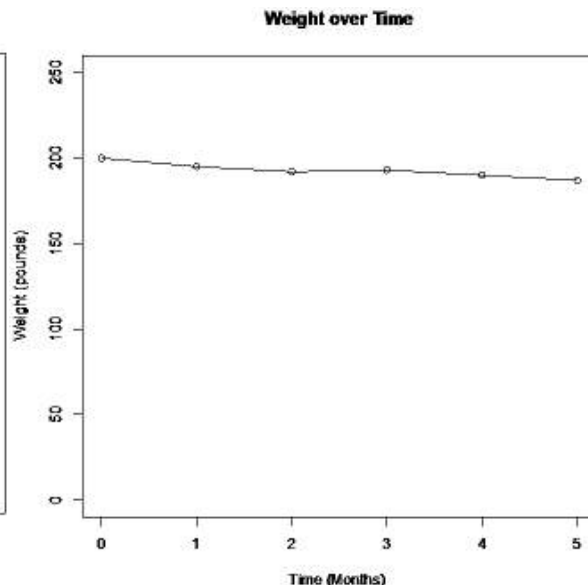


Figure 2-34

Example 2-26: What is misleading about the graph in Figure 2-35?



Figure 2-35

<https://www.mediamatters.org/blog/2014/03/31/dishonest-fox-charts-obamacare-enrollment-editi/198679>.

Solution:

The y-axis scale is different for each bar and there are no units on the axis. The first bar has each tic mark as 2 billion, the second bar has each tick as less than 1 billion.

This exaggerates the difference. If they used square scaling as in Figure 2-36, there would not be such an extreme difference between the height of the bars.

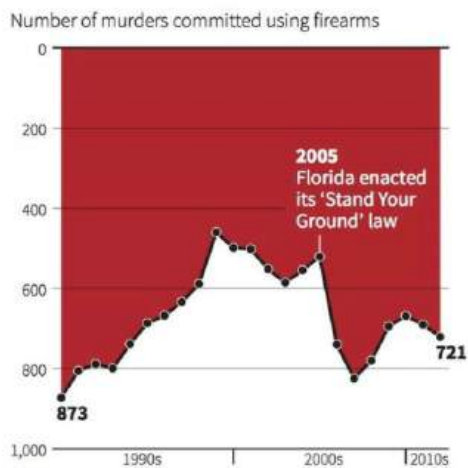


<https://www.mediamatters.org/blog/2014/03/31/dishonest-fox-charts-obamacare-enrollment-edited/198679>

Figure 2-36

Example 2-27: What is misleading about the graph in Figure 2-37?

Gun deaths in Florida



Source: Florida Department of Law Enforcement

<https://www.livescience.com/45083-misleading-gun-death-chart.html>

Figure 2-37

Solution: The graph has the y-axis reversed. What looks like an increasing trend line really is decreasing when you correct the y-axis. The red background is also an effect to raise alarm, almost like a curtain of blood.

Example 2-28: What is misleading about the graph shown in a Lanacane commercial, shown in Figure 2-38?

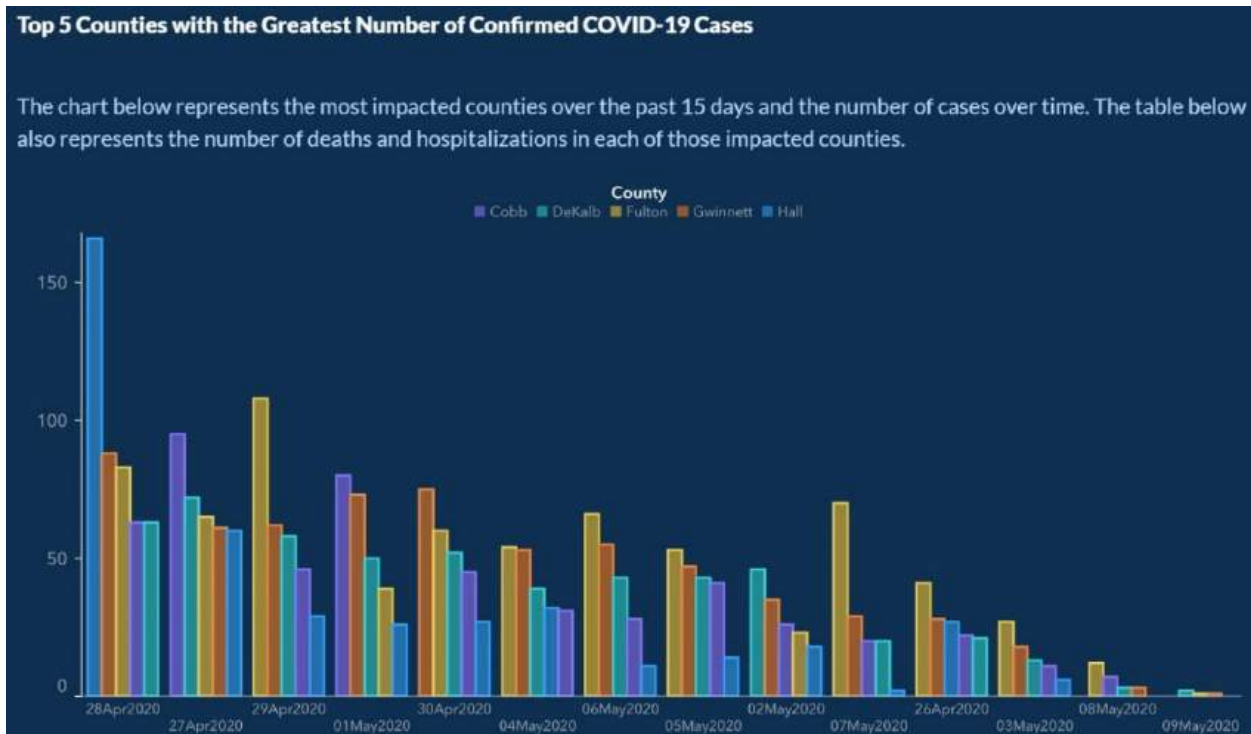


Retrieved 7/2/2021 from <https://youtu.be/IODapkQ-c1I?t=17>

Figure 2-38

Solution: It appears that Lanacane is better than regular hydrocortisone cream at relieving itching. However, note that there are no units or labels to the axis.

Example 2-29: What is misleading about the graph published Georgia’s Department of Public Health website in May 2020, shown in Figure 2-39?



Retrieved 7/3/2021 from

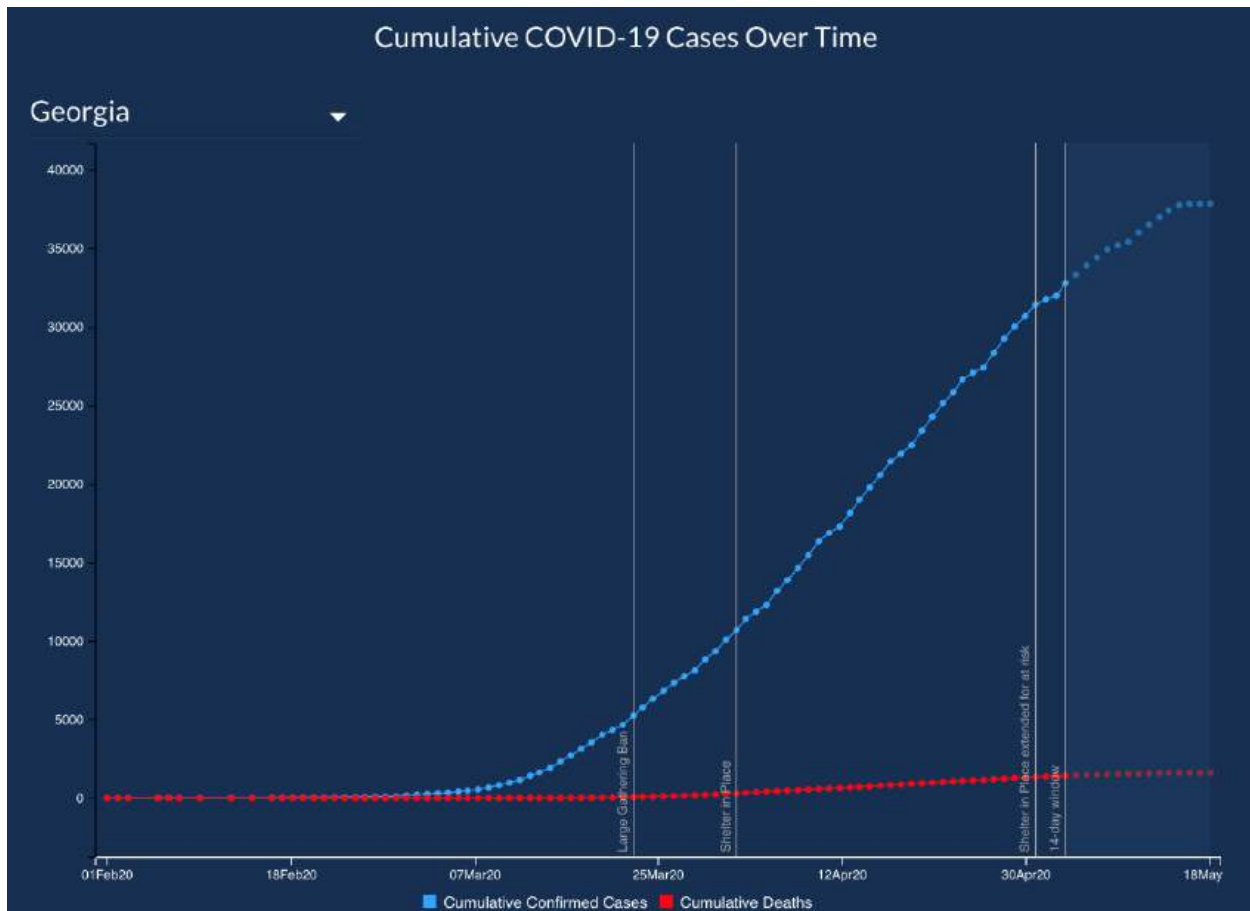
<https://www.vox.com/covid-19-coronavirus-us-response-trump/2020/5/18/21262265/georgia-covid-19-cases-declining-reopening>

Figure 2-39

Solution: There are two misleading items for this graph. The horizontal axis is time, yet the dates are out of sequence starting with April 28, April 27, April 29, May 1, April 30, May 4, May 6, May 5, May 2, May 7, April 26, May 3, May 8, May 9. The first date of April 26 is presented almost at the end of the axis. The graph at first glance would deceive viewers in cases going down over time. A Pareto style chart should never be used for time series data.

The second misleading item is the graph’s title and no label on the y-axis. What does the height of each bar represent? Is the height the number of cases for each county, or is the height the number of deaths and hospitalizations?

The website later corrected the graphic as shown in Figure 2-40.



Retrieved 7/3/2021 from

<https://www.vox.com/covid-19-coronavirus-us-response-trump/2020/5/18/21262265/georgia-covid-19-cases-declining-reopening>

Figure 2-40

Graphical representations of data also play a crucial role in statistical ethics.

- **Enhanced Transparency:** Graphs provide a visual representation of data, making it easier for researchers, analysts, and stakeholders to understand and interpret the information. Transparency is essential in statistical analysis, as it allows others to scrutinize the data, methods, and findings, promoting openness and accountability.
- **Mitigating Bias and Misinterpretation:** Graphs help reduce the risk of bias and misinterpretation of statistical results. By presenting data visually, graphs can reveal patterns, trends, and relationships that may not be apparent in raw numbers alone. This visual clarity enables researchers to identify potential biases and challenges in the data, ensuring more accurate and reliable analyses.
- **Effective Communication:** Graphical representations simplify complex statistical concepts and findings, making them accessible to a wider audience. This is particularly important in ethical statistical analysis, as it promotes effective communication of research outcomes to policymakers, stakeholders, and the general public. Clear and visually appealing graphs facilitate the dissemination of information and support informed decision-making.
- **Ethical Visualization Choices:** Ethical considerations come into play when selecting appropriate graph types and design elements. Ethical visualization choices prioritize accuracy, honesty, and fairness in representing data. This includes using appropriate scales, labeling axes clearly, avoiding distortion or manipulation, and ensuring that the visual presentation does not mislead or deceive the audience.

- **Encouraging Critical Thinking:** Graphs encourage critical thinking and promote a deeper understanding of statistical concepts and analysis. Ethical statistical analysis requires researchers and analysts to critically evaluate data, methods, and results. Graphical representations facilitate this process by allowing for visual comparisons, identifying outliers or anomalies, and enabling exploratory data analysis.

In summary, graphical representations of data are essential in statistical ethics as they enhance transparency, mitigate bias, facilitate effective communication, promote ethical visualization choices, and encourage critical thinking. Large data sets need to be summarized in order to make sense of all the information. It is the role of the researcher or data scientist to make accurate graphical representations that can help make sense of this in the context of the data. By utilizing graphs appropriately and ethically, researchers can ensure that statistical analysis is conducted with integrity, accuracy, and fairness. Tables and graphs can summarize data, but they alone are insufficient. In the next chapter we will look at describing data numerically.

Chapter 2 Exercises

- Which types of graphs are used for quantitative data? Select all that apply.
 - Ogive
 - Pie Chart
 - Histogram
 - Stem-and-Leaf Plot
 - Bar Graph
- Which types of graphs are used for qualitative data? Select all that apply.
 - Pareto Chart
 - Pie Chart
 - Dotplot
 - Stem-and-Leaf Plot
 - Bar Graph
 - Time Series Plot
- The bars for a histogram should always touch, true or false?
- A sample of rents found the smallest rent to be \$600 and the largest rent was \$2,500. What is the recommended class width for a frequency table with 7 classes?
- An instructor had the following grades recorded for an exam.

96	66	65	82	85
82	87	76	80	85
83	69	79	70	83
63	81	94	71	83
99	75	73	83	86

- Create a stem-and-leaf plot.
- Complete the following table.

Class	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
60-69				
70-79				
80-89				
90-99				
Total	25			

- What should the relative frequencies always add up to?
 - What should the last value always be in the cumulative frequency column?
 - What is the frequency for students that were in the C range of 70-79?
 - What is the relative frequency for students that were in the C range of 70-79?
 - Which is the modal class?
 - Which class has a relative frequency of 12%?
 - What is the cumulative frequency for students that were in the B range of 80-89?
 - Which class has a cumulative relative frequency of 40%?
- Eyeglassomatic manufactures eyeglasses for different retailers. The number of lenses for different activities is in the following table.

Activity	Grind	Multi-coat	Assemble	Make Frames	Receive Finished	Unknown
Number of lenses	18,872	12,105	4,333	25,880	26,991	1,508

Grind means that they ground the lenses and put them in frames, multi-coat means that they put tinting or scratch resistance coatings on lenses and then put them in frames, assemble means that they receive frames and lenses from other sources and put them together, make frames means that they make the frames and put lenses in from other sources, receive finished means that they received glasses from another source, and unknown means they do not know where the lenses came from.

- Make a relative frequency table for the data.
- How many of the eyeglasses did Eyeglassomatic assemble?
- How many of the eyeglasses did Eyeglassomatic manufacture all together?
- What is the relative frequency for the assemble category?
- What percent of eyeglasses did Eyeglassomatic grind?

7. The following table is from a sample of five hundred homes in Oregon asked the primary source of heating in their residential homes.

Type of Heat	Percent
Electricity	33
Heating Oil	4
Natural Gas	50
Firewood	8
Other	5

- How many of the households heat their home with firewood?
- What percent of households heat their home with natural gas?

8. The following table is from a sample of 50 undergraduate PSU students.

Class	Relative Frequency Percent
Freshman	18
Sophomore	13
Junior	23
Senior	46

- What percent of students are below a senior class?
- What is the cumulative frequency of the junior class?

9. A sample of heights of 20 people in cm is recorded below. Make a stem-and-leaf plot.

Height (cm)					
167	201	170	185	175	162
182	186	172	173	188	154
185	178	177	184	178	165
169	171	185	178	175	176

10. The stem-and-leaf plot below is for pulse rates before and after exercise.

Before	Pulse Rates	After
9 8 8 7 6 5 2	6	
9 8 8 8 6 5 5 5 1 1 0 0	7	
8 8 7 5 4 2	8	5 6 6 7 8 9
4 0	9	0 1 1 2 3 4 5 5 6 8
4	10	0 1 4 6 7
	11	6 7
	12	4 5 8

- Was pulse rate higher on average before or after exercise?
- What was the fastest pulse rate of the before exercise group?
- What was the slowest pulse rate of the after-exercise group?

11. The following is a sample of 25 temperatures in the summer for Portland, OR in °F.

78	88	93	73	86
85	95	76	89	80
92	83	81	91	74
72	79	90	77	94
68	87	84	82	75

- Create a frequency table with 5 classes.
- Create a histogram.
- Create an ogive.

12. A survey was conducted to determine the preferred genres of books among readers. The following table shows the number of respondents who prefer each genre.

Book Genre	Frequency	Relative Frequency
Fantasy	23	
Nonfiction	21	
Mystery	45	
Romance	57	
Science Fiction	18	
Thriller	11	

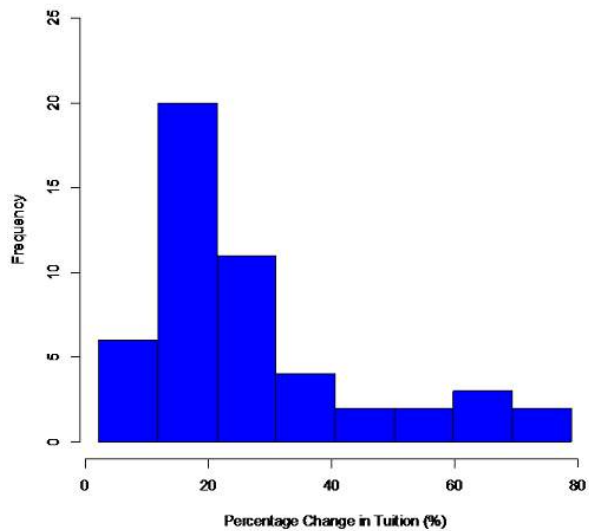
- Calculate the missing relative frequencies.
- Create a bar graph using the frequencies.
- Create a bar graph using the relative frequencies.
- Create a pie graph.

13. The following data represents the percent change in tuition levels at public, four-year colleges (inflation adjusted) from 2008 to 2013 (Weissmann, 2013). Below is the frequency distribution and histogram.

Class Limits	Frequency	Relative Frequency
2.2 – 11.7	6	0.12
11.8 – 21.3	20	0.40
21.4 – 30.9	11	0.22
31.0 – 40.5	4	0.08
40.6 – 50.1	2	0.04
50.2 – 59.7	2	0.04
59.8 – 69.3	3	0.06
69.4 – 78.9	2	0.04

- How many colleges were sampled?
- What was the approximate value of the highest change in tuition?
- What was the approximate value of the most frequent change in tuition?

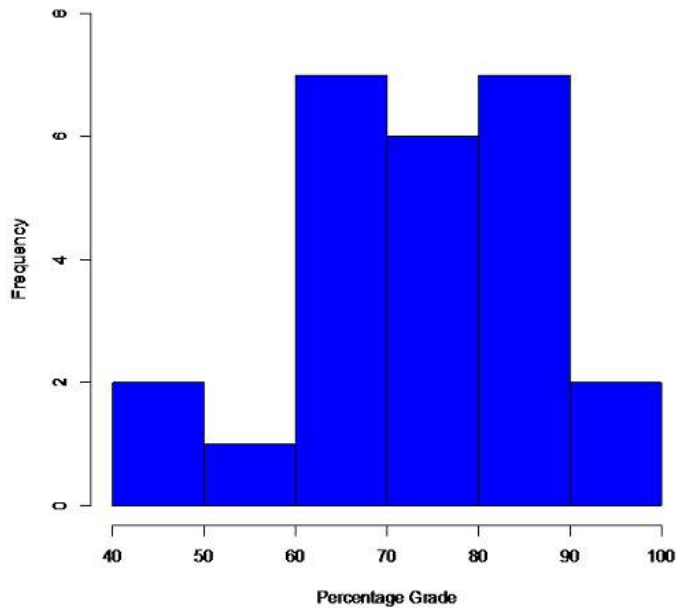
Percentage Change in Tuition Levels (Inflation Adjusted) 2008 to 2013



14. The following data and graph represent the grades in a statistics course.

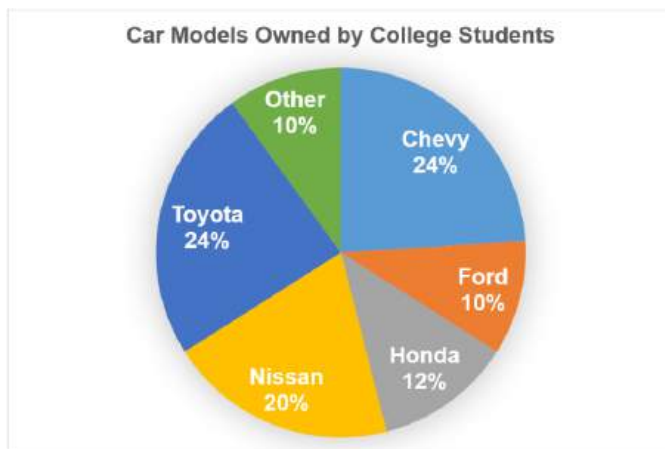
Class Limit	Class Midpoint	Frequency	Relative Frequency
40 – 49.9	45	2	0.08
50 – 59.9	55	1	0.04
60 – 69.9	65	7	0.28
70 – 79.9	75	6	0.24
80 – 89.9	85	7	0.28
90 – 99.9	95	2	0.04

Grades for Statistics Class

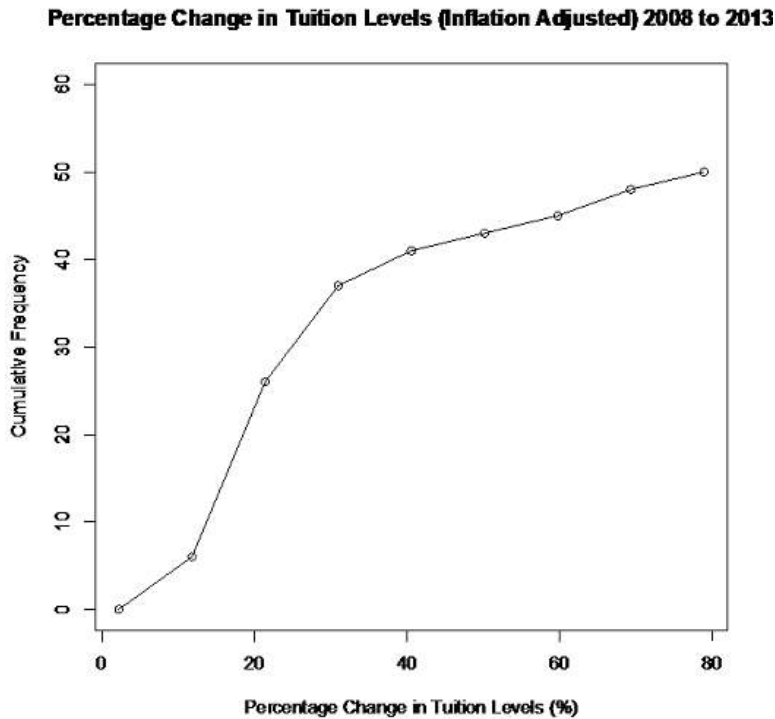


- How many students were in the class?
- What was the approximate lowest and highest grade in the class?
- What percent of students had a passing grade of 70% or higher?

15. The following graph represents a random sample of car models driven by college students. What percent of college students drove a Nissan?



16. The following graph and data represent the percent change in tuition levels at public, four-year colleges (inflation adjusted) from 2008 to 2013 (Weissmann, 2013).



Class Limits	Cumulative Frequency
2.2 – 11.7	6
11.8 – 21.3	26
21.4 – 30.9	37
31.0 – 40.5	41
40.6 – 50.1	43
50.2 – 59.7	45
59.8 – 69.3	48
69.4 – 78.9	50

- How many colleges were sampled?
 - What class of percent changes had the most colleges in that range?
 - How many colleges had a percent change below 50.2% change in tuition?
 - What is the cumulative relative frequency for the 50.2% – 59.7% change in tuition class?
 - Calculate the frequencies based off of the given cumulative frequencies.
17. The table below shows the percentage of market share held by each brand of smartphone. Create a pie graph for the market share of different smartphone brands.

Smartphone Brand	Market Share (%)
Apple	40%
Samsung	30%
Huawei	15%
Xiaomi	10%
Others	5%

18. The table below shows the percentage of respondents who chose each mode of transportation. Create a pie graph for the following preferred modes of transportation among a group of commuters.

Mode of Transportation	Percentage
Car	45%
Public Transportation	30%
Bicycle	15%
Walking	10%

19. A manufacturing company wants to analyze the causes of equipment failures. The following table shows the number of failures caused by different factors in the past month.

Factor	Number of Failures
Human Error	5
Lack of Maintenance	10
Equipment Malfunction	8
Environmental Factors	6
Other	3

- Create a relative frequency table.
 - Create a pie graph for the data.
 - Create a Pareto chart for the data.
 - Identify the most significant factor contributing to the failures.
20. A retail store wants to analyze the reasons for customer returns. The following table shows the frequency of returns categorized by the type of issue. Create a Pareto chart for the data.

Return Issue	Frequency
Customer Dissatisfaction	5
Defective or Damaged Product	20
Size or Fit Issues	8
Incorrect Item Shipped	12
Other	3

21. Eyeglassomatic manufactures eyeglasses for different retailers. The number of lenses by activity is in the table.

Activity	Grind	Multi-coat	Assemble	Make Frames	Receive Finished	Unknown
Number of lenses	18,872	12,105	4,333	25,880	26,991	1,508

- Make a pie chart.
 - Make a bar chart.
 - Make a Pareto chart.
22. A customer support department wants to identify the main reasons for customer complaints. The following table shows the frequency of different complaint categories received in the past week.

Complaint Category	Frequency
Shipping Delays	15
Product Defects	8
Poor Customer Service	10
Billing Errors	4
Other	2

- Create a pie graph for the data.

- b) Create a bar graph using the frequencies for the data.
- c) Create a bar graph using the relative frequency percentages for the data.
- d) Create a Pareto chart.
- e) Which graph would best identify the main reasons for customer complaints?

23. A restaurant is interested in analyzing customer feedback to improve service quality. The following table shows the number of customer complaints received in the past month, categorized by the type of complaint.

Complaint Category	Frequency
Cleanliness	5
Food Quality	2
Noise Level	7
Slow Service	9
Other	2

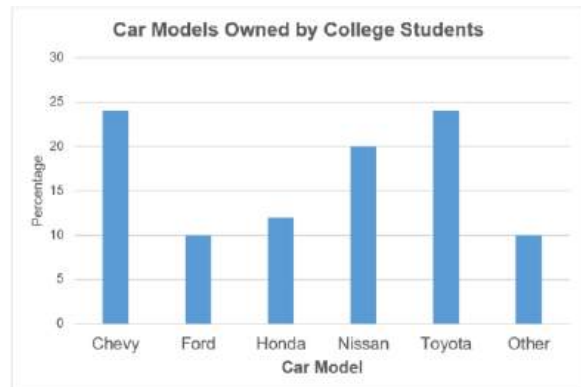
- a) Create a pie graph for the data.
- b) Create a bar graph using relative frequencies for the data.
- c) Create a Pareto graph.

24. The daily sales using different sales strategies is shown in the graph below.



- a) Which strategy generated the most sales?
- b) Was there a particular strategy that worked well for one product, but not for another product?

25. The following graph represents a random sample of car models driven by college students. What was the most common car model?



26. A company conducted a survey to collect data on customer satisfaction levels for its two product lines: Product A and Product B. The following table summarizes the survey results by customer satisfaction rating (Highly Satisfied, Satisfied, Neutral, Dissatisfied, Highly Dissatisfied) for each product.

Product	Highly Satisfied	Satisfied	Neutral	Dissatisfied	Highly Dissatisfied	Total
A	25	35	10	5	2	77
B	20	30	15	8	3	76
Total	45	65	25	13	5	153

- What is the name of this type of table?
- Create a relative frequency contingency table.
- Create a side by side bar graph using the relative frequencies.
- Which product line would you choose based off of the graph?

27. A high school principal conducted a survey to gather data on students' preferred extracurricular activities: Sports, Music, Art, and Science Club. The table below shows the number of students who chose each activity based on their grade level (Grade 9, Grade 10, Grade 11, Grade 12).

Activity	Grade 9	Grade 10	Grade 11	Grade 12
Sports	50	40	30	25
Music	30	25	20	18
Art	20	15	12	10
Science Club	15	12	10	8

- Create a side by side bar graph for the data by day of the week.
- Create a stacked column graph for the data.
- Create a 100% stacked column graph for the data.

28. A restaurant wants to analyze the sales distribution of different menu items. The table below shows the frequency of items sold over several month by each item category and day of the week.

Item Category	Wednesday	Thursday	Friday	Saturday
Appetizers	250	189	364	205
Main Courses	483	407	672	591
Desserts	128	103	238	196

- Create a side by side bar graph for the data by day of the week.
- Create a stacked column graph for the data.
- Create a 100% stacked column graph for the data.

29. The following data show the all-time Olympic medals count by continent.

Continent	Gold	Silver	Bronze
Africa	126	146	169
America	1542	1390	1370
Asia	772	734	832
Europe	3949	4095	4393
Oceania	232	218	273

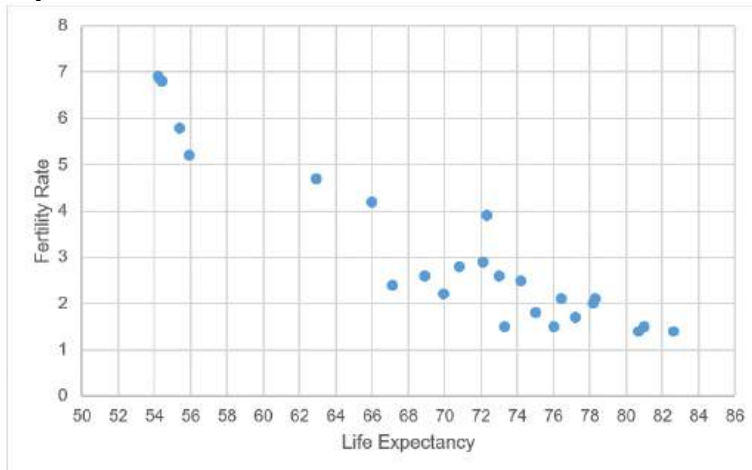
<https://www.olympiandatabase.com/index.php?id=21633&L=1>

- Create a side by side bar graph for the data by type of medal.
- Create a 100% stacked column graph for the data where the continent is on the horizontal axis.

30. The Australian Institute of Criminology gathered data on the number of deaths (per 100,000 people) due to firearms during the period 1983 to 1997. The data is in table below. Create a time-series plot of the data. What is the overall trend over time? (2013, September 26). Retrieved from <http://www.statsci.org/data/oz/firearms.html>.

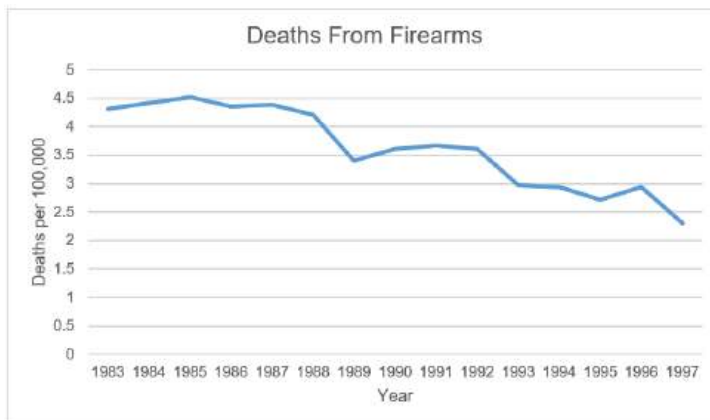
Year	Rate
1983	4.31
1984	4.42
1985	4.52
1986	4.35
1987	4.39
1988	4.21
1989	3.4
1990	3.61
1991	3.67
1992	3.61
1993	2.98
1994	2.95
1995	2.72
1996	2.95
1997	2.3

31. A scatter plot for a random sample of 24 countries shows the average life expectancy and the average number of births per woman (fertility rate). What is the approximate fertility rate for a country that has a life expectancy of 76 years?



(2013, October 14). Retrieved from <http://data.worldbank.org/indicator/SP.DYN.TFRT.IN>.

32. The Australian Institute of Criminology gathered data on the number of deaths (per 100,000 people) due to firearms during the period 1983 to 1997. The time-series plot is below. What year had the highest rate of deaths?

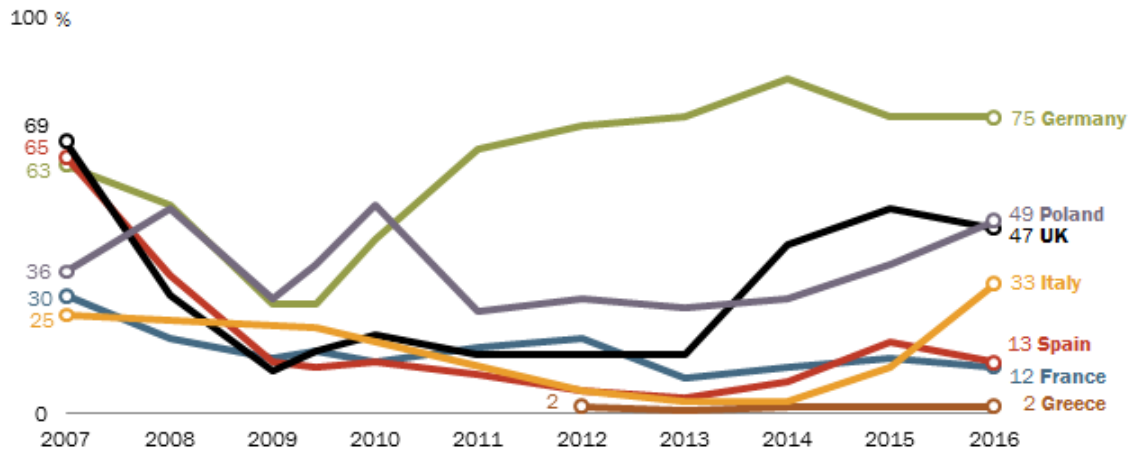


(2013, September 26). Retrieved from <http://www.statsci.org/data/oz/firearms.html>.

33. A survey by the Pew Research Center, conducted in 16 countries among 20,132 respondents from April 4 to May 29, 2016, before the United Kingdom's Brexit referendum to exit the EU. The following is a time series graph for the proportion of survey respondents by country that responded that the current economic situation in their country was good.

Some European publics view economy on the rebound, but others remain negative

The current economic situation in our country is good



Source: Spring 2016 Global Attitudes Survey, Q3.

PEW RESEARCH CENTER

<http://www.pewglobal.org/2016/08/09/views-on-national-economies-mixed-as-many-countries-continue-to-struggle/>

- a) Which country had the most favorable outlook of their country's economic situation in 2010?
 b) Which country had the least favorable outlook of their country's economic situation in 2016?
34. The following data represent the leaching rates (percent of lead extracted vs. time in minutes) for lead in solutions of magnesium chloride ($MgCl_2$).

Time (x)	4	8	16	30	60	120
Percent Extracted (y)	1.2	1.6	2.3	2.8	3.6	4.4

- a) Create a time-series plot of the data.
 b) Is there a negative or positive trend to the data?

35. Bone mineral density and cola consumption have been recorded for a sample of patients. Let x represent the number of colas consumed per week and y the bone mineral density in grams per cubic centimeter. Create a scatter plot for the following data.

x	y
1	0.883
2	0.8734
3	0.8898
4	0.8852
5	0.8816
6	0.863
7	0.8634
8	0.8648
9	0.8552
10	0.8546
11	0.862

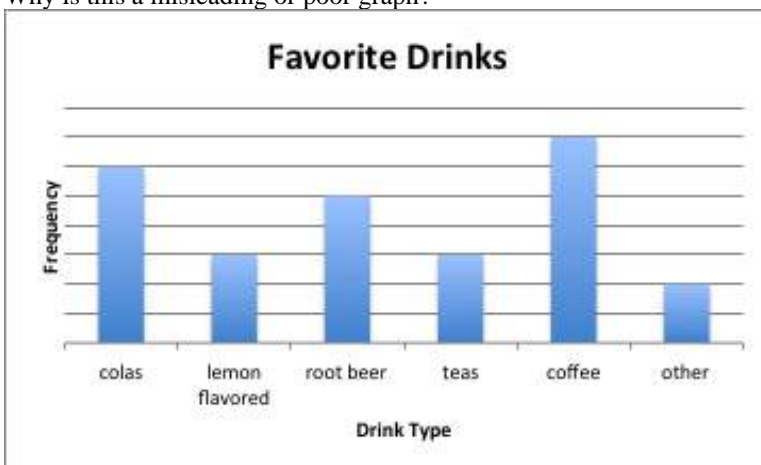
36. An object is thrown from the top of a building. The following data measure the height (ft) of the object from the ground for a five-second period. Create a scatter plot for the following data.

Seconds	Height
0.5	112.5
1	110.875
1.5	106.8
2	100.275
2.5	91.3
3	79.875
3.5	70.083
4	59.83
4.5	30.65
5	0

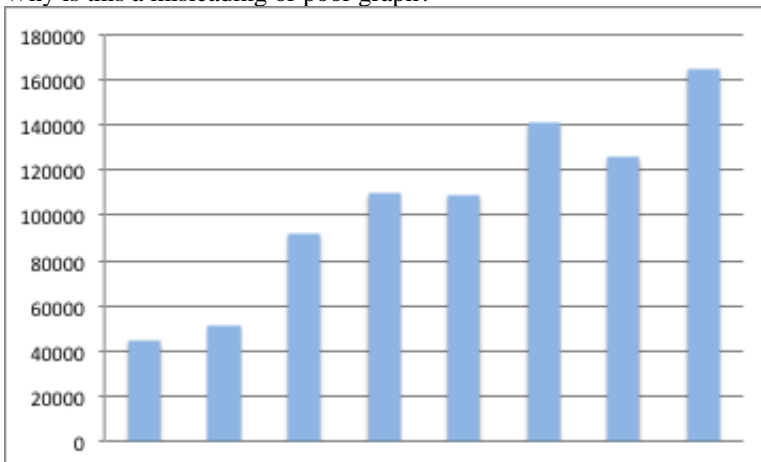
37. Create a scatter plot for the predicted average high temperature ($^{\circ}\text{F}$) per month by the Farmer's Almanac (x) in Portland, Oregon and the actual high (y) temperature per month that occurred.

Farmer's Almanac	45	50	57	62	69	72	81	90	78	64	51	48
Actual High	46	52	60	61	72	78	82	95	85	68	52	49

38. Why is this a misleading or poor graph?



39. Why is this a misleading or poor graph?

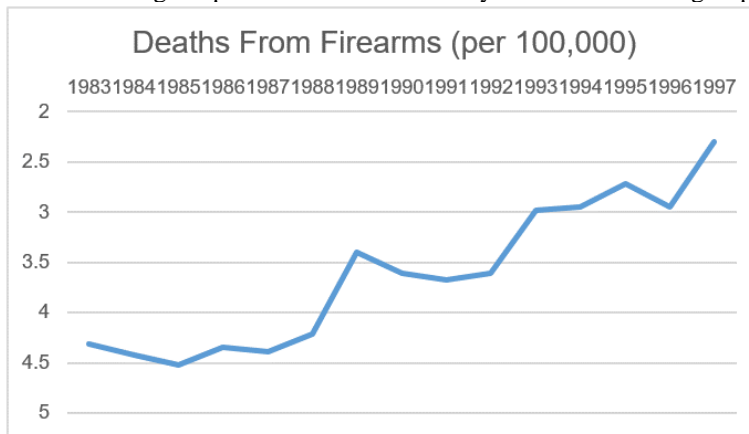


40. Why is this a misleading or poor graph?



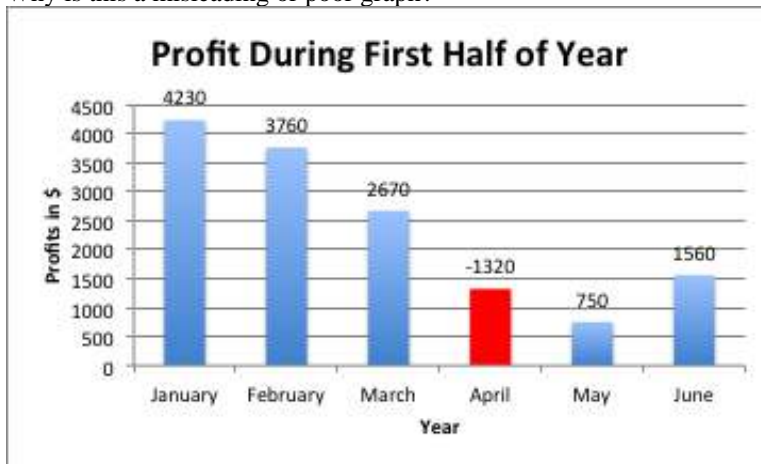
United States unemployment. (2013, October 14). Retrieved from <http://www.tradingeconomics.com/united-states/unemployment-rate>

41. The Australian Institute of Criminology gathered data on the number of deaths (per 100,000 people) due to firearms during the period 1983 to 1997. Why is this a misleading or poor graph?



(2013, September 26). Retrieved from <http://www.statsci.org/data/oz/firearms.html>.

42. Why is this a misleading or poor graph?



Chapter 3

Descriptive Statistics



3.1 Measures of Center

- 3.1.1 Mode
- 3.1.2 Mean
- 3.1.3 Weighted Mean
- 3.1.4 Median
- 3.1.5 Outliers
- 3.1.6 Distribution Shapes

3.2 Measures of Spread

- 3.2.1 Range
- 3.2.2 Variance & Standard Deviation
- 3.2.3 Coefficient of Variation

3.3 Measures of Placement

- 3.3.1 Z-Scores
- 3.3.2 Percentiles
- 3.3.3 Quartiles
- 3.3.4 Five Number Summary & Outliers
- 3.3.5 Modified Box-and-Whisker Plot

3.4 Correlation and Linear Regression

- 3.4.1 Correlation
- 3.4.2 Linear Regression

3.1 Measures of Center

Both graphical and numerical methods of summarizing data make up the branch of statistics known as **descriptive statistics**. Later, descriptive statistics will be used to estimate and make inferences about population parameters using methods that are part of the branch called **inferential statistics**. This section introduces numerical measurements to describe sample data.

This section focuses on measures of central tendency. Many times, you are asking what to expect “on average.” Such as when you pick a career, you would probably ask how much you expect to earn in that field. If you are trying to buy a home, you might ask how much homes are selling for in your area. If you are planting vegetables in your garden, you might want to know how long it will be until you can harvest. These questions, and many more, can be answered by knowing the center of the data set. The three most common measures of the “center” of the data are called the mode, mean, and median.

3.1.1 Mode

To find the mode, you count how often each data value occurs, and then determine which data value occurs most often.

The **mode** is the data value that occurs the most frequently in the data.

There may not be a mode at all, or you may have more than one mode. If there is a tie between two values for the greatest number of times then both values are the mode and the data is called bimodal (two modes). If every data point occurs the same number of times, there is no mode. If there are more than two numbers that appear the most times, then usually we write there is no mode. When looking at grouped data in a frequency distribution or a histogram then the largest frequency is called the **modal class**.

Below is a dotplot showing the height of some 3-year-old children in cm and we would like to answer the question, “How tall are 3-year-olds?”

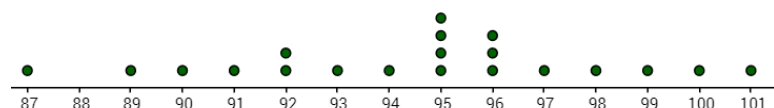


Figure 3-1

From the graph, we can see that the most frequent value is 95 cm. This is not exactly the middle of the distribution, but it is the most common height and is close to the middle in this case. We call this most frequent value the mode.

For larger data sets, use software to find the mode or at least sort the data so that you can see grouping of numbers. Excel reports a mode at the first repetitive value, so be careful in Excel with bimodal data or data with many multiples that would really have no mode at all.

Note that zero may be the most frequent value in a data set. The mode = 0 is not the same as “no mode” in the data set.

The mode is the observation that occurs most often.

- **Example 3-1:** -5 4 8 3 4 2 0 mode = 4
- **Example 3-2:** 3 -6 0 1 -2 1 0 5 0 mode = 0
- **Example 3-3:** 18 25 15 32 10 27 no mode (Excel writes N/A)
- **Example 3-4:** 15 23 18 15 24 23 17 modes = 15, 23 (bimodal)
- **Example 3-5:** 100 125 100 125 130 140 130 140 no mode (Excel gives 100)

Summation Notation

Throughout this course, we will be using **summation notation**, also called sigma notation. The capital Greek letter Σ “sigma” means to add. For example, Σx means to sum up all of the x values where X is the variable name.

Example 3-6: A random sample of households had the following number of children living at home 4, -3, 2, 1, and 3. Calculate Σx .

Solution: Let $x_1 = 4, x_2 = -3, x_3 = 2, x_4 = 1, x_5 = 3$. Start with the first value $i = 1$ up to the n^{th} value $i = 5$ to get $\sum_{i=1}^n x_i = 4 + -3 + 2 + 1 + 3 = 7$.

To make things simpler we will drop the subscripts and write $\sum_{i=1}^n x_i$ as Σx_i or Σx .

The order of operations is important in summation notation.

For example, $\Sigma x^2 = (4)^2 + (-3)^2 + (2)^2 + (1)^2 + (3)^2 = 39$.

When we insert parentheses $(\Sigma x)^2 = (4 + -3 + 2 + 1 + 3)^2 = (7)^2 = 49$.

Note that $\Sigma x^2 \neq (\Sigma x)^2$.

“‘One of the interesting things about space,’ Arthur heard Slartibartfast saying to a large and voluminous creature who looked like someone losing a fight with a pink duvet and was gazing raptly at the old man’s deep eyes and silver beard, ‘is how dull it is.’
‘Dull?’ said the creature, and blinked her rather wrinkled and bloodshot eyes.
‘Yes,’ said Slartibartfast, ‘staggeringly dull. Bewilderingly so. You see, there’s so much of it and so little in it. Would you like me to quote some statistics?’
‘Er, well...’
‘Please, I would like to. They, too, are quite sensationally dull.’”
(Adams, 2002)

3.1.2 Mean

The **mean** is the arithmetic average of the numbers. This is the center that most people call the average.

Distinguishing between a population and a sample is very important in statistics. We frequently use a representative sample to generalize about a population.

A **statistic** is any characteristic or measure from a sample. A **parameter** is any characteristic or measure from a population. We use sample statistics to make inferences about population parameters.

The **sample mean** = \bar{x} (pronounced “x bar”) of a sample of n observations $x_1, x_2, x_3, \dots, x_n$ taken from a population, is given by the formula:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

The **population mean** = μ (pronounced “mu”) is the average of the entire population, is given by the formula:

$$\mu = \frac{\Sigma x}{N} = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N}$$

Most cases, you cannot find the population parameter, so you use the sample statistic to estimate the population parameter. Since μ cannot be calculated in most situations, the value for \bar{x} is used to estimate μ . You should memorize the symbol μ and what it represents for future reference.

Example 3-7: Find the mean for the following sample of house prices (\$1,000): 325, 375, 385, 395, 420, and 825.

Solution: Before starting any mathematics problem, it is always a good idea to define the unknown in the problem. In this case, you want to define the variable. The symbol for the variable is x . The variable is x = price of a house in \$1,000.

$$\bar{x} = \frac{\sum x}{n} = \frac{325+375+385+395+420+825}{6} = 454.1\bar{6}$$

The sample mean house price is \$454,166.67.

We can use technology to find the mean. Directions for the TI calculators are in the next section. In Excel, use the cell function AVERAGE(array). For this example, we can type the data into column A and then in a blank cell =AVERAGE(A1:A6).

	A	B
1	325	
2	375	
3	385	
4	395	
5	420	
6	825	=AVERAGE(A1:A6)
7		

3.1.3 Weighted Mean

Weighted averages are used quite often in real life. Some teachers use them in calculating your grade in the course, or your grade on a project. Some employers use them in employee evaluations. The idea is that some components of a mean are more important than others are. As an example, a full-time teacher at a community college may be evaluated on their service to the college, their service to the community, whether their paperwork is turned in on time, and their teaching. However, teaching is much more important than whether their paperwork is turned in on time. When the evaluation is completed, more weight needs to be given to the teaching and less to the paperwork. This is a weighted average.

Weighted Mean = $\frac{\text{sum of the scores times their weights}}{\text{sum of all the weights}} = \frac{\sum(xw)}{\sum w}$, where w is the weight of the data value x .

Example 3-8: In your biology class, your final grade is based on several things: a lab score, scores on two major tests, and your score on the final exam. There are 100 points available for each score. The lab score is worth 15% of the course, the two exams are worth 25% of the course each, and the final exam is worth 35% of the course. Suppose you earned scores of 95 on the labs, 83 and 76 on the two exams, and 84 on the final exam. Compute your weighted average for the course.

Solution: Variable: x = score

$$\text{The weighted mean is } \frac{\sum(xw)}{\sum w} = \frac{95(0.15)+83(0.25)+76(0.25)+84(0.35)}{0.15+0.25+0.25+0.35} = \frac{83.4}{1.00} = 83.4.$$

The course average is 83.4 %.

Example 3-9: A faculty evaluation process at Portland State University rates a faculty member on the following activities: teaching, publishing, committee service, community service, and submitting paperwork in a timely manner. The process involves reviewing student evaluations, peer evaluations, and supervisor evaluation for each teacher and awarding them a score on a scale from 1 to 10 (with 10 being the best). The weights for each activity are 20 for teaching, 18 for publishing, 6 for committee service, 4 for community service, and 2 for paperwork.

- a) One faculty member had the following ratings: 8 for teaching, 9 for publishing, 2 for committee work, 1 for community service, and 8 for paperwork. Compute the weighted average of their evaluation.

Solution:

Variable: x = rating

$$\text{The weighted average is } \frac{\sum(xw)}{\sum w} = \frac{8(20)+9(18)+2(6)+1(4)+8(2)}{20+18+6+4+2} = \frac{354}{50} = 7.08.$$

The average evaluation score is 7.08.

- b) Another faculty member had ratings of 6 for teaching, 8 for publishing, 9 for committee work, 10 for community service, and 10 for paperwork. Compute the weighted average of their evaluation.

Solution:

$$\text{The weighted average is } \frac{\sum(xw)}{\sum w} = \frac{6(20)+8(18)+9(6)+10(4)+10(2)}{20+18+6+4+2} = \frac{378}{50} = 7.56.$$

The average evaluation score is 7.56.

- c) Which faculty member had the higher average evaluation?

Solution:

The second faculty member has a higher average evaluation.

3.1.4 Median

Another statistic that measures the center of a distribution is the median.

The **median** is the data value in the middle of the ordered data that has 50% of the data below that point and 50% of the data above that point. The median is also referred to as the 50th percentile and is the midpoint of a distribution.

To find the median:

1. Arrange the observations from smallest to largest.
2. If the number of observations n is odd, the middle observation is the median.
3. If the number of observations n is even, the mean of the two middle observations is the median.

Example 3-10: Find the median for the following sample of ages: 15, 23, 18, 15, 24, 23, and 17.

Solution: First, sort the data: 15, 15, 17, 18, 23, 23, and 24. The sample size is odd so the median will be the middle number. Use your fingers to cover outside numbers, one pair at a time until you get to 18. Median = 18 years old.

Example 3-11: Find the median for the following sample of house prices (in \$1,000): 325, 375, 385, 395, 420, and 825.

Solution: The data is already ordered from smallest to largest. The sample size is even so take the average of the two middle values $\frac{385+395}{2} = 390$. The median house price is \$390,000.

We can use technology to find the median. Directions for the TI calculators are in the next section. In Excel the median is found using the cell function MEDIAN(array). For this example, we can type the data into column A and then in a blank cell =MEDIAN(A1:A6).

Recall that the sample mean house price is \$454,167. Note that the median is much lower than the mean for this example. The observation of 825 is an outlier and is very large compared to the rest of the data. The sample mean is sensitive to unusual observations, i.e., outliers. The median is resistant to outliers.

3.1.5 Outliers

An **outlier** is a data value that is very different from the rest of the data and is far enough from the center. If there are extreme values in the data, the median is a better measure of the center than the mean. The mean is not a resistant measure because it is moved in the direction of the outlier. The median and the mode are resistant measures because they are not affected by extreme values.

As a consumer, you need to be aware that people choose the measure of center that best supports their claim. When you read an article in the newspaper and it talks about the “average,” it usually means the mean but sometimes it refers to the median. Some articles will use the word “median” instead of “average” to be more specific. If you need to make an important decision and the information says “average,” it would be wise to ask if the “average” is the mean or the median before you decide.

As an example, suppose that a company administration wants to use the mean salary as the average salary for the company. This is because the high salaries of the administration will pull the mean higher. The company can say that the employees are paid well because the average is high. However, the employees’ union wants to use the median since it discounts the extreme values of the administration and will give a lower value of the average. This will make the salaries seem lower and that a raise is in order.

Why use the mean instead of the median? When multiple samples are taken from the same population, the sample means tend to be more consistent than other measures of the center. The sample mean is the more reliable measure of center.

3.1.6 Distribution Shapes

Remember that there are varying levels of skewness and symmetry. Sample data is rarely exactly symmetric, but is approximately symmetric. Outliers will pull the mean in the direction of the outlier. If the distribution has a skewed tail to the left, the mean will be smaller than the median. If the distribution has a skewed tail to the right, the mean will be larger than the median. The mode, or modal class, is the tallest point(s), highest frequency, of the distribution. The following show examples of different distribution shapes. Figures 3-2 to 3-5 show example distribution shapes.

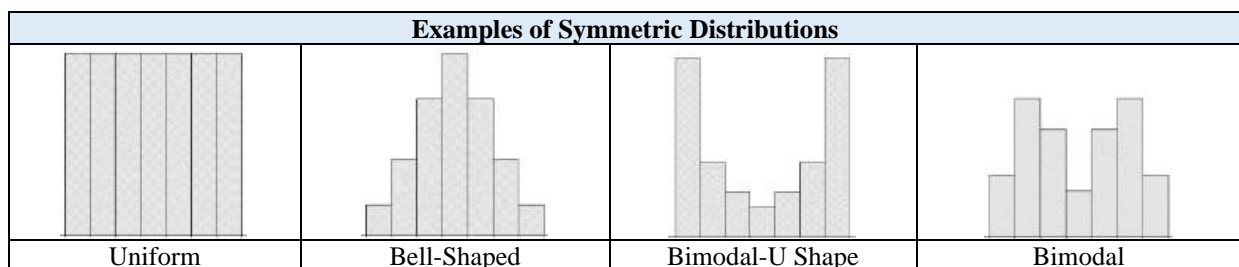


Figure 3-2

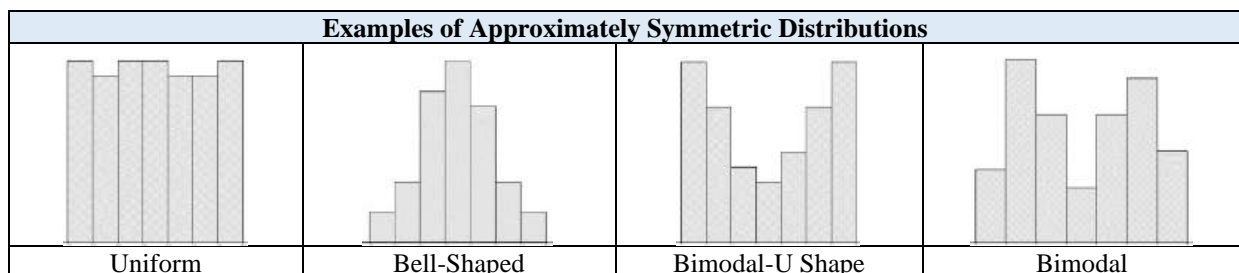


Figure 3-3

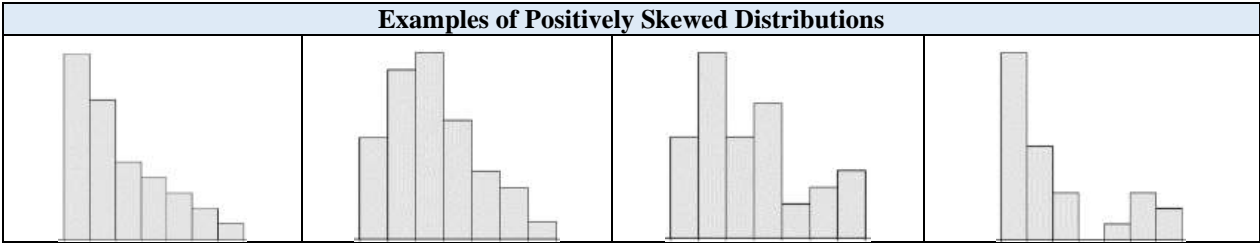


Figure 3-4

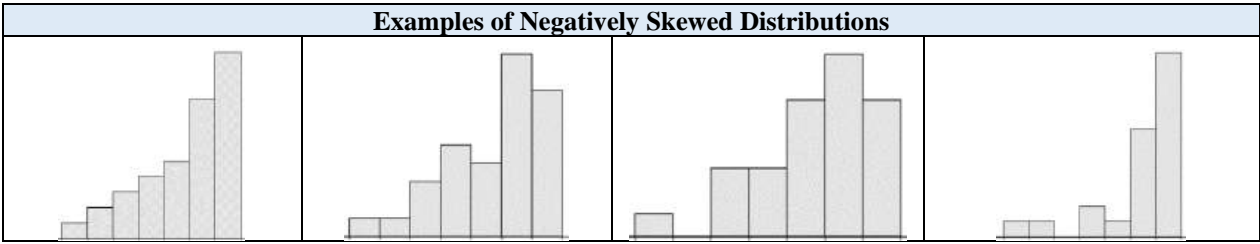


Figure 3-5

Comparing the mean and the median provides useful information about the distribution shape.

- If the mean is equal to the median, the data is symmetric, see Figure 3-6.

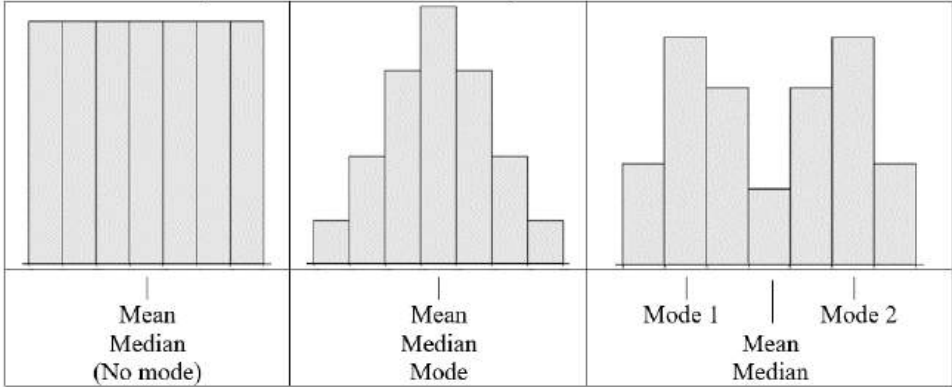


Figure 3-6

If the mean is larger than (to the right of) the median, the data is right skewed or positively skewed, see Figure 3-7.

If the mean is smaller than (to the left of) the median, the data is left skewed, or negatively skewed, see Figure 3-8.

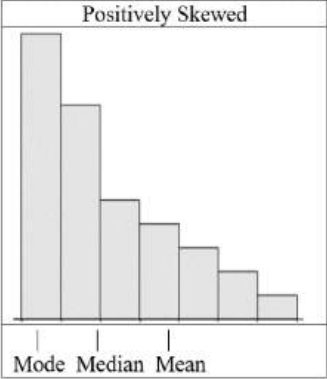


Figure 3-7

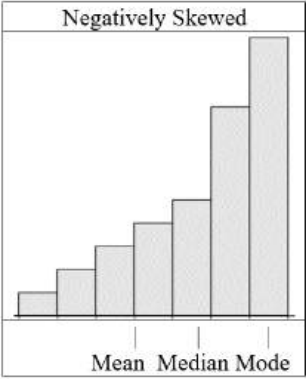


Figure 3-8

Example 3-12: Figure 3-9 is a histogram for a random sample of student rent prices from Example 2-12. Comment on the distribution shape.

1500	1350	350	1200	850	900
1500	1150	1500	900	1400	1100
1250	600	610	960	890	1325
900	800	2550	495	1200	690

Solution: If we were to use Excel to find the mean and median, we would get that the mean house rental price is \$1,082.08 and the median house rental price is \$1,030. The mean is larger than the median and is being pulled to the right by the outlier of \$2,550.

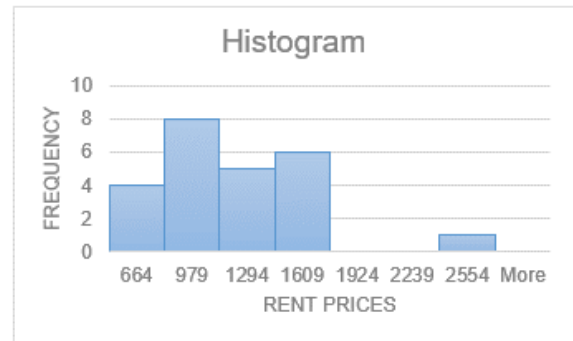


Figure 3-9

If you were to draw a curve around the bars as in Figure 3-10, you would get a tail for the one data point on the right. The outlier on the right is the direction of the skewness.

This distribution is skewed to the right, or positively skewed.



Figure 3-10

Which measure of center is used on which type of data?

- Mode can be found on nominal, ordinal, interval, and ratio data, since the mode is just the data value that occurs most often. You are just counting the data values.
- Median can be found on ordinal, interval, and ratio data, since you need to put the data in order. As long as there is order to the data, you can find the median.
- Mean can be found on interval and ratio data, since you must have numbers to add together.

3.2 Measures of Spread

Variability is an important idea in statistics. If you were to measure the height of everyone in your classroom, every student gives you a different value. That means not every student has the same height. Thus, there is variability in people's heights. If you were to take a sample of the income level of people in a town, every sample gives you different information. There is variability between samples too. Variability describes how the data are spread out. If the data are very close to each other, then there is low variability. If the data are very spread out, then there is high variability. How do you measure variability? It would be good to have a number that measures it. This section will describe some of the different measures of variability, also known as variation.

Numerical statistics for variation can show how spread out data is. The variation of data is relative, and is usually used when comparing two sets of similar data. When we are making inferences about an average, we can make better estimates when there is less variation in the data. The four most common measures of the "spread" of data are called the range, variance, standard deviation, and coefficient of variation.

A sample of house prices (in \$1,000): 325, 375, 385, 395, 420, and 825, found the mean house price of \$454,167. How much does this tell you about the price of all houses? Can you tell if most of the prices were close to the mean

or were the prices really spread out? What is the highest price and the lowest price? All you know is that the center of the price is \$454,167. What if you were approved for only \$400,000 for a home loan, could you buy a home in this area? You need more information.

3.2.1 Range

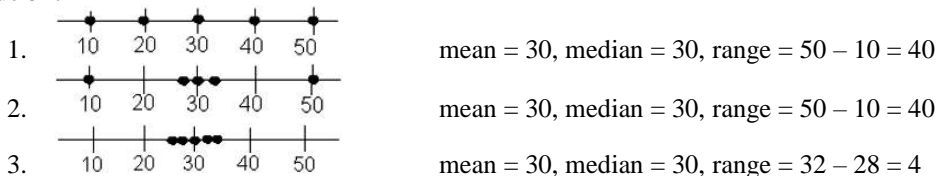
The **range** of a set of data is the difference between the highest and the lowest data values (or maximum and minimum values). Note in statistics we only report a single number which represents the spread from the lowest to highest value.

$$\text{Range} = \text{Max} - \text{Min.}$$

Example 3-13: Look at the following three sets of data. Find the mean, median and range of each of data set.

1. 10, 20, 30, 40, 50
2. 10, 29, 30, 31, 50
3. 28, 29, 30, 31, 32

Solution:



Based on the mean, median, and range, the first two distributions are the same, but you can see from the graphs that they are distributed differently. In part 1, the data are spread out equally. In part 2, the data has a clump in the middle and a single value at each end. The mean and median are the same for part 3, but the range is much smaller. All the data is clumped together in the middle.

3.2.2 Variance & Standard Deviation

The range does not really provide a very detailed picture of the variability. A better way to describe how the data is spread out is needed. Instead of looking at the distance as the highest value from the lowest, how about looking at the distance each value is from the mean? This spread is called the deviation.

Example 3-14: Suppose a vet wants to analyze the weights of cats. The weights (in pounds) of five cats are 6.8, 8.2, 7.5, 9.4, and 8.2. Compute the deviation for each of the data values. The deviation is how far each data point is from the mean. To be consistent always subtract the data point minus the mean.

Solution: Variable: X = weight of a cat. First, find the mean for the data set. The mean is $\bar{x} = \frac{\sum x}{n} = \frac{(6.8+8.2+7.5+9.4+8.2)}{5} = 8.02$ pounds. Subtract the mean from each data point to get the deviations, see Figure 3-11.

Deviations of Weights of Cats

x	$x - \bar{x}$
6.8	$6.8 - 8.02 = -1.22$
8.2	$8.2 - 8.02 = 0.18$
7.5	$7.5 - 8.02 = -0.52$
9.4	$9.4 - 8.02 = 1.38$
8.2	$8.2 - 8.02 = 0.18$

Figure 3-11

Now average the deviations. Add the deviations together, see Figure 3-12.

Sum of Deviations of Weights of Cats

x	$x - \bar{x}$
6.8	$6.8 - 8.02 = -1.22$
8.2	$8.2 - 8.02 = 0.18$
7.5	$7.5 - 8.02 = -0.52$
9.4	$9.4 - 8.02 = 1.38$
8.2	$8.2 - 8.02 = 0.18$
Total	0

Figure 3-12

The average distance from the mean cannot be zero. The reason the deviations add to 0 is that there are some positive and negative values. The sum of the deviations from the mean will always be zero.

To get rid of the negative signs square each deviation, see Figure 3-13.

Squared Deviations of Weights of Cats

x	$x - \bar{x}$	$(x - \bar{x})^2$
6.8	$6.8 - 8.02 = -1.22$	1.4884
8.2	$8.2 - 8.02 = 0.18$	0.0324
7.5	$7.5 - 8.02 = -0.52$	0.2704
9.4	$9.4 - 8.02 = 1.38$	1.9044
8.2	$8.2 - 8.02 = 0.18$	0.0324
Total	0	3.728

Figure 3-13

Then average the total of the squared deviations. The only thing is that in statistics there is a strange average here. Instead of dividing by the number of data values, you divide by the number of data values minus one. This $n - 1$ is called the degrees of freedom and will be discussed more later in the text. When we divide by the degrees of freedom, this gives an unbiased statistic. In this case, you would have the following:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{3.728}{5 - 1} = \frac{3.728}{4} = 0.932 \text{ pounds}^2$$

Notice that this statistic is denoted as s^2 . This statistic is called the sample variance and it is a measure of the average squared distance from the mean. If you now take the square root, you will get the average distance from the mean. The square root of the variance is called the sample standard deviation, and is denoted with the letter s .

$$s = \sqrt{0.932} = 0.9654 \text{ pounds}$$

The standard deviation is the average (mean) distance from a data point to the mean. It can be thought of as how much a typical data point differs from the mean.

The **sample variance** formula: $s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$.

Where \bar{x} is the sample mean, n is the sample size, and Σ means to find the sum.

The **sample standard deviation** formula: $s = \sqrt{s^2} = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$.

The $n - 1$ in the denominator has to do with a concept called degrees of freedom (df). Dividing by the df makes the sample standard deviation a better approximation of the population standard deviation than dividing by n .

We rarely will find a population variance or standard deviation, but you will need to know the symbols.

The **population variance** formula: $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$.

The **population standard deviation** formula: $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$.

The lower-case Greek letter σ pronounced “sigma” and σ^2 represents the population variance, μ is the population mean, and N is the size of the population.

Note: the sum of the deviations should always be zero. Try not to round too much in the calculations for standard deviation since each rounding causes a slight error.

Example 3-15: Suppose that a manager wants to test two new training programs. They randomly select 5 people for each training type and measures the time it takes to complete a task after the training. The times for both trainings are in table below. Which training method is more consistent?

Time to Finish Task in Minutes

Training 1	56	75	48	63	59
Training 2	60	58	66	59	58

Solution: It is important that you define what each variable is since there are two of them.

Variable 1: X_1 = productivity from training 1

Variable 2: X_2 = productivity from training 2

The units and scale are the same for both groups. To answer which training method better, first you need some descriptive statistics. Start with the mean for each sample.

$$\bar{x}_1 = \frac{56 + 75 + 48 + 63 + 59}{5} = 60.2 \text{ minutes}$$

$$\bar{x}_2 = \frac{60 + 58 + 66 + 59 + 58}{5} = 60.2 \text{ minutes}$$

Since both means are the same values, you cannot answer the question about which is better.

Now calculate the standard deviation for each sample, see Figures 3-14 and 3-15.

Squared Deviations for Training 1 Squared Deviations for Training 2

x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$
56	-4.2	17.64
75	14.8	219.04
48	-12.2	148.84
63	2.8	7.84
59	-1.2	1.44
Total	0	394.8

Figure 3-14

x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
60	-0.2	0.04
58	-2.2	4.84
66	5.8	33.64
59	-1.2	1.44
58	-2.2	4.84
Total	0	44.8

Figure 3-15

The variance for each sample is:

$$s_1^2 = \frac{394.8}{4} = 98.7 \text{ minutes}^2$$

$$s_2^2 = \frac{44.8}{4} = 11.2 \text{ minutes}^2.$$

The standard deviations are:

$$s_1 = \sqrt{98.7} = 9.9348 \text{ minutes}$$

$$s_2 = \sqrt{11.2} = 3.3466 \text{ minutes.}$$

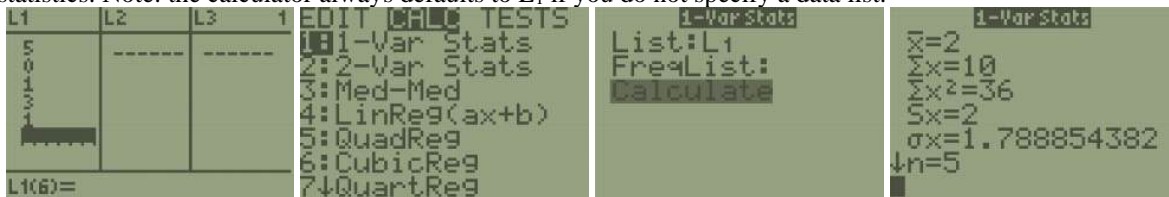
Comparing the standard deviations, the second training method seemed to be the better training since the data is less spread out. This means it is more consistent. It would be better for the managers in this case to have a training program that produces more consistent results so they know what to expect for the time it takes to complete the task.

Descriptive statistics can be time-consuming to calculate by hand so use technology.

One Variable Statistics on the TI Calculator

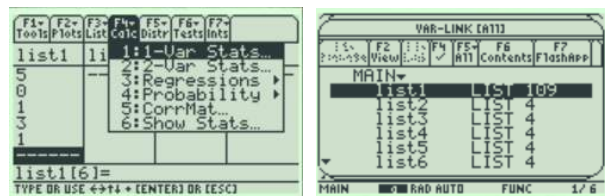
The procedure for calculating the sample mean (\bar{x}) and the sample standard deviation (s_x) for the TI calculator are shown below. Note, the TI calculator also gives you the population standard deviation (σ_x) because it does not know whether the data you input is a population or a sample. You need to decide which value you need to use, based on whether you have a population or sample. In almost all cases you have a sample and will be using s_x . In addition, the calculator uses the notation of s_x instead of just s . It is just a way for it to denote the information.

TI-84: Enter the data in a list and then press [STAT]. Use cursor keys to highlight CALC. Press 1 or [ENTER] to select **1:1-Var Stats**. Press [2nd], then press the number key corresponding to your data list. Press [Enter] to calculate the statistics. Note: the calculator always defaults to L₁ if you do not specify a data list.



s_x is the sample standard deviation. You can arrow down and find more statistics. Use the min and max to calculate the range by hand. To find the variance simply square the standard deviation.

TI-89: Press [APPS], select **FlashApps** then press [ENTER]. Highlight **Stats/List Editor** then press [ENTER]. Press [ENTER] again to select the main folder. To clear a previously stored list of data values, arrow up to the list name you want to clear, press [CLEAR], then press enter.

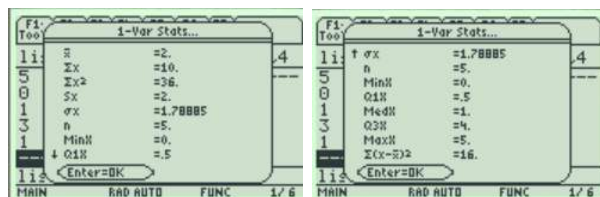


Press [F4], select 1: 1-Var Stats. To get the list name to the List box, press [2nd] [Var-Link], arrow down to list1 and press [Enter]. This will bring list1 to the List box. Press [Enter] to enter the list name and then enter again to calculate.

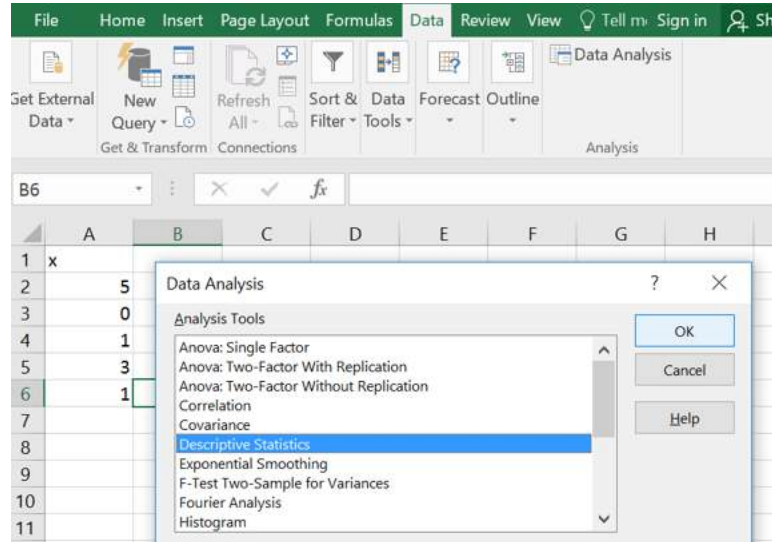


Use the down arrow key to see all the statistics.

S_x is the sample standard deviation. You can arrow down and find more statistics. Use the min and max to calculate the range by hand. To find the variance simply square the standard deviation or take the last sum of squares divided by $n - 1$.

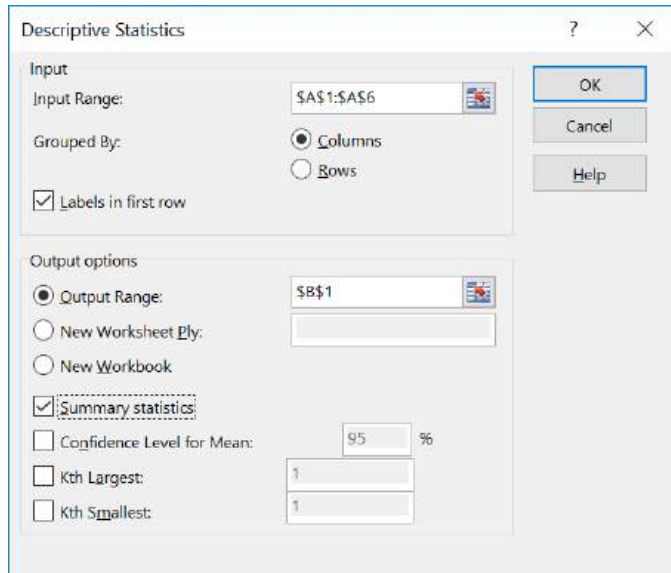


Excel: Type in the data into one column, select the Data tab, and choose Data Analysis. Select Descriptive Statistics, and then select OK.



Highlight the data for the Input Range, if you highlighted a label; check the Labels in first row box. Select the circle to the left of Output Range, then click into the box to the right of the Output Range and select one cell where you want the top left-hand corner of your summary table to start. Select the box next to Summary statistics, then select OK, see below.

We get the following summary statistics:



x	
Mean	2
Standard Error	0.8944
Median	1
Mode	1
Standard Deviation	2
Sample Variance	4
Kurtosis	-0.1875
Skewness	0.9375
Range	5
Minimum	0
Maximum	5
Sum	10
Count	5

In general, a “small” standard deviation means the data are close together (more consistent) and a “large” standard deviation means the data is spread out (less consistent). Sometimes you want consistent data and sometimes you do not. As an example, if you are making bolts, you want the lengths to be very consistent so you want a small standard deviation. If you are administering a test to see who can be a pilot, you want a large standard deviation so you can tell whom the good and bad pilots are.

What do “small” and “large” mean? To a bicyclist whose average speed is 20 mph, $s = 20$ mph is huge. To an airplane whose average speed is 500 mph, $s = 20$ mph is nothing. The “size” of the variation depends on the size of the numbers in the problem and the mean. Another situation where you can determine whether a standard deviation is small or

large is when you are comparing two different samples. A sample with a smaller standard deviation is more consistent than a sample with a larger standard deviation.

If we were to compare the variability between two histograms. The standard deviation and variance measure the average spread from left to right. Take a moment and see if you can order the following histograms from the smallest to the largest standard deviation.

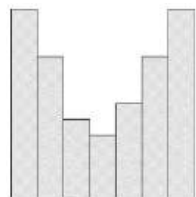


Figure 3-16

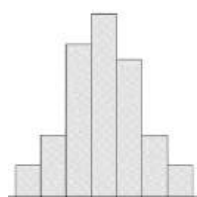


Figure 3-17

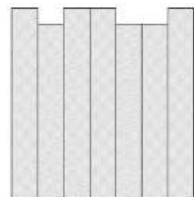


Figure 3-18

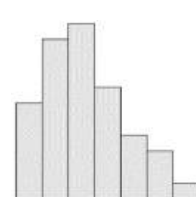


Figure 3-19

The histogram that has more of the data close to the mean will have the smallest standard deviation. The histogram that has more of the data towards the end points will have a larger standard deviation. Figure 3-16 will have the largest standard deviation since more of the data is grouped in the first and last class. Figure 3-17 will have the smallest standard deviation since more of the data is grouped in the center class which will be close to the mean in a symmetric distribution. Figures 3-18 and 3-19 are harder to compare without also having access to the mean and median to indicate skewness. However, Figure 3-19 does have smaller frequencies in the first and last three classes compared to Figure 3-18.

The correct order from smallest to largest standard deviation would be Figure 3-17, Figure 3-19, Figure 3-18, and then Figure 3-16.

One should not compare the range, standard deviation or variance of different data sets that have different units or scale.

3.2.3 Coefficient of Variation

The **coefficient of variation**, denoted by CVar or CV, is the standard deviation divided by the mean. The units on the numerator and denominator cancel with one another and the result is usually expressed as a percentage. The coefficient of variation allows you to compare variability among data sets when the units or scale are different.

$$\text{Coefficient of Variation} = \text{CVar} = \left(\frac{s}{\bar{x}} \cdot 100 \right) \%$$

Example 3-16: The following is a sample of the alcohol content and calories for 12 oz. beers. Is the alcohol content (alcohol by volume ABV) or calories more variable?

Name	Brewery	ABV	Calories in 12 oz.
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	Miller Coors	8.10%	222
O'Doul's	Anheuser Busch	0.40%	70
Coors Light	Miller Coors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Breakside Pilsner	Breakside	5.00%	158
Dark Ale	Alberta Brewing Company	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution: Type in the data to Excel and run descriptive statistics on both data sets to get the following:

<i>Alcohol Content</i>	
Mean	0.05005
Standard Error	0.006326
Median	0.05
Mode	0.047
Standard Deviation	0.020003
Sample Variance	0.0004

<i>Calories in 12 oz.</i>	
Mean	160.2
Standard Error	14.67257
Median	160
Mode	158
Standard Deviation	46.39875
Sample Variance	2152.844

Next, compute the coefficient of variation using the mean and standard deviation for both data sets.

$$\text{Alcohol Content CVar} = \left(\frac{0.020003}{0.05005} \cdot 100 \right) \% = 39.97\%$$

$$\text{Calories CVar} = \left(\frac{46.39875}{160.2} \cdot 100 \right) \% = 28.96\%$$

The alcohol content varies more than the number of calories.

There is no shortcut on the calculator or Excel for CVar, but you can find s and \bar{x} then simply divide.

Example 3-17: Average price of a latte is \$3.67 with a standard deviation of \$0.33. The average house price is \$286,000 with a standard deviation of \$20,000. Which price is more variable?

Solution: The latte and house prices are both in dollars but of different magnitudes. To compare the variability, we want to see how varied each value is from its corresponding mean.

$$\text{Latte CVar} = \left(\frac{0.33}{3.67} \cdot 100 \right) \% = 8.99\% \quad \text{House CVar} = \left(\frac{20000}{286000} \cdot 100 \right) \% = 6.99\%$$

Latte prices have a smaller standard deviation, however since the coefficient of variation is larger, the lattes have more variability in price.

It is tempting to use standard deviation or variance to compare variability since both data sets use the same units of dollars. However, the two data sets use different scales and the large values of house prices will naturally have larger values for the standard deviation and variance. When you use the coefficient of variation, the units cancel each other out in the fraction so that we can compare data sets with different units, scale or magnitude.

3.3 Measures of Placement

3.3.1 Z-Scores

A **z-score** is the number of standard deviations an observation x is above or below the mean. Z-scores are used to compare placement of a value compared to the mean.

If x is an observation from a sample, then the standardized value of x is the **z-score** where $z = \frac{x - \bar{x}}{s}$.

If x is an observation from a population, then the standardized value of x is the **z-score** where $z = \frac{x - \mu}{\sigma}$.

If the z-score is negative, x is less than the mean. If the z-score is positive, x is greater than the mean.

There are no shortcuts on the calculator in Excel for z -score, but you can find s and \bar{x} then simply subtract and divide. The number of standard deviations that a data value is from the mean is frequently used when comparing position of values. If a z -score is zero, then the data value is the same as the mean. If the z -score is one, then the data value x is one standard deviation above the mean. If the z -score is -3.5 , then the data value is three and a half standard deviations below the mean. The shaded area in Figure 3-20 represents one standard deviation from the mean.

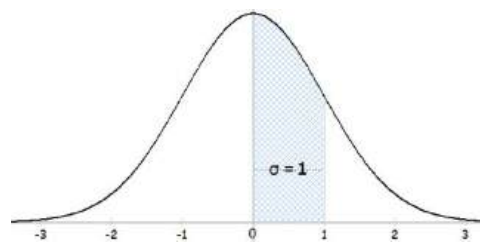


Figure 3-20

Example 3-18: For a random sample, the mean time to make a cappuccino is 2.8 minutes with a standard deviation of 0.86 minutes. Find the z -score for someone that makes their cappuccino in 4.95 minutes.

Solution: $z = \frac{x - \bar{x}}{s} = \frac{4.95 - 2.8}{0.86} = 2.5$. Their time is 2.5 standard deviations above average.

Example 3-19: On a math test, a student scored 45. The class average was 50 with a standard deviation of 3. The same student scored an 80 on a history test, and the class average was 85 with a standard deviation of 2.5. Which exam did the student perform better on compared with the rest of the class?

Solution: $z_m = \frac{45 - 50}{3} = -1.67$ $z_h = \frac{80 - 85}{2.5} = -2$

Test scores are “better” when they are larger, so whichever has the largest z -score did better. The student did better on the math test than the history test, compared to the rest of the class.

Be careful with the word “better,” depending on the context, better may be smaller rather than larger. For example, golf scores, time running a race, and cholesterol levels would be better if they were smaller values.

Example 3-20: The length of a human pregnancy has a mean of 272 days. A pregnancy lasting 281 days or more has a z -score of one. How many standard deviations above the mean is a pregnancy lasting 281 days or more?

Solution: One, since by definition the z -score is the number of standard deviations from the mean.

Example 3-21: The length of a human pregnancy has a mean of 272 days. A pregnancy lasting 281 days or more has a z -score of one. What is the standard deviation of human pregnancy length?

Solution: We know the z -score = 1 and mean = 272. Replace these two numbers in the z -score formula then solve for the standard deviation.

$$1 = \frac{281 - 272}{\sigma} \Rightarrow 1 = \frac{9}{\sigma} \Rightarrow \sigma = 9.$$

3.3.2 Percentiles

Along with the center and variability, another useful numerical measure is the ranking of a number. A **percentile** is a measure of ranking. It represents a location measurement of a data value to the rest of the values. Many standardized tests give the results as a percentile. Doctors use percentiles graphs to show height and weight standards.

Interpreting Percentiles

The p^{th} percentile is the value that separates the bottom $p\%$ from the upper $(100 - p)\%$ of the ordered (smallest to largest) data. For example, the 75th percentile is the value that separates the bottom 75% from the upper 25% of the data. There are several methods used to find percentiles. You may get different percentile values depending on which software or calculator you use. For example, Excel has two methods, both of which are not the same method as the TI calculators.

Example 3-22: What does a score of the 90th percentile represent?

Solution: This means that 90% of the scores were at or below this score. (A person did the same as or better than 90% of the test takers.)

Example 3-23: What does a score of the 70th percentile represent?

Solution: This means that 70% of the scores were at or below this score.

Percentile versus Score

If the test was out of 100 points and you scored at the 80th percentile, what was your score on the test? You do not know! All you know is that you scored the same as or better than 80% of the people who took the test. If all the scores were low, you could have still failed the test. On the other hand, if many of the scores were high you could have gotten a 95% or so.

Note there is more than one method to find percentiles. This rounding rule in Excel is not the same as used on your TI calculators.

Finding a Percentile:

Step 1: Arrange the data in order from lowest to highest.

Step 2: Substitute into the formula $i = \frac{(n+1)p}{100}$ where n = sample size and p = percentile.

Step 3A: If i is a whole number, count out i places from the lowest number to find the percentile. For example, if you get $i = 3$, then the 3rd value is the percentile.

Step 3B: If i is not a whole number, then take the weighted average between the i^{th} and $i^{\text{th}} + 1$ data value as the percentile. For example, if $i = 3.25$, this would be 25% of the distance between the 3rd and the 4th data values as the percentile. Percentile = i^{th} data value + $(i^{\text{th}} + 1$ data value – i^{th} data value) $\cdot(0.##)$ where ## is the remainder percent.

Example 3-24: Compute the 10th percentile of the random sample of 13 ages: 15, 18, 22, 25, 26, 31, 33, 35, 38, 46, 51, 53, and 95.

Solution: Data is already ordered, so next find $i = \frac{(n+1)p}{100} = \frac{14 \cdot 10}{100} = 1.4$. Since i is not a whole number use Step 3B, take the weighted average of 40% of the way between the 1st and 2nd values. This would be $15 + (18 - 15) \cdot 0.4 = 16.2$ and this is your 10th percentile, $P_{10} = 16.2$.

In Excel use =PERCENTILE.EXC(array, k) where array is the cell reference to where the data is located and k is the percentile as a decimal between 0 and 1. Note you do not have to sort the data prior to typing it in to Excel.

For this example, if you type in the data into column A, then use the formula =PERCENTILE.EXC(A1:A13, 0.1) = 16.2.

	A	B	C	D	E
1	Age				
2	15		=PERCENTILE.EXC(A2:A14,0.1)		
3	18				
4	22				
5	25				
6	26				
7	31				
8	33				
9	35				
10	38				
11	46				
12	51				
13	53				
14	95				

3.3.3 Quartiles

There are special percentiles called quartiles. Quartiles are numbers that divide the data into fourths. One fourth (or a quarter) of the data falls between consecutive quartiles. There are three quartiles Q_1 , Q_2 , and Q_3 that subsequently divide the ordered data into the 4 pieces of approximately equal size, or 25% each. Thus, 25% of the values are less than Q_1 , 25% of the data values are between Q_1 and Q_2 , 25% of the data values are between Q_2 and Q_3 , and 25% are of the data values are greater than Q_3 .

Use the dollar as an example. If we make change for a dollar, we would get four quarters to make one dollar. Hence, quarter for quartiles. The quartiles would represent the three spaces between the quarters.



To find the quartiles use the same rules as percentiles where we:

1. Arrange the observations from smallest to largest and use the previous percentile rule.
2. Then find all three quartiles.
 - $Q_1 = \text{first quartile} = 25^{\text{th}} \text{ percentile}$
 - $Q_2 = \text{second quartile} = \text{median} = 50^{\text{th}} \text{ percentile}$
 - $Q_3 = \text{third quartile} = 75^{\text{th}} \text{ percentile}$

Example 3-25: Compute all three quartiles for the random sample of 13 ages: 15, 18, 22, 25, 26, 31, 33, 35, 38, 46, 51, 53, and 95.

Solution: For the first quartile $i = \frac{(n+1) \cdot p}{100} = \frac{14 \cdot 25}{100} = 3.5$. Since i is not a whole number take the weighted average of half way between the 3rd and 4th data values $22 + (25 - 22) \cdot 0.5 = 23.5$, so $Q_1 = 23.5$.

In Excel you could use the percentile formula, but there is also a quartile formula: =QUARTILE.EXC(array, quartile), where array is the cell reference to the data and quartile is either 1, 2 or 3 for the 3 possible quartiles.

In this example we would have =QUARTILE.EXC(A1:A13, 1) = 23.5.

To find the second quartile: $i = \frac{(n+1) \cdot p}{100} = \frac{14 \cdot 50}{100} = 7$. Since i is a whole number use the 7th value for Q_2 , so $Q_2 = 33$.

Or use the Excel formula =QUARTILE.EXC(A1:A13, 2) = 33.

For the third quartile, $i = \frac{(n+1) \cdot p}{100} = \frac{14 \cdot 75}{100} = 10.5$. Since i is not a whole number use the weighted average of half way between the 10th and 11th values $46 + (51 - 46) \cdot 0.5 = 48.5$.

Or use the Excel formula =QUARTILE.EXC(A1:A13, 3) = 48.5, so $Q_3 = 48.5$.

Example 3-26: The high school graduating class of 2016 in Oregon had the following ACT quartile scores. Interpret what the number 26 under the Composite column represents.

Quartile	English	Mathematics	Reading	Science	Composite
Q_3 (75th Percentile)	25	26	27	25	26
Q_2 (50th Percentile)	21	21	22	22	21
Q_1 (25th Percentile)	16	17	17	18	17

https://www.act.org/content/dam/act/unsecured/documents/P_38_389999_S_S_N00_ACT-GCPR_Oregon.pdf

Solution: From the report we can see that the third quartile for composite score is 26, this means that 75% of Oregon students that took the ACT exam scored 26 or below.

Other Types of Percentiles

Quintiles break a data set up into five equal pieces. We will not be using these, but be aware that percentiles come in different forms. Deciles break a data set up into ten equal pieces and are found using the percentile rule. For example, the 6th decile = $D_6 = 60^{\text{th}}$ percentile.

Use the dollar as an example. If we make change for a dollar, we would get ten dimes to make one dollar. Hence, a dime might help you remember deciles.



Example 3-27: Earlier in Example 2-14, we made an ogive using Excel with the following sample of 35 ages shown in Figure 3-21. Use the ogive to find the age for the 8th decile.

46	47	49	25	46	22	42
24	46	40	39	27	25	30
33	27	46	21	29	20	26
25	25	26	35	49	33	26
32	31	39	30	39	29	26

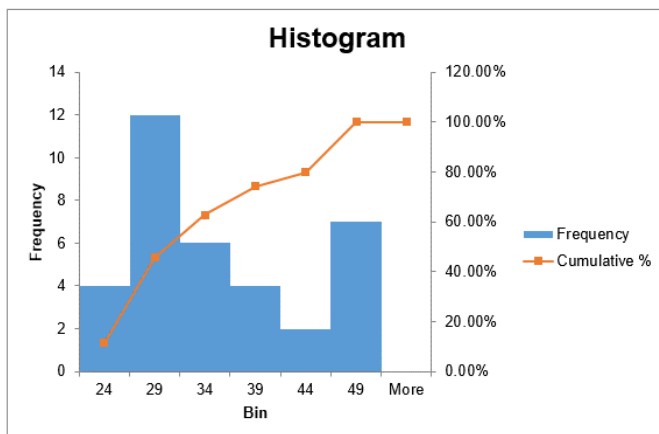


Figure 3-21

Solution: The cumulative % represents the cumulative relative frequencies which are equivalent to the percentiles for each class. The red line is the ogive and the percentiles correspond to the vertical axis on the right side.

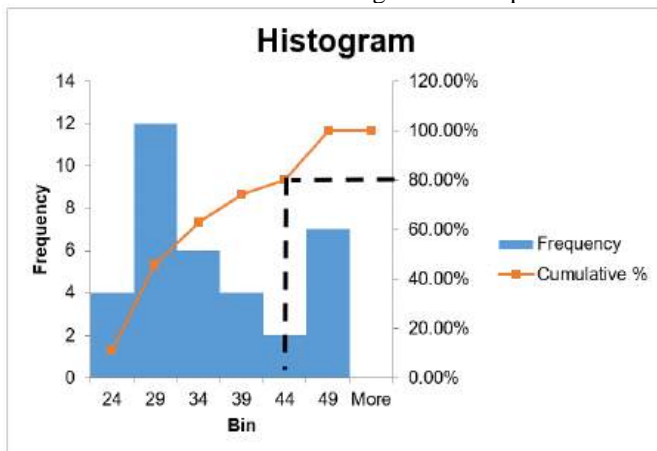


Figure 3-22

If we wanted to know what age the 80th percentile was in the sample, we could use the ogive to get an approximate value. Starting on the right at 80% make a horizontal line until you hit the red cumulative % line, and then make a

vertical line from there down to the axis to get the approximate age. See Figure 3-22. In this case, the 80th percentile = 8th decile would be approximately 44.

3.3.4 Five Number Summary & Outliers

If you record the quartiles together with the minimum and maximum values from a data set, you have five numbers. These five numbers are known as the five-number summary consisting of the minimum, the first quartile (Q_1), the median (Q_2), the third quartile (Q_3), and the maximum (in that order).

The **interquartile range**, IQR, is the difference between the first and third quartiles, Q_1 and Q_3 . Half of the data (50%) falls in the interquartile range. If the IQR is “large,” the data is spread out and if the IQR is “small,” the data is closer together.

The interquartile range (IQR) = $Q_3 - Q_1$

Not only does the IQR give a range of the middle 50% of the data, but is also used to determine outliers in a sample.

To find these outliers we first find what are called a lower and upper limit sometimes called fences.

The lower limit, or inner fence, is $Q_1 - (1.5 \cdot \text{IQR})$. Any values that are less than the lower limit is considered an outlier.

Similarly, the upper limit, or outer fence, is $Q_3 + (1.5 \cdot \text{IQR})$. Any values that are more than the upper limit are considered outliers.

If all the numbers in the sample fall between the lower and upper limit, including the endpoints, then there are no outliers in the sample. Any values outside these limits would be considered outliers.

3.3.5 Modified Box-and-Whisker Plot

A boxplot (or box-and-whisker plot) is a graphical display of the five-number summary. A boxplot can be drawn vertically or horizontally. The modified boxplot shows outliers, whereas a regular boxplot does not show outliers. The basic format of the plot is a box drawn from Q_1 to Q_3 , a vertical line drawn inside the box for the median, and horizontal lines (called whiskers) extending out of the middle of each end of the box to the minimum and maximum. The box should not touch the number line. The modified boxplot extends the left line to the smallest value greater than the lower fence, and extends the right line to the largest value less than the upper fence. Dots, circles or asterisks represent any outlier. We will make modified boxplots for this course. Like always, label the tick marks on the number line and give the graph a title.

A **boxplot** is a graph of the 5-number summary, see Figure 3-23.

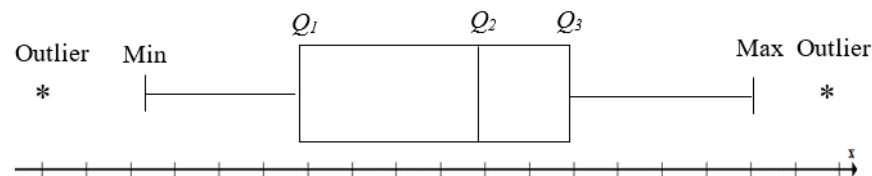


Figure 3-23

It is important to note that when you are making the boxplot the limits for finding outliers are not graphed in the plot, they were only used to find the outliers. The whiskers would go to the next largest (or smallest) value in the data set after you removed the outlier(s).

If the sample has a symmetrical distribution, then the boxplot will be visibly symmetrical. If the data distribution has a left skew or a right skew, the line on that side of the boxplot will be visibly long in the direction of skewness. If the four quartiles are all about the same distance apart, then the data are likely a near uniform distribution. If a boxplot is symmetrical, and both outside lines are noticeably longer than the Q_1 to median and median to Q_3 distance, the distribution is then probably bell-shaped.

Example 3-28: Make a modified box-and-whisker plot for the random sample of 13 ages: 15, 18, 22, 25, 26, 31, 33, 35, 38, 46, 51, 53, and 95.

Solution: Use Excel to compute the three quartiles as:

$$Q_1 = \text{QUARTILE.EXC}(A1:A13, 1) = 23.5$$

$$Q_2 = \text{QUARTILE.EXC}(A1:A13, 2) = 33$$

$$Q_3 = \text{QUARTILE.EXC}(A1:A13, 3) = 48.5$$

The 5-number summary values are 15, 23.5, 33, 48.5 and 95. Each of these numbers will need to be incorporated into the box-and-whisker plot, and any outliers to graph the modified box-and-whisker plot.

To find the outliers, first find the IQR, and then find the lower and upper limits. The IQR = $Q_3 - Q_1 = 48.5 - 23.5 = 25$. The lower limit is $Q_1 - (1.5 \cdot \text{IQR}) = 23.5 - 1.5(25) = -14$. The upper limit is $Q_3 + (1.5 \cdot \text{IQR}) = 48.5 + 1.5(25) = 86$.

Any value in our data set that is not between the lower and upper limits $[-14, 86]$ is an outlier. By observation, we have one number that is outside the range so the outlier is 95.

The whiskers would be drawn to the next largest (or smallest) value in the data set after you removed the outlier(s). For this example, the next largest value in the data set is 53.

Now put that all together to get the following graph in Figure 3-24.

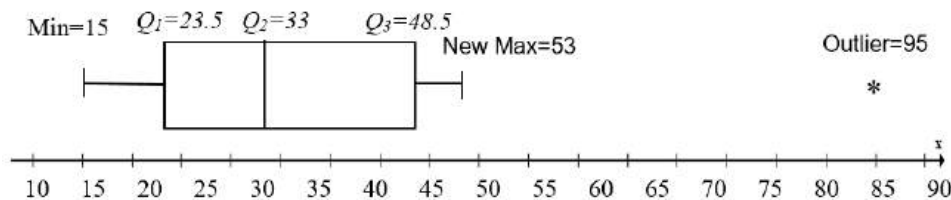
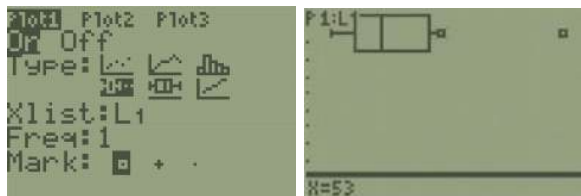


Figure 3-24

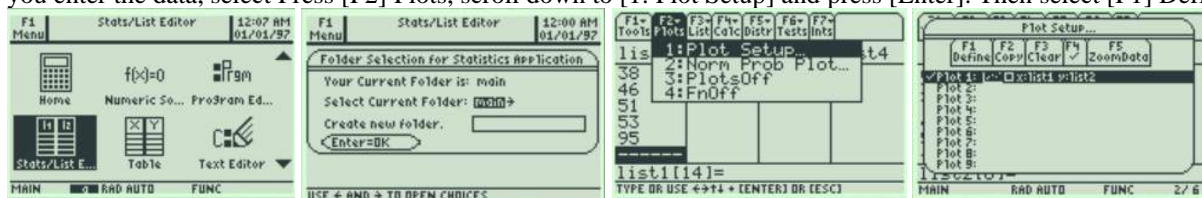
The TI-calculator and newer versions of Excel will make a modified boxplot. Note that the quartile rules used in the TI calculators are slightly different than in Excel and what is presented in this content. They do not use a weighted mean between values, just half way between values.

TI-84: First, enter your data in to list 1. Next, press $2^{\text{nd}} > \text{STAT PLOT}$, then choose the first plot. Note that your calculator may say Plot1...Off or show a different type of graph than the screenshot. Using your arrow keys, turn the plot on. Choose the modified boxplot which is the first of the two boxplot options with the small dots to the right of the whiskers representing outliers. Make sure your Xlist: is on L1, keep frequency as a one, and any mark will work, but the square shows up best.

Here is screen shot from the calculator for the last example. You can use Trace on the boxplot from the TI-84 calculator below to see where each quartile, whisker and outlier are.

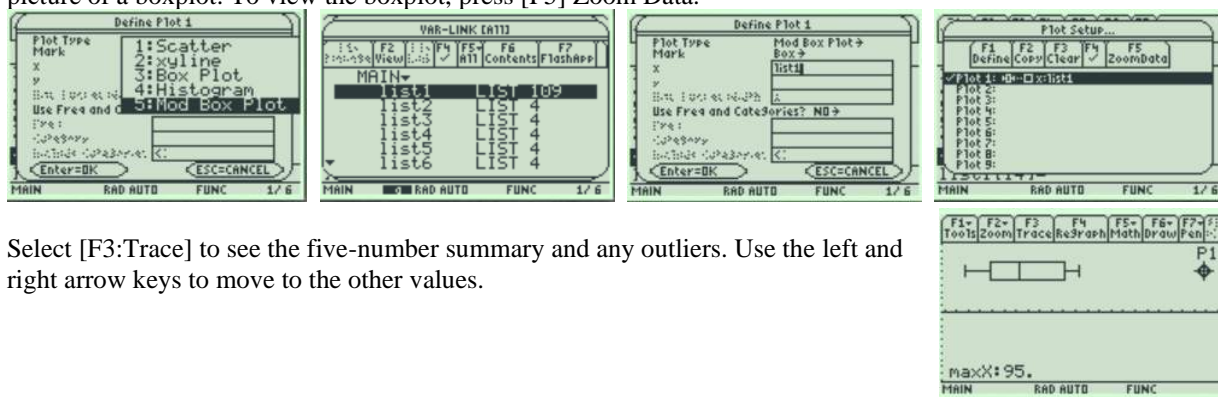


TI-89: Enter the data into the Stat/List editor under list 1. Press [APP] then scroll down to Stat/List Editor; on the older style TI-89 calculators, go into the Flash/App menu, and then scroll down the list. Make sure the cursor is in the list, not on the list name, and type the desired values pressing [ENTER] after each one. To clear a previously stored list of data values, arrow up to the list name you want to clear, press [CLEAR], and then press enter. After you enter the data, select Press [F2] Plots, scroll down to [1: Plot Setup] and press [Enter]. Then select [F1] Define.



Use your arrow keys to select Mod Box Plot for Type, and then scroll down to the x-variable box. Press [2nd] [Var-Link] this key is above the + sign. Then arrow down until you find your List1 name under the Main file folder.

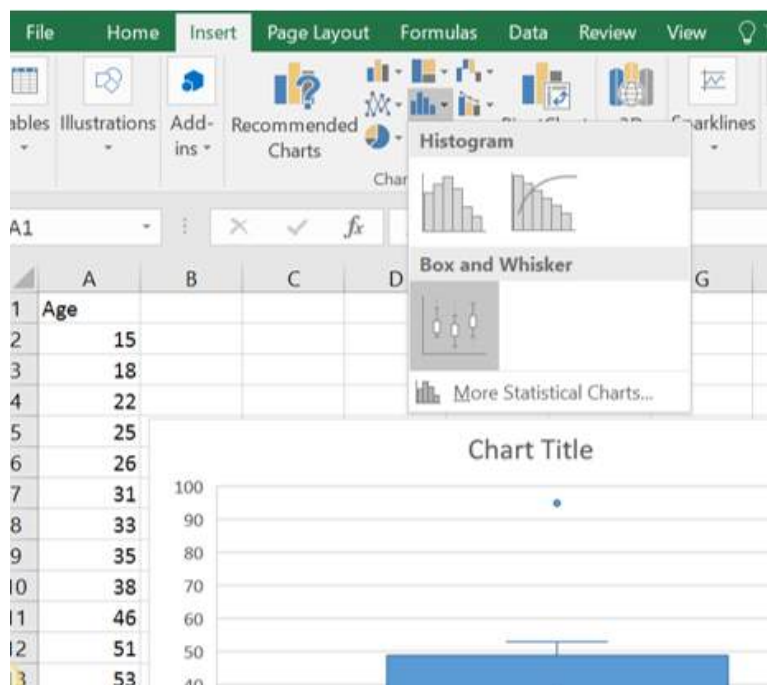
Then press [Enter] and this will bring the name List1 back to the menu. You will now see that Plot1 has a small picture of a boxplot. To view the boxplot, press [F5] Zoom Data.



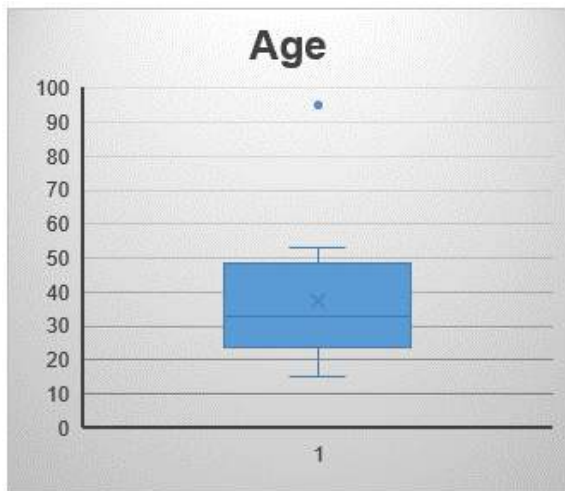
Select [F3:Trace] to see the five-number summary and any outliers. Use the left and right arrow keys to move to the other values.

Excel: Note this example is on a PC running Excel 2019. Older versions of Excel may not have a boxplot option. First, type your sample data into column A in any order. Highlight the data, and then select the Insert tab. Under the graphing options, the picture shaped like a histogram called statistical charts, select Box and Whisker.

You can change the formatting options and add the chart title as needed.



Below is the finished Excel boxplot. Note that Excel does a vertical boxplot rather than the traditional horizontal number line.



Excel marks an \times just above the median where the mean would fall. Usually, one would not include the mean on a boxplot. Remember that when the mean is greater than the median the distribution is usually skewed to the right.

When a boxplot has outliers only on one side, then we can also say the distribution is skewed in the direction of the outlier, which also indicates that these ages are skewed to the right.

Side by side boxplots are great at comparing quartiles and distribution shapes for several samples using the same units and scale.

Example 3-29: There are four franchises in different parts of town. Compare the weekly sales over a year for each of the four franchises. Compare the boxplots shown in Figure 3-25.

Solution: We can see that Store 2 in Figure 3-25 has the highest sales since the median for this store is higher than the third quartile for all the other stores. Store 2 also has sales that are more consistent from week to week with the smaller range and has a symmetric distribution. The lowest performing store, Store 1, has the lowest median sales and is skewed to the right. Both Stores 3 and 4 have moderate sales and are skewed left.



Figure 3-25

3.4 Correlation and Linear Regression

We are often interested in the relationship between two variables. This chapter determines whether a linear relationship exists between sets of quantitative data and making predictions for a population. For instance, the relationship between the number of hours of study time and an exam score, or calories burned and time exercising.

A predictor variable (also called the independent or explanatory variable; usually we use the letter x) explains or causes changes in the response variable. The predictor variable can be manipulated or changed by the researcher.

A response variable (also called the dependent variable; usually we use the letter y) measures the outcome of a study. The different outcomes for a dependent variable are measured or observed by the researcher. For instance, suppose we are interested in how much time spent studying affects the scores on an exam. In this study, study time is the predictor variable, and exam score is the response variable.

In data from an experiment, it is much easier to know which variable we should use for the independent and dependent variables. This can be harder to distinguish in observational data. Think of the dependent variable as the variable that you are trying to learn about.

If we were observing the relationship between unemployment rate and economic growth rate, it may not be clear which variable should be x and y . Do we want to predict the unemployment rate or the economic growth rate? One should never jump to a cause and effect reasoning with observational data. Just because there is a strong relationship between unemployment rate and economic growth rate does not mean that one causes the other to change directly. There may be many other contributing factors to both of these rates changing at the same time, such as retirements or pandemics.

3.4.1 Correlation

Correlation is a way to see if two things are related or connected to each other. Correlation is like finding a pattern or a connection between two quantitative variables. A correlation coefficient is a number that tells us how strong the relationship is between two things. It is a statistical measure that ranges from -1 to 1 , with 0 indicating no correlation, -1 indicating a perfect negative correlation (meaning that as one variable increases, the other variable decreases), and 1 indicating a perfect positive correlation (meaning that as one variable increases, the other variable also increases).

The correlation coefficient is useful because it allows us to quantify the strength and direction of the relationship between two variables. This can help us make predictions and better understand how the variables are related. However, it's important to note that correlation does not necessarily imply causation, meaning that just because two variables are correlated does not necessarily mean that one causes the other. Sometimes, things might seem like they're related, but there might not actually be a correlation. Other times, there might be a really strong correlation, but another variable is causing the two correlated variables to change.

Correlation Coefficient

The sample correlation coefficient measures the direction and strength of the linear relationship between two quantitative variables. There are several different types of correlations. We will be using the Pearson Product Moment Correlation Coefficient (PPMCC). We will just use the lower-case r for short when we want to find the sample correlation coefficient.

Interpreting the Correlation Coefficient:

- A positive r indicates a positive association (positive linear slope).
- A negative r indicates a negative association (negative linear slope).
- r is always between -1 and 1 , inclusive.
- If r is close to 1 or -1 , there is a strong linear relationship between x and y .
- If r is close to 0 , there is a weak linear relationship between x and y . There may be a non-linear relation or there may be no relation at all.
- Like the mean, r is strongly affected by outliers.

When r is equal to -1 or 1 , all the dots in the scatterplot line up in a straight line. As the points disperse, r gets closer to zero. The correlation tells the direction of a linear relationship only. It does not tell you what the slope of the line is, nor does it recognize nonlinear relationships.

For instance, in Figure 3-26, there are three scatterplots overlaid on the same set of axes. All three data sets would have $r = 1$ even though they all have different slopes.

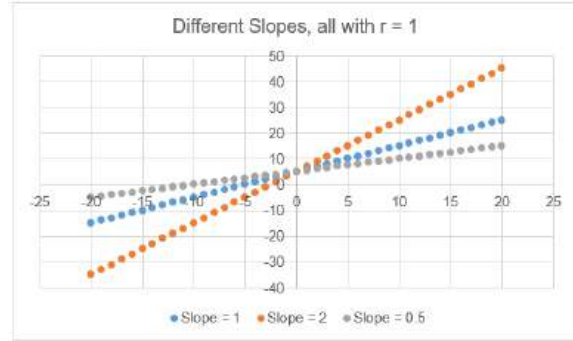


Figure 3-26

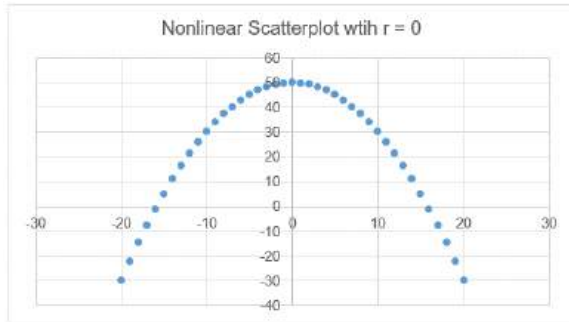


Figure 3-27

For the next example in Figure 3-27, $r = 0$ would indicate no linear relationship, however there is clearly a non-linear pattern with the data.

Figure 3-28 shows a correlation $r = 0.874$, which is pretty close to one, indicating a strong linear relationship. However, there is an outlier, called a leverage point, which is inflating the value of the slope.

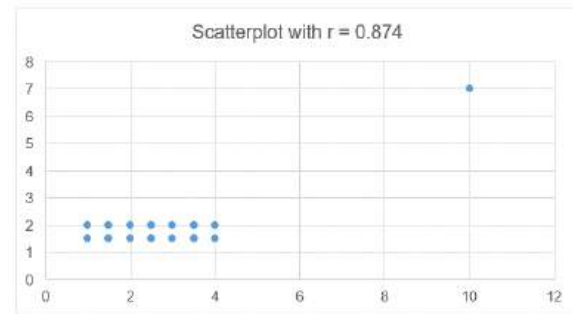


Figure 3-28

If you remove the outlier then $r = 0$, and there is no up or down trend to the data.

Figure 3-29 gives examples of correlations with their corresponding scatterplots.

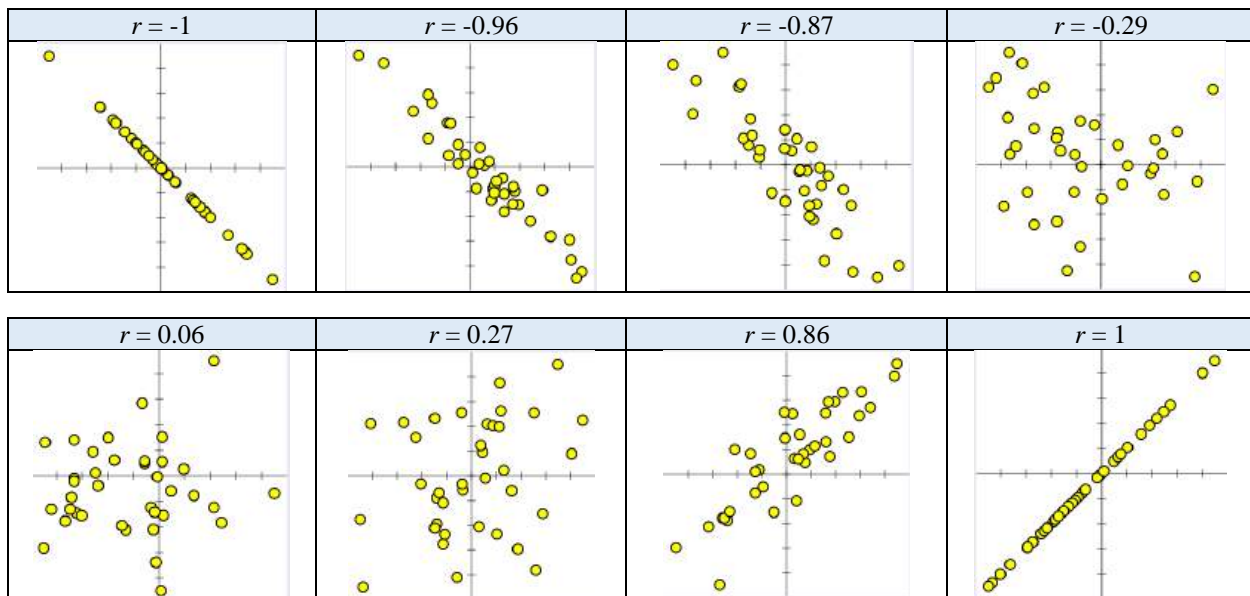
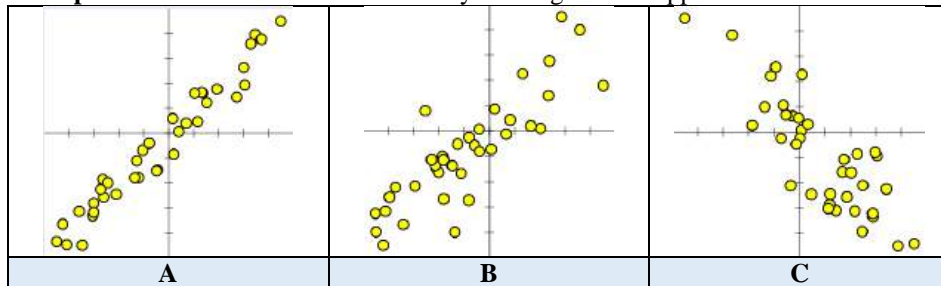


Figure 3-29

When you have a correlation that is very close to -1 or 1 , then the points on the scatter plot will line up in an almost perfect line. The closer r gets to 0 the more scattered your points become.

Example 3-30: Take a moment and see if you can guess the approximate value of r for the scatter plots below.



Solution:

In Example 3-30: Scatter Plot A: $r = 0.98$, Scatter Plot B: $r = 0.85$, Scatter Plot C: $r = -0.85$.

Calculating a Correlation

To calculate the correlation coefficient by hand we would use the following formula.

Sample Correlation Coefficient

$$r = \frac{\sum((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}} = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}}$$

Instead of doing all of these sums by hand we can use the output from summary statistics. Recall that the formula for a variance of a sample is $s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$. If we were to multiply both sides of the sample variance formula by the degrees of freedom, we would get $\sum(x_i - \bar{x})^2 = (n - 1)s_x^2$.

We use these sums of squares $\sum(x_i - \bar{x})^2$ frequently, so for shorthand, we will use the notation $SS_{xx} = (n - 1)s_x^2$. The same would hold true for the y variable, just changing the letter the variance of y would be $s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$, therefore $SS_{yy} = (n - 1)s_y^2$. Some texts will use a single S for sum of squares, or a single x or y in the subscript as $SS_{xx} = SS_x = S_{xx}$, not to be confused with the sample variance s_x^2 , or the sample standard deviation s_x .

The numerator of the correlation formula is taking in the distance each data point is horizontally from the mean of the x values, times the vertical distance each point is away from the mean of the y values. This is time-consuming to find so we will use an algebraically equivalent formula $\sum((x_i - \bar{x})(y_i - \bar{y})) = \sum(xy) - n \cdot \bar{x} \cdot \bar{y}$ and for short we will use the notation $SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y}$.

To start each problem, use descriptive statistics to find the sum of squares.

Sum of Squares (SS) used for Correlation and Linear Regression

$$SS_{xx} = (n - 1)s_x^2$$

$$SS_{yy} = (n - 1)s_y^2$$

$$SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y}$$

Example 3-31: Use the following data to calculate the correlation coefficient.

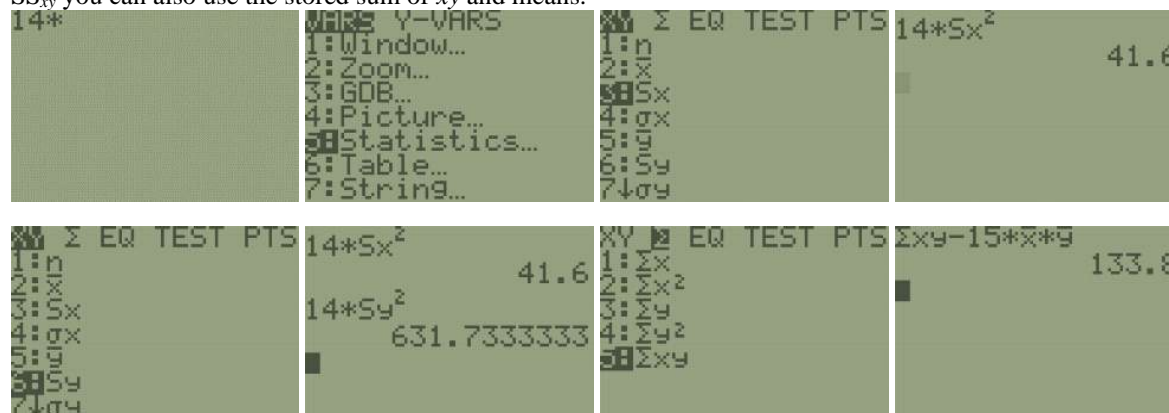
Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: We could show all the work the long way by hand using the shortcut formula. On the TI-84 press the [STAT] key and then the [EDIT] function, type the x values into L_1 and the y values into L_2 . Press the [STAT] key again and arrow over to highlight [CALC], select 2-Var Stats, then press [ENTER]. This will return the descriptive stats.

The TI calculator can run descriptive statistics and quickly get everything we need to find the sum of squares. Go to STAT > CALC > 2-Var Stats. For TI-83, you may need to enter your list names separated by a comma, for example 2-Var Stats L_1, L_2 then press enter. On the TI-89, open the Stats/List Editor. Enter all x -values in one list. Enter all corresponding y -values in a second list. Press F4, then select 2-Var Stats, then press [ENTER]. This will return the descriptive stats. Use the down arrow to see everything.



Once you do this the statistics are stored in your calculator so you can use the VARS key, go to Statistics, then select the standard deviation for x , repeat for the y -variable. This will reduce rounding errors by using exact values. For the SS_{xy} you can also use the stored sum of xy and means.



This gives:

$$SS_{xx} = (n - 1)s_x^2 = (15 - 1)1.723783215^2 = 41.6$$

$$SS_{yy} = (n - 1)s_y^2 = (15 - 1)6.717425811^2 = 631.7333$$

$$SS_{xy} = \Sigma(xy) - n \cdot \bar{x} \cdot \bar{y} = 20087 - (15 \cdot 16.6 \cdot 80.133333) = 133.8$$

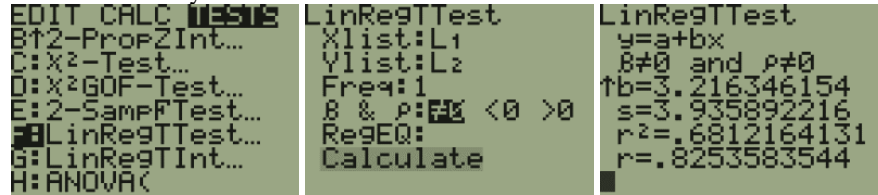
Note that both SS_{xx} and SS_{yy} will always be positive, but SS_{xy} could be negative or positive. For the TI-89, you will see the sum of squares at the very bottom of the descriptive statistics $\Sigma(x - \bar{x})^2 = 41.6$ and $\Sigma(y - \bar{y})^2 = 631.7333$.

To find the correlation, substitute the three sums of squares into the formula to get: $r = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}} = \frac{133.8}{\sqrt{(41.6 \cdot 631.7333)}}$
 $= 0.8524$. Try this now on your calculator to see if you are getting your order of operations correct.

For our example, $r = 0.8524$ is close to 1, therefore it looks like there is positive linear relationship between the number of hours studying for an exam and the grade on the exam.

Most software has a built-in correlation function.

TI-84: Press the [STAT] key and then the [EDIT] function, type the x values into L_1 and the y values into L_2 . Press the [STAT] key again and arrow over to highlight [TEST], select LinRegTTest, then press [ENTER]. The default is Xlist: L_1 , Ylist: L_2 , Freq:1, β and ρ : $\neq 0$. Arrow down to Calculate and press the [ENTER] key. Scroll down to the bottom until see you r .



TI-89: On the TI-89, open the Stats/List Editor. Enter all x -values in one list. Enter all corresponding y -values in a second list. Press F6, then select LinRegTTest, then press [ENTER]. Scroll down to the bottom of the output to see r .



Excel:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14				
2	Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76				=CORREL(B1:P1,B2:P2)

$$r = \text{CORREL}(\text{array1}, \text{array2}) = \text{CORREL}(B1:P1, B2:P2) = 0.8254$$

Correlation is Not Causation

Just because two variables are significantly correlated does not imply a cause and effect relationship. There are several relationships that are possible. It could be that x causes y to change. You can actually swap x and y in the fields and get the same r value and y could be causing x to change. There could be other variables that are affecting the two variables of interest. For instance, you can usually show a high correlation between ice cream sales and home burglaries. Selling more ice cream does not “cause” burglars to rob homes. More home burglaries do not cause more ice cream sales. We would probably notice that the temperature outside may be affecting both ice cream sales to increase and people leaving their windows open. This third variable is called a lurking variable and causes both x and y to change, making it look like the relationship is just between x and y .

There are also highly correlated variables that seemingly have nothing to do with one another. These seemingly unrelated variables are called spurious correlations.

The following website has some examples of spurious correlations (a slight caution that the author has some gloomy examples): <http://www.tylervigen.com/spurious-correlations>.

Note that the plots on this site are time series, graphing two different variables. We don’t want to find the correlation between time and each variable, but instead use the two variable measurements paired up at each point of time.

A Time series plot is not ideal for correlation or regression for several reasons:

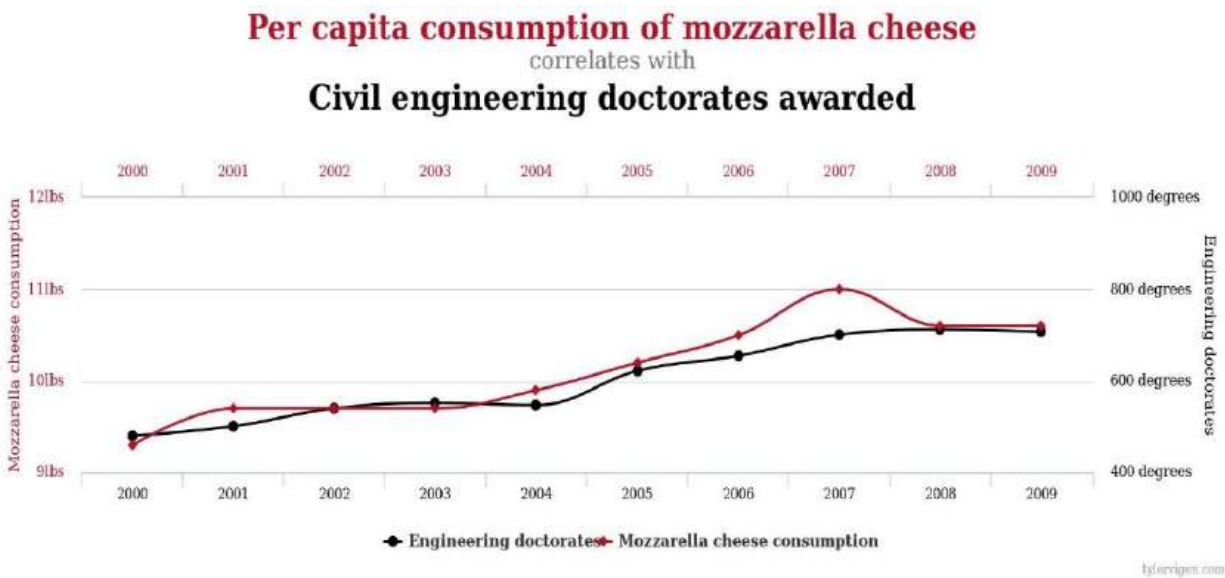
- Time dependence: Time series analysis assumes data are ordered in time, where each observation is dependent on the preceding ones. On the other hand, regression analysis assumes that the data points are independent and not influenced by their order, treating them as random and unrelated to their position in the dataset. Linear regression assumes that observations are independent.

- Trend and seasonality: Time series data often includes seasonal trends or patterns, which linear regression does not adequately capture. Linear regression assumes a constant relationship between the dependent and independent variables, while time series data may have changing patterns over time.
- Autocorrelation: Time series data may exhibit autocorrelation (the correlation between observations), leading to incorrect estimates.
- Model interpretation: In linear regression, the coefficients represent the change in the dependent variable for a one-unit change in the independent variable. In time series analysis, interpreting coefficients in this manner may not be meaningful.

Instead of linear regression, specialized time series analysis techniques should be used which are not discussed in this text.

You can play around with different variables in the <http://tylervigen.com/discover> page.

Figure 3-30 is one of their examples:



Spurious Correlations. (6/25/2020) Retrieved from http://tylervigen.com/view_correlation?id=28726.

Figure 3-30

Example 3-32: Use the data from Figure 3-30 to make a scatterplot and calculate the correlation coefficient.

Solution: If we were to take out each pair of measurements by year from the time-series plot in Figure 3-30, we would get the following data.

Year	Engineering Doctorates	Mozzarella Cheese Consumption
2000	480	9.3
2001	501	9.7
2002	540	9.7
2003	552	9.7
2004	547	9.9
2005	622	10.2
2006	655	10.5
2007	701	11
2008	712	10.6
2009	708	10.6

Using Excel, we get the scatterplot shown in Figure 3-31.

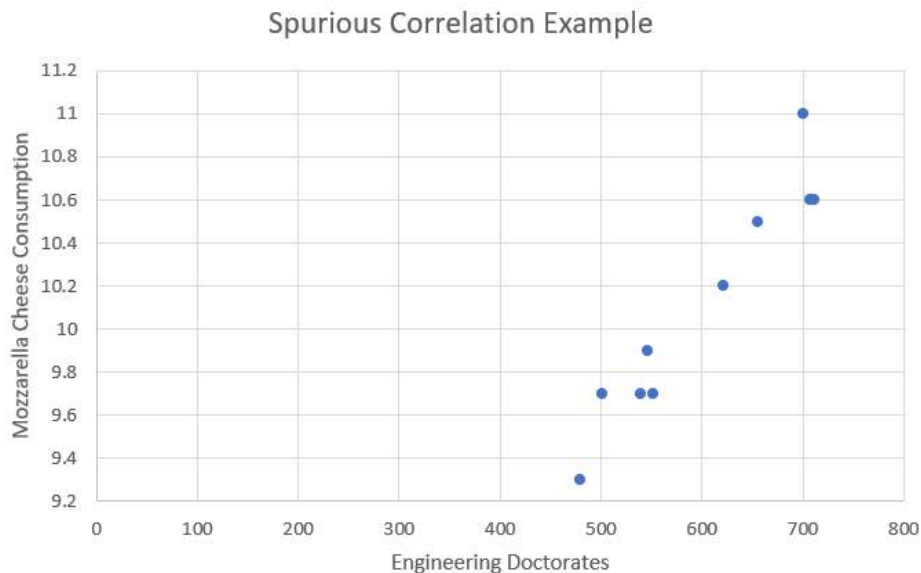


Figure 3-31

The correlation is $r = 0.9586$. There is strong correlation between the number of engineering doctorate degrees earned and mozzarella cheese consumption over time, but earning your doctorate degree does not cause one to eat more cheese. Nor does eating more cheese cause one to earn a doctorate degree. Most likely these items are both increasing over time and therefore show a spurious correlation to one another.

When two variables are correlated, it does not imply that one variable causes the other variable to change.

Correlation is causation is an incorrect assumption that because something correlates, there is a causal relationship. Causality is the area of statistics that is most commonly misused, and misinterpreted, by people. Media, advertising, politicians and lobby groups often leap upon a perceived correlation, and use it to “prove” their own agenda. They fail to understand that, just because results show a correlation, there is no proof of an underlying causality. Many people assume that because a poll, or a statistic, contains many numbers, it must be scientific, and therefore correct. The human brain is built to try and subconsciously establish links between many pieces of information at once. The brain often tries to construct patterns from randomness, and may jump to conclusions, and assume that a cause and effect relationship exists. Relationships may be accidental or due to other unmeasured variables. Overcoming this tendency to jump to a cause and effect relationship is part of academic training for students and in most fields, from statistics to the arts.

3.4.2 Linear Regression

A linear regression is a straight line that describes how the values of a response variable y change as the predictor variable x changes.

The equation of a line, relating x to y uses the slope-intercept form of a line, but with different letters than what you may be used to in a math class. We let b_0 represent the sample y -intercept (the value of y when $x = 0$), b_1 represent the sample slope (rise over run), and \hat{y} represent the predicted value of y for a specific value of x . The equation is written as $\hat{y} = b_0 + b_1x$.

Some textbooks and the TI calculators use the letter a to represent the y -intercept and b to represent the slope and the equation is written as $\hat{y} = a + bx$. These letters are just symbols representing the placeholders for the numeric values for the y -intercept and slope.

If we were to fit the best line that was closest to all the points on the scatterplot, we would get what we call the “line of best fit,” also known as the “regression equation” or “least squares regression line.”

Figure 3-32 is a scatterplot with just five points.

Figure 3-33 shows the least-squares regression line of y on x is the line that minimizes the squared vertical distance from all of the data. If we were to fit the line that best fits through the points, we would get the following picture.

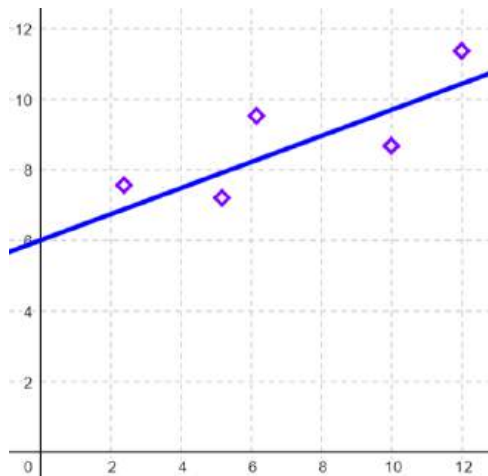


Figure 3-33

What we want to look for is the minimum of the squared vertical distance between each point and the regression equation, called a residual. This is where the name of least squares regression line comes from. Figure 3-34 shows the squared residuals.

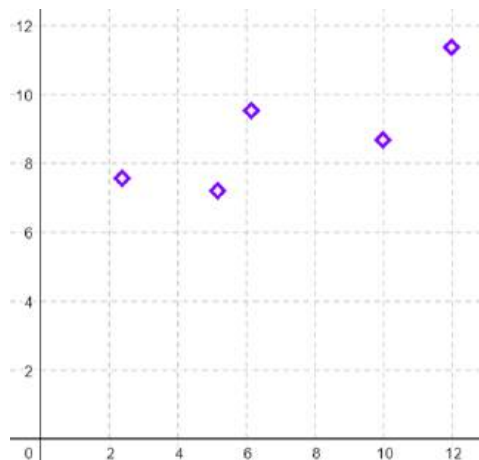


Figure 3-32

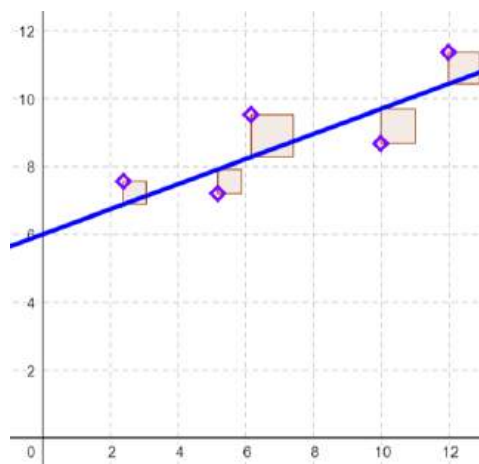


Figure 3-34

To find the slope and y-intercept for the equation of the least-squares regression line $\hat{y} = b_0 + b_1x$ we use the following formulas: slope $= b_1 = \frac{SS_{xy}}{SS_{xx}}$, y-intercept: $b_0 = \bar{y} - b_1\bar{x}$.

To compute the least squares regression line, you will need to first find the slope. Then substitute the slope into the following equation of the y-intercept: $b_0 = \bar{y} - b_1\bar{x}$, where \bar{x} = the sample mean of the x 's and \bar{y} = the sample mean of the y 's.

Once we find the equation for the regression line, we can use it to estimate the response variable y for a specific value of the predictor variable x . Note: we would only want to use the regression equation for prediction if there is a significant correlation between x and y . We will use inferential statistics later in the text to determine how far r needs to be away from zero to be statistically significant.

Example 3-33: Use the following data to find the line of best fit. Graph the regression equation on a scatterplot for the data. Then predict a student's grade when they studied 19 hours.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: Start with finding the 2-Var Stats and sum of squares as shown in the steps for correlation.

$$SS_{xx} = (n - 1)s_x^2 = (15 - 1)1.723783215^2 = 41.6$$

$$SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y} = 20087 - (15 \cdot 16.6 \cdot 80.133333) = 133.8$$

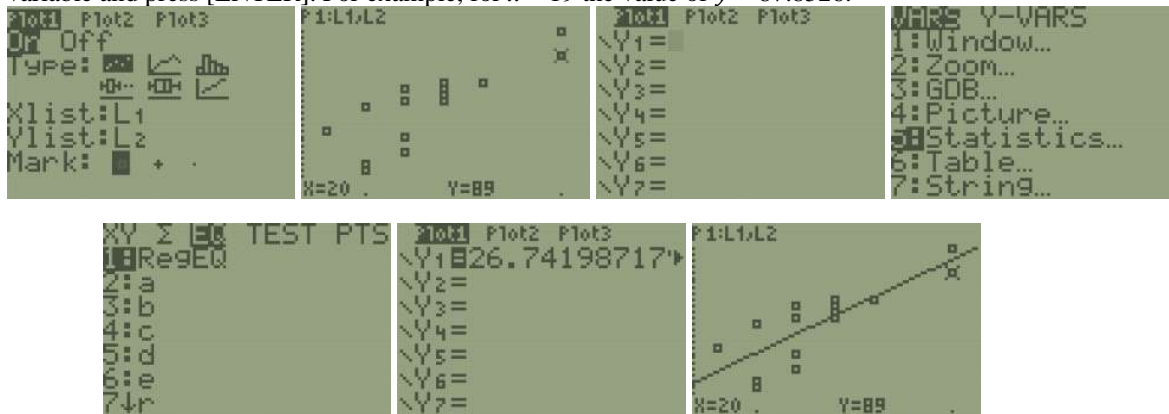
Calculate the slope: $b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{133.8}{41.6} = 3.216346$.

Calculate the y-intercept: $b_0 = \bar{y} - b_1 \cdot \bar{x} = 80.133333 - 3.216346 \cdot 16.6 = 26.742$.

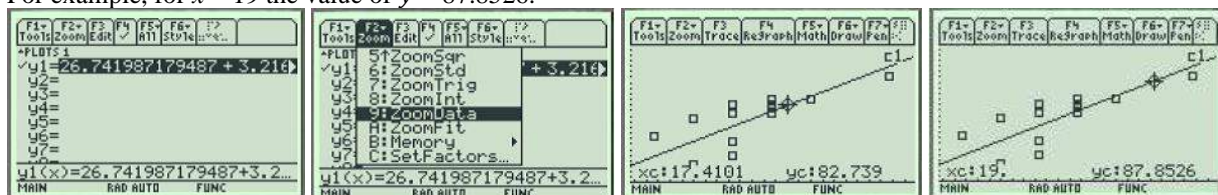
Put these numbers back into the regression equation and write your answer as: $\hat{y} = 26.742 + 3.216346x$.

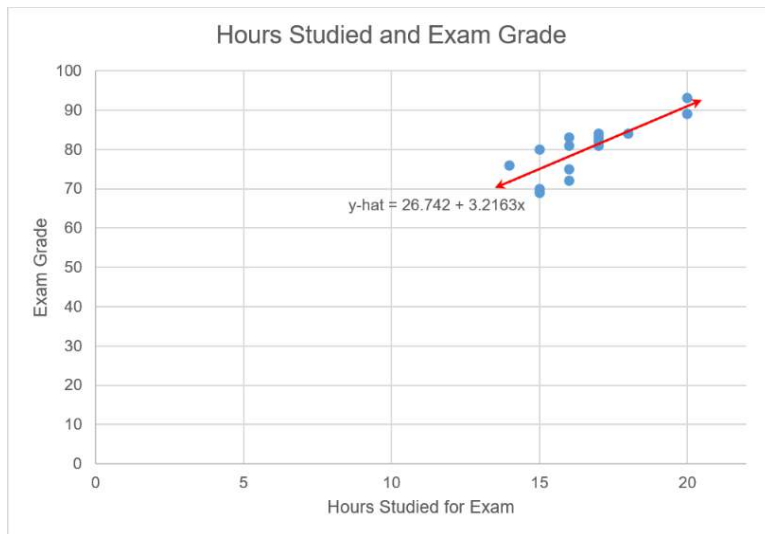
Adding the Regression Line to the Scatterplot

TI-84: Make a scatterplot using the directions from the previous section. Turn your STAT scatter plot on. Press [Y=] and clear any equations that are in the y-editor. Into Y1, enter the least-squares regression equation manually as found above. Or, press the VARS key, go to option 5: Statistics, arrow over to EQ for equation, then choose the first option RegEQ. This will bring the equation over to the Y= menu without rounding error. Press [GRAPH]. You can press [TRACE] and use the arrow keys to scroll left or right. Pressing up or down on the arrow keys will change between tracing the scatterplot and the regression line. You can use the regression line to predict values of the response variable for a given value of the explanatory variable. While tracing the regression line type the value of the explanatory variable and press [ENTER]. For example, for $x = 19$ the value of $\hat{y} = 87.8526$.



TI-89: Make a scatterplot and find the regression line using the directions in the previous section. If you press [♦] then [F1] (Y=) you will notice the regression equation has been stored into y1 in the y-editor. Press [F2] **Trace** and use the left and right arrow keys to trace along the plot. Use the up and down arrow keys to toggle between the regression line and the scatterplot. You can use the regression line to predict values of the response variable for a given value of the explanatory variable. While tracing the regression line type the value of the explanatory variable and press [ENTER]. For example, for $x = 19$ the value of $\hat{y} = 87.8526$.





Interpreting the y-intercept coefficient: When $x = 0$, note that $\hat{y} = 26.742$, this means that we would expect a failing midterm score of 26.742 for students who had studied zero hours.

Interpreting the slope coefficient: For each additional hour studied for the exam, we would expect an increase in the midterm grade of 3.2163 points.

In general, when interpreting the slope coefficient, for each additional 1 unit increase in x , the predicted \hat{y} value will change by b_1 units.

If you were to predict a student's exam grade when they studied 19 hours, substitute $x = 19$ into the regression equation and you would get a predicted grade of $\hat{y} = 26.742 + 3.216346 \cdot 19 = 87.8526$.

Extrapolation is the use of a regression line for prediction far outside the range of values of the independent variable x . As a general rule, one should not use linear regression to estimate values too far from the given data values. The further away you move from the center of the data set, the more variable results become. For instance, we would not want to estimate a student's grade for someone that studied way less than 14 hours or more than 20 hours.

Example 3-34: The following data gives a sample of countries from 2019. Use the independent variable as the percentage share of the country's population that is below the \$3.65 a day poverty line. Use the country's life expectancy for the dependent variable. Use Excel to answer the following questions.

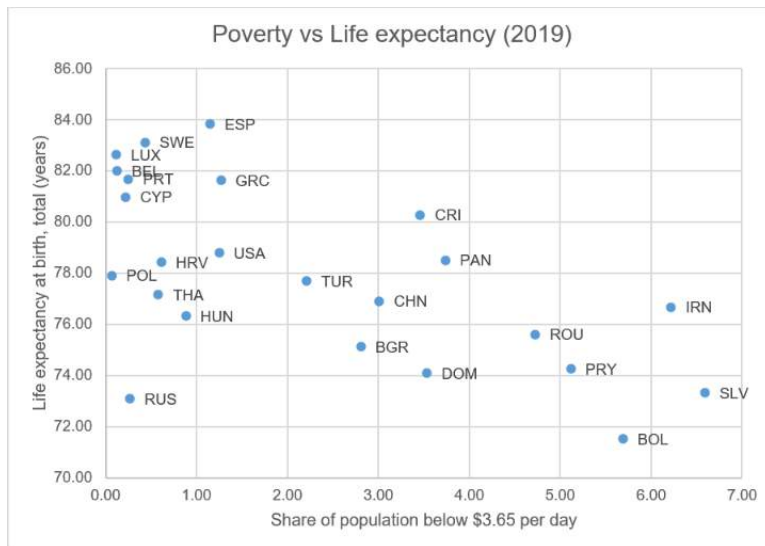
Country	Code	Year	\$3.65 a day - share of population below poverty line	Life expectancy at birth, total (years)
Belgium	BEL	2019	0.13	82.00
Bolivia	BOL	2019	5.69	71.51
Bulgaria	BGR	2019	2.81	75.11
China	CHN	2019	3.01	76.91
Costa Rica	CRI	2019	3.46	80.28
Croatia	HRV	2019	0.61	78.42
Cyprus	CYP	2019	0.22	80.98
Dominican Republic	DOM	2019	3.54	74.08
El Salvador	SLV	2019	6.59	73.32
Greece	GRC	2019	1.27	81.64
Hungary	HUN	2019	0.88	76.32
Iran	IRN	2019	6.22	76.68
Luxembourg	LUX	2019	0.12	82.64
Panama	PAN	2019	3.74	78.51
Paraguay	PRY	2019	5.12	74.25
Poland	POL	2019	0.07	77.90
Portugal	PRT	2019	0.25	81.68

Country	Code	Year	\$3.65 a day - share of population below poverty line	Life expectancy at birth, total (years)
Romania	ROU	2019	4.73	75.61
Russia	RUS	2019	0.27	73.08
Spain	ESP	2019	1.15	83.83
Sweden	SWE	2019	0.43	83.11
Thailand	THA	2019	0.58	77.15
Turkey	TUR	2019	2.21	77.69
United States	USA	2019	1.25	78.79

Our World in Data (5/12/23). Retrieved from <https://ourworldindata.org/grapher/poverty-vs-life-expectancy>

a) Make a scatterplot for the data.

Solution: Start by copying the data into an Excel spreadsheet. Highlight the two columns for the independent and dependent data. Select the Insert tab, then select the scatter plot icon. Add labels to the axis and rename the plot. See completed scatterplot below.



b) Calculate the correlation coefficient.

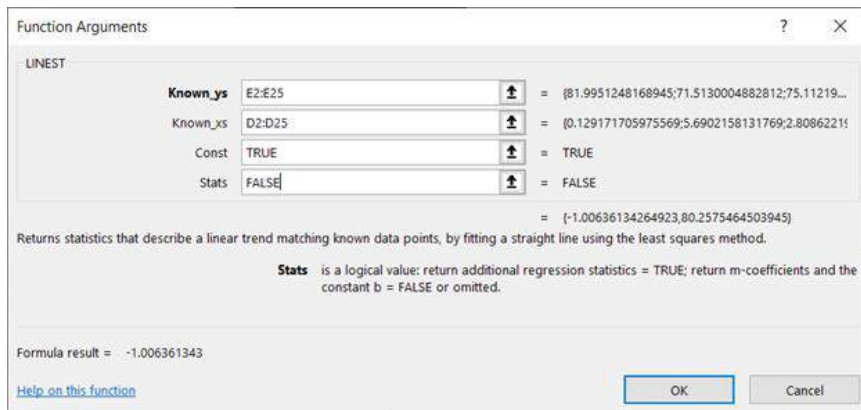
Solution: In a blank cell type in an equal sign, then the correlation function =CORREL(. Then highlight the independent data column. Type in a comma, then highlight the dependent data column. Select enter, see below.

	A	B	C	D	E	F	G	H	I
1	Country	Code	Year	\$3.65 a day - share of population below poverty line	Life expectancy at birth, total (years)				
2	Belgium	BEL	2019	0.13	82.00				
3	Bolivia	BOL	2019	5.69	71.51				
4	Bulgaria	BGR	2019	2.81	75.11				
5	China	CHN	2019	3.01	76.91				
6	Costa Rica	CRI	2019	3.46	80.28				

The correlation $r = -0.6237$.

c) Calculate the regression equation.

Solution: Select an empty cell. Select the formulas tab, then select More Formulas. From the Statistical menu select the LINEST function. Select the dependent data for the known_ys, the independent data for the known_xs, type in TRUE under Const and type in FALSE under Stats. Then press enter, see below.



You could also just type in the formula $\text{=LINEST}(E2:E25,D2:D25,TRUE,FALSE)$ in a blank cell. The first number in the output is the slope, and the second number is the y-intercept. The regression equation would be written as $\hat{y} = 80.2576 - 1.0064x$.

d) Interpret the y-intercept coefficient in this context.

Solution: Interpreting the y-intercept coefficient: When $x = 0$, note that $\hat{y} = 80.2576$, this means that we would expect a country at the poverty line would have a life expectancy of 80.2576 years.

e) Interpret the slope coefficient in this context.

Solution: Interpreting the slope coefficient: For each additional 1% below the poverty line, we would expect the life expectancy to decrease by 1.0064 years.

f) Predict the life expectancy for a country with a 0.34 share below the poverty line.

Solution: Substitute $x = 0.34$ into the regression equation. $\hat{y} = 80.2576 - 1.0064x = 80.2576 - 1.0064 \cdot 0.34 = 79.92$. We would predict that a country with a 0.34 share below the poverty line would have a life expectancy of 79.92 years.

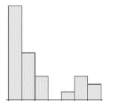
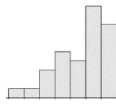
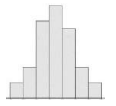
g) In 2019 Denmark had a life expectancy of 81.45 with a 0.34 share below the poverty line. How close was the regression equation at correctly predicting Denmark's life expectancy?

Solution: The difference between Denmark's life expectancy of 81.45 and the predicted life expectancy found in part f of 79.92 is called the residual. The residual is $y - \hat{y} = 81.45 - 79.92 = 1.53$ years.

Summary

When looking at correlations, start with a scatterplot to see if there is a linear relationship prior to finding a correlation coefficient. If there is a linear relationship in the scatterplot then we can find the correlation coefficient to tell the strength and direction of the relationship. Clusters of dots forming a linear uphill pattern from left-to-right will have a positive correlation. The closer the dots in the scatterplot are to a straight line the closer r will be to 1. If the cluster of dots in the scatterplots go downhill from left-to-right in linear pattern, then there is a negative relationship. The closer those dots in the scatterplot are to a straight line going downhill, the closer r will be to -1 . As sample sizes get larger, smaller values of r become statistically significant. Be careful with outliers, which can heavily influence correlations. Most importantly, correlation is not causation. When x and y are significantly correlated, this does not mean that x causes y to change. A simple linear regression should only be performed if you observe visually that there is a linear pattern in the scatterplot and that there is a statistically significant correlation between the independent and dependent variables. Use technology to find the numeric values for the y-intercept $a = b_0$ and slope $b = b_1$, then make sure to use the correct notation when substituting your numbers back in the regression equation $\hat{y} = b_0 + b_1x$.

Chapter 3 Formulas

Sample Mean: $\bar{x} = \frac{\sum x}{n}$	Population Mean: $\mu = \frac{\sum x}{N}$
Weighted Mean: $\bar{x} = \frac{\sum(xw)}{\sum w}$	Range = Max – Min
Sample Standard Deviation: $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	Population Standard Deviation = σ
Sample Variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$	Population Variance = σ^2
Coefficient of Variation: $\text{CVar} = \left(\frac{s}{\bar{x}} \cdot 100\right) \%$	Z-Score: $z = \frac{x-\bar{x}}{s}$
<p>Percentile Index: $i = \frac{(n+1) \cdot p}{100}$</p> <ul style="list-style-type: none"> If i is a whole number, count out i places from the lowest number to find the percentile. For example, if you get $i = 3$, then the 3rd value is the percentile. If i is not a whole number, then take the weighted average between the i^{th} and $i^{\text{th}} + 1$ data value as the percentile. For example, if $i = 3.25$, this would be 25% of the distance between the 3rd and the 4th data values as the percentile. Percentile = i^{th} data value + $(i^{\text{th}} + 1 \text{ data value} - i^{\text{th}} \text{ data value}) \cdot (0.\#\#)$ where $\#\#$ is the remainder percent. 	<p>Mean > Median indicates right skew</p>  <p>Mean < Median indicates left skew</p>  <p>Mean = Median indicates symmetric</p> 
$Q_1 = P_{25}$ $Q_2 = P_{50} = \text{Median}$ $Q_3 = P_{75}$	Interquartile Range: $\text{IQR} = Q_3 - Q_1$
Outlier Lower Limit: $Q_1 - (1.5 \cdot \text{IQR})$	Outlier Upper Limit: $Q_3 + (1.5 \cdot \text{IQR})$
<p>Correlation Coefficient</p> $SS_{xx} = (n - 1)s_x^2$ $SS_{yy} = (n - 1)s_y^2$ $SS_{xy} = \sum(xy) - (n \cdot \bar{x} \cdot \bar{y})$ $r = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}}$	<p>Slope</p> $b_1 = \frac{SS_{xy}}{SS_{xx}}$ <p>y-Intercept</p> $b_0 = \bar{y} - b_1 \bar{x}$ <p>Regression Equation</p> $\hat{y} = b_0 + b_1 x$

Chapter 3 Exercises

1. A sample of eight cats found the following weights in kg.

4.0	4.1	3.2	4.0	3.8	3.6	3.7	3.4
-----	-----	-----	-----	-----	-----	-----	-----

- Compute the mode.
 - Compute the median.
 - Compute the mean.
2. Cholesterol levels in milligrams (mg) of cholesterol per deciliter (dL) of blood, were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985). Retrieved from <http://www.statsci.org/data/general/cholest.html>.

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

- Compute the mode.
 - Compute the median.
 - Compute the mean.
3. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed below.

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

Data from <http://www.statsci.org/data/oz/nzrivers.html>.

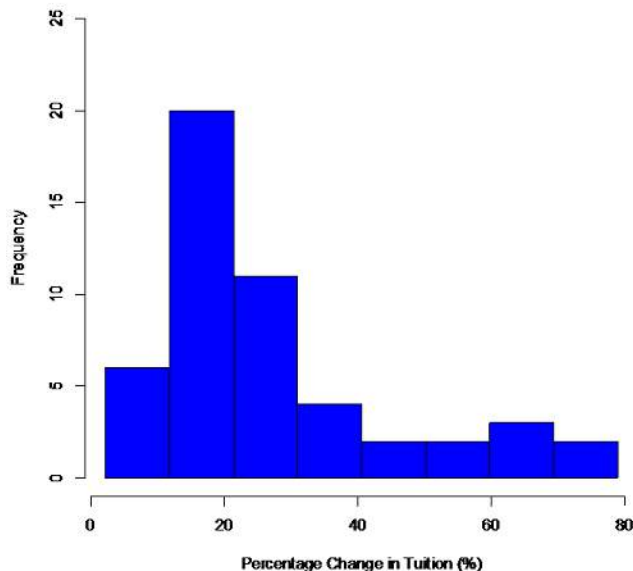
- Compute the mode.
 - Compute the median.
 - Compute the mean.
4. A university assigns letter grades with the following 4-point scale: A = 4.00, A- = 3.67, B+ = 3.33, B = 3.00, B- = 2.67, C+ = 2.33, C = 2.00, C- = 1.67, D+ = 1.33, D = 1.00, D- = 0.67, F = 0.00. Calculate the grade point average (GPA) for a student who took in one term a 4 credit math course and received a B+, a 1 credit seminar course and received an A, a 3 credit history course and received an A- and a 5 credit writing course and received a D.
5. A university assigns letter grades with the following 4-point scale: A = 4.00, A- = 3.67, B+ = 3.33, B = 3.00, B- = 2.67, C+ = 2.33, C = 2.00, C- = 1.67, D+ = 1.33, D = 1.00, D- = 0.67, F = 0.00. Calculate the grade point average (GPA) for a student who took in one term a 3 credit biology course and received a C+, a 1 credit lab course and received a B, a 4 credit engineering course and received an A- and a 4 credit chemistry course and received a C+.

6. An employee at Clackamas Community College (CCC) is evaluated based on goal setting and accomplishments toward the goals, job effectiveness, competencies, and CCC core values. Suppose for a specific employee, goal 1 has a weight of 30%, goal 2 has a weight of 20%, job effectiveness has a weight of 25%, competency 1 has a weight of 4%, competency 2 has weight of 3%, competency 3 has a weight of 3%, competency 4 has a weight of 3%, competency 5 has a weight of 2%, and core values has a weight of 10%. Suppose the employee has scores of 3.0 for goal 1, 3.0 for goal 2, 2.0 for job effectiveness, 3.0 for competency 1, 2.0 for competency 2, 2.0 for competency 3, 3.0 for competency 4, 4.0 for competency 5, and 3.0 for core values. Compute the weighted mean score for this employee. If an employee has a score less than 2.5, they must have a Performance Enhancement Plan written. Does this employee need a plan?
7. A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives an 85 on test 1, a 76 on test 2, an 83 on test 3, a 74 on the homework, a 65 on the project, and a 61 on the final, what grade did the student earn in the course? All the assignments were out of 100 points.
8. A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives a 25 out of 30 on test 1, a 20 out of 30 on test 2, a 28 out of 30 on test 3, a 120 out of 140 points on the homework, a 65 out of 100 on the project, and a 31 out of 35 on the final, what grade did the student earn in the course?
9. A sample of eight cats found the following weights in kg.

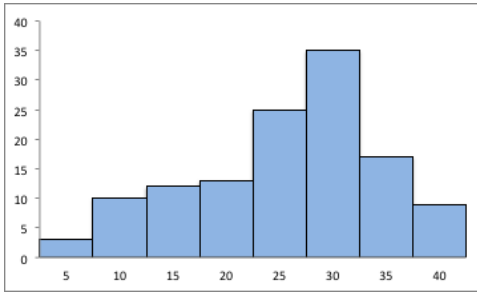
3.7	4.1	3.2	4.0	3.8	3.6	3.7	3.4
-----	-----	-----	-----	-----	-----	-----	-----

- a) Compute the range.
 b) Compute the variance.
 c) Compute the standard deviation.
10. The following data represents the percent change in tuition levels at public, four-year colleges (inflation adjusted) from 2008 to 2013 (Weissmann, 2013). To the right is the histogram. What is the shape of the distribution?

Percentage Change in Tuition Levels (Inflation Adjusted) 2008 to 2013



11. The following is a histogram of quiz grades.



- What is the shape of the distribution?
- Which is higher, the mean or the median?

12. Cholesterol levels were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985). Retrieved from <http://www.statsci.org/data/general/cholest.html>.

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

- Compute the range.
- Compute the variance.
- Compute the standard deviation.

13. Suppose that a manager wants to test two new training programs. The manager randomly selects five people for each training type and measures the time it takes to complete a task after the training. The times for both trainings are in table below. Which training method is more variable?

Training 1	56	75	48	63	59
Training 2	60	58	66	59	58

14. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed below.

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

Data from <http://www.statsci.org/data/oz/nzrivers.html>.

- Compute the range.
- Compute the variance.
- Compute the standard deviation.

15. Here are pulse rates before and after exercise. Which group has the larger range?

Before	Pulse Rates	After
9 8 8 7 6 5 2	6	
9 8 8 8 6 5 5 5 1 1 0 0	7	
8 8 7 5 4 2	8	5 6 6 7 8 9
4 0	9	0 1 1 2 3 4 5 5 6 8
4	10	0 1 4 6 7
	11	6 7
	12	4 5 7

16. A midterm in a statistics course had a mean score of 70 with a standard deviation 5. A quiz in a biology course had an average of 20 with a standard deviation of 5.

- Compute the coefficient of variation for the statistics midterm exam.
- Compute the coefficient of variation for the biology quiz.
- Aaliyah scored a 75 on the statistics midterm exam. Compute Aaliyah's z -score.
- Viannie scored a 25 on the biology quiz. Compute Viannie's z -score.
- Which student did better on their respect test? Why?
- Was there more variability for the midterm or the quiz scores? Justify your answer with statistics.

17. The following is a sample of quiz scores.

Score	17	44.5	16.1	37.2	42.8	37.5	19.5	28.2
-------	----	------	------	------	------	------	------	------

- Compute \bar{x} .
- Compute s^2 .
- Compute the median.
- Compute the coefficient of variation.
- Compute the range.

18. The time it takes to fill an online order was recorded and the following descriptive statistics were found using Excel. What is the coefficient of variation?

<i>Time</i>	
Mean	7
Standard Error	0.8085
Median	6.5
Standard Deviation	3.43
Sample Variance	11.7647
Range	12
Minimum	3
Maximum	15
Sum	126
Count	18

19. The following is the height and weight of a random sample of baseball players.

Height (inches)	Weight (pounds)
76	212
76	224
72	180
74	210
75	215
71	200
77	235
78	235
77	194
76	185
72	180
72	170
75	220
74	228
73	210
72	180
70	185
73	190
71	186
74	200
74	200
75	210
78	240
72	208
75	180

- a) Compute the coefficient of variation for both height and weight.
 - b) Is there more variation in height or weight?
20. A midterm in a statistics course had a mean score of 70 with a standard deviation 5. A quiz had an average of 20 with a standard deviation of 5. A student scored a 73 on their midterm and 22 on their quiz. On which test did the student do better on compared to the rest of the class? Justify your answer with statistics.
21. The length of a human pregnancy is normally distributed with a mean of 272 days with a standard deviation of 9.1 days. William Hunnicut was born in Portland, Oregon, at just 181 days into his gestation. What is the z -score for William Hunnicut's gestation?
Retrieved from: http://digitalcommons.georgefox.edu/cgi/viewcontent.cgi?article=1149&context=gfc_life.
22. Arm span (sometimes referred to as wingspan) is the physical measurement of the length of an individual's arms from fingertip to fingertip. The average arm span of a man is 70 inches with a standard deviation of 4.5 inches. The Olympic gold medalists Michael Phelps has an arm span of 6 feet 7 inches, which is three inches more than his height. What is the z -score for Michael Phelps arm span?
23. The average time to run the Pikes Peak Marathon 2017 was 7.44 hours with a standard deviation of 1.34 hours. Rémi Bonnet won the Pikes Peak Marathon with a run time of 3.62 hours. Retrieved from: <http://pikespeakmarathon.org/results/ppm/2017/>.

The Tevis Cup 100-mile one-day horse race for 2017 had an average finish time of 20.38 hours with a standard deviation of 1.77 hours. Tennessee Lane won the 2017 Tevis cup in a ride time of 14.75 hours. Retrieved from: <https://aerc.org/rpts/RideResults.aspx>.

- a) Compute the z -score for Rémi Bonnet's time.

- b) Compute the z -score for Tennessee Lane's time.
- c) Which competitor did better compared to their respective events?

24. Cholesterol levels were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985). Retrieved from <http://www.statsci.org/data/general/cholest.html>.

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

- a) Compute the 25th percentile.
- b) Compute the 90th percentile.
- c) Compute the 5th percentile.
- d) Compute Q_3 .

25. A sample of eight cats found the following weights in kg. Compute the 5-number summary.

3.7	4.1	3.2	4.0	3.8	3.6	3.7	3.4
-----	-----	-----	-----	-----	-----	-----	-----

26. The following data represent the grade point averages for a sample of 15 PSU students.

GPA		
3.37	2.61	2.02
3.33	2.93	2.98
3.5	3.81	3.77
2.91	2.51	3.33
3.11	3.32	3.87

- a) Compute the lower and upper limits.
- b) Identify if there are any outliers.
- c) Draw a modified box-and-whisker plot.

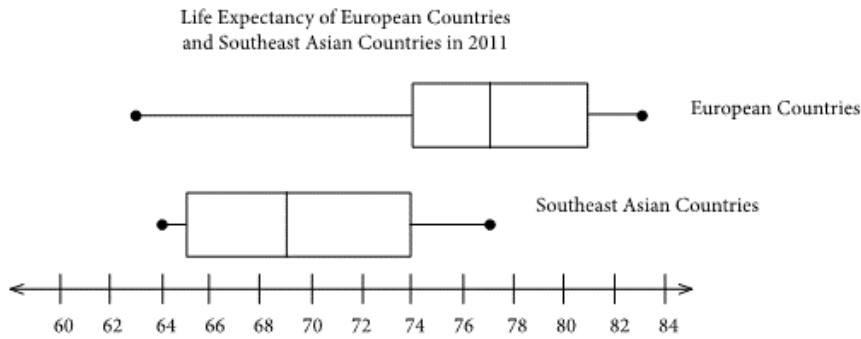
27. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed below.

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

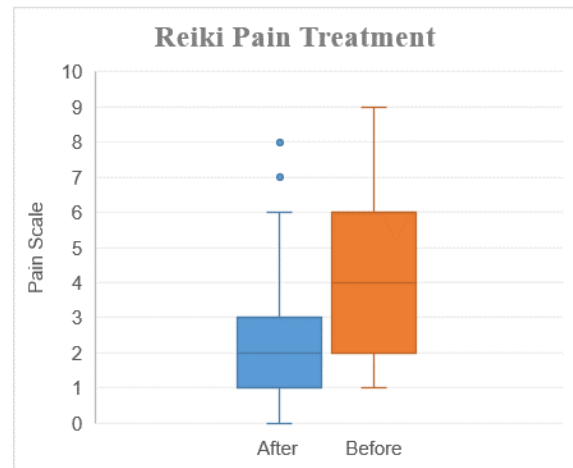
Data from <http://www.statsci.org/data/oz/nzrivers.html>.

- a) Compute the 5-number summary.
- b) Compute the lower and upper limits and any outlier(s) if any exist.
- c) Make a modified box-and-whisker plot.

28. The following are box-and-whiskers plot for life expectancy for European countries and Southeast Asian countries from 2011. What is the distribution shape of the European countries' life expectancy?

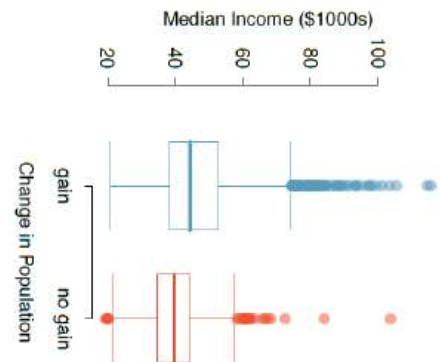


29. To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) immediately before and after the Reiki treatment (Olson & Hanson, 1997). Higher numbers mean the patients had more pain.



- Use the box-and-whiskers plots to determine the IQR for the before treatment measurements.
- Use the box-and-whiskers plots of the before and after VAS ratings to determine if the Reiki method was effective in reducing pain.

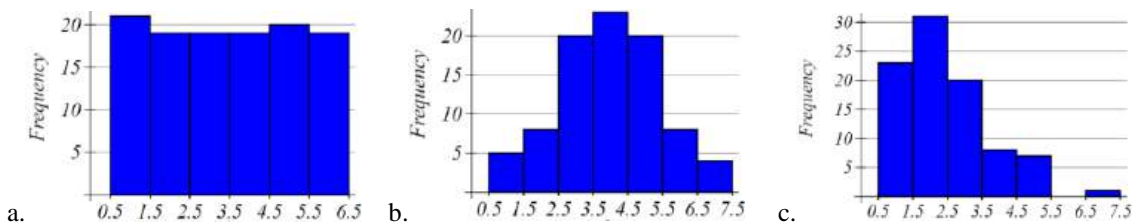
30. The median household income (in \$1,000s) from a random sample of 100 counties that gained population over 2000-2010 are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.



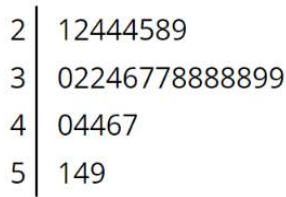
What is the distribution shape for the counties with no population gain?

(OpenIntro Statistics, 2016)

31. Sort the following histograms from the smallest standard deviation to the largest and comment on the shape of each histogram.



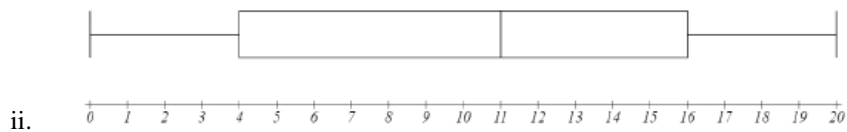
32. The following stem-and-leaf plot shows a sample of weights for 30 dogs in pounds. Calculate the following statistics.

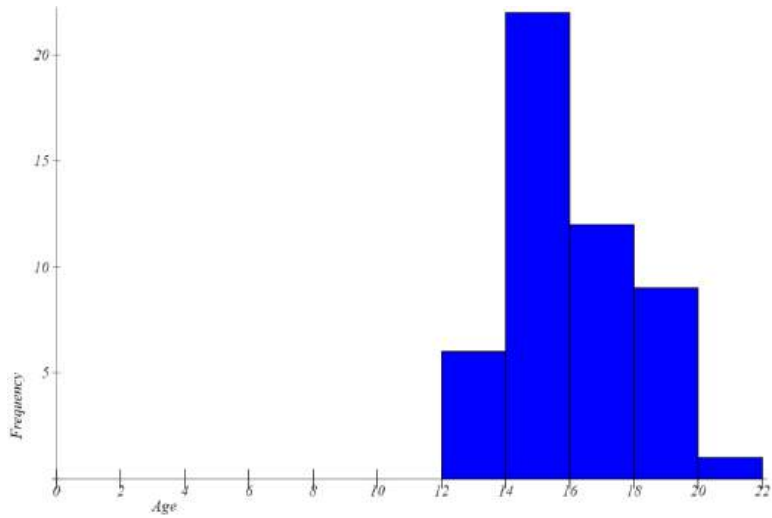


- a) \bar{x}
- b) s^2
- c) Standard Deviation
- d) Coefficient of Variation
- e) Range
- f) P_{20}
- g) Median
- h) Q_1
- i) Q_3
- j) IQR
- k) Lower and Upper limit to identify outliers. Are there any outliers?

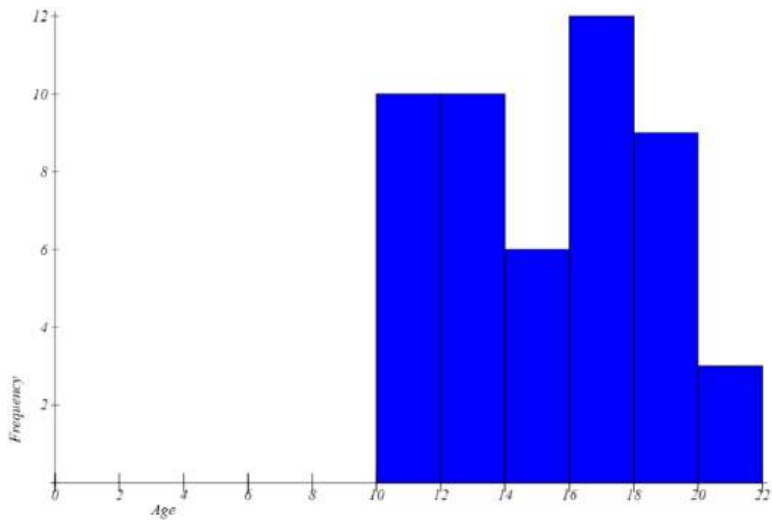
33. Match the correct descriptive statistics to the letter of the corresponding histogram and boxplot. Choose the correct letter for the corresponding histogram and Roman numeral for the corresponding boxplot. You should only use the visual representation, the definition of standard deviation and measures of central tendency to match the graphs with their respective descriptive statistics.

	Mean	Median	Standard Deviation	Histogram Letter	Boxplot Numeral
1	10.4	11	6.2		
2	16	15.8	1.9		
3	14.8	15	3.1		

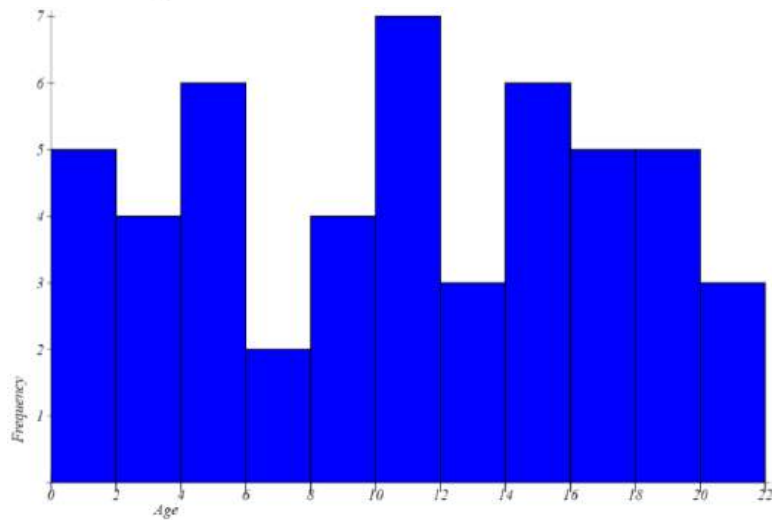




a)



b)



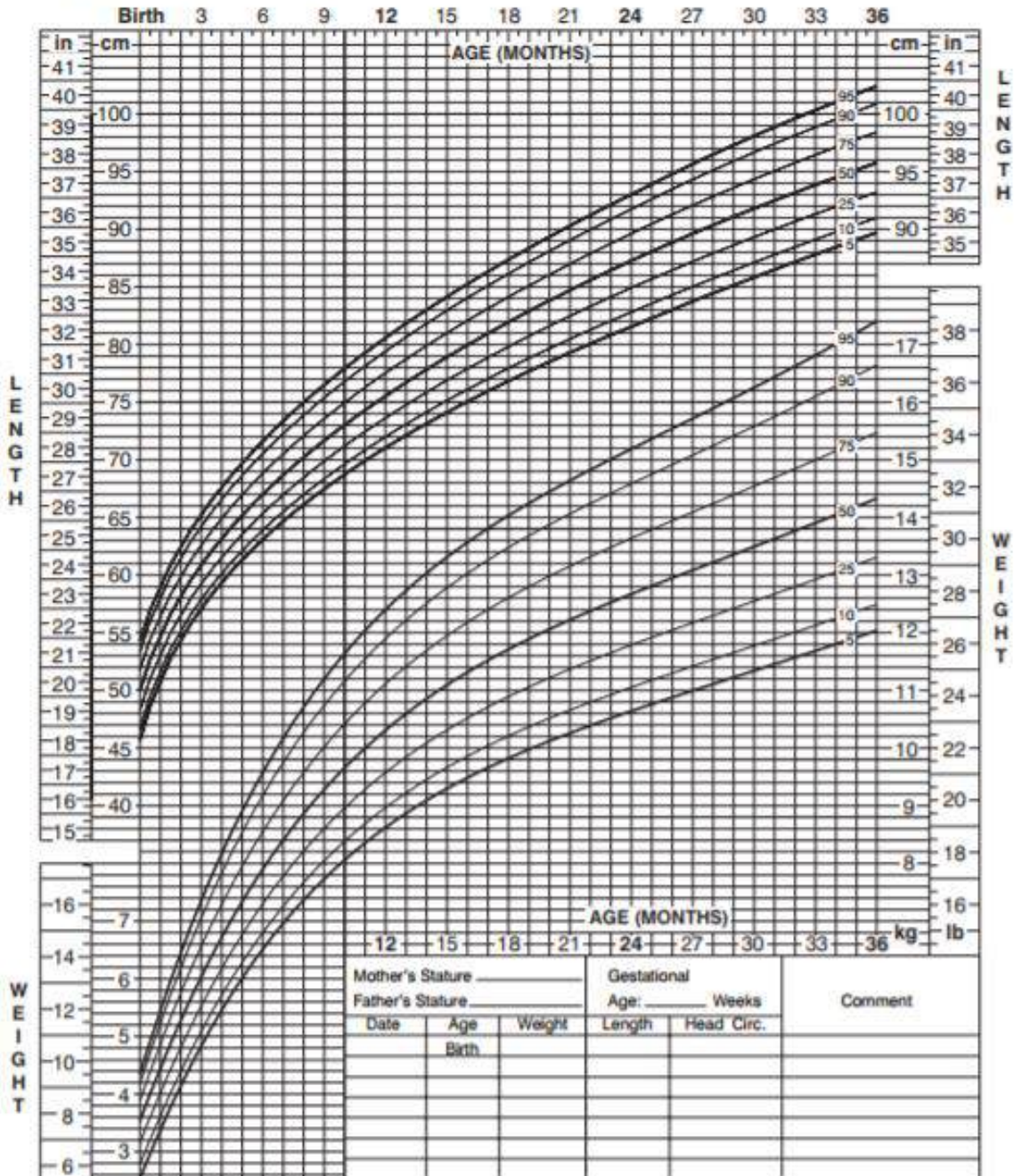
c)

34. The following is an infant weight percentile chart. What is the 50th percentile height in cm for a 10-month old boy?

Birth to 36 months: Boys
Length-for-age and Weight-for-age percentiles

NAME _____

RECORD # _____



Retrieved from: <https://www.cdc.gov/growthcharts/data/set1clinical/cj411017.pdf>.

35. The correlation coefficient, r , is a number between _____.

- a) -1 and 1
- b) -10 and 10
- c) 0 and 10
- d) 0 and ∞
- e) 0 and 1
- f) $-\infty$ and ∞

36. Which of the following is not a valid linear regression equation?

- a) $\hat{y} = -5 + \frac{2}{9}x$
- b) $\hat{y} = 3x + 2$
- c) $\hat{y} = \frac{2}{9} - 5x$
- d) $y = 5 + 0.4x$

37. Bone mineral density and cola consumption have been recorded for a sample of patients. Let x represent the number of colas consumed per week and y the bone mineral density in grams per cubic centimeter. Assume the data is normally distributed. Calculate the correlation coefficient.

x	y
1	0.883
2	0.8734
3	0.8898
4	0.8852
5	0.8816
6	0.863
7	0.8634
8	0.8648
9	0.8552
10	0.8546
11	0.862

38. An object is thrown from the top of a building. The following data measure the height of the object from the ground for a five-second period. Calculate the correlation coefficient.

Seconds	Height
0.5	112.5
1	110.875
1.5	106.8
2	100.275
2.5	91.3
3	79.875
3.5	70.083
4	59.83
4.5	30.65
5	0

39. A teacher believes that the third homework assignment is a key predictor of how well students will do on the midterm. Let x represent the third homework score and y the midterm exam score. A random sample of last term's students were selected and their grades are shown below. Assume scores are normally distributed.

HW3	Midterm
13.1	59
21.9	87
8.8	53
24.3	95
5.4	39
13.2	66
20.9	89
18.5	78

HW3	Midterm
6.4	43
20.2	79
21.8	84
23.1	92
22	87
11.4	54
14.9	71
18.4	76

HW3	Midterm
20	86
15.4	73
25	93
9.7	52
15.1	70
15	65
16.8	77
20.1	78

- a) Compute the correlation coefficient.
- b) Compute the regression equation.
- c) Compute the predicted midterm score when the homework 3 score is 15.

40. Bone mineral density and cola consumption have been recorded for a sample of patients. Let x represent the number of colas consumed per week and y the bone mineral density in grams per cubic centimeter. Assume the data is normally distributed. Compute the regression equation. Which is the best interpretation of the slope coefficient?

x	y
1	0.883
2	0.8734
3	0.8898
4	0.8852
5	0.8816
6	0.863
7	0.8634
8	0.8648
9	0.8552
10	0.8546
11	0.862

- a) For every additional average weekly soda consumption, a person's bone density increases by 0.0031 grams per cubic centimeter.
- b) For every additional average weekly soda consumption, a person's bone density decreases by 0.0031 grams per cubic centimeter.
- c) For an increase of 0.8893 in the average weekly soda consumption, a person's bone density decreases by 0.0031 grams per cubic centimeter.
- d) For every additional average weekly soda consumption, a person's bone density decreases by 0.8893 grams per cubic centimeter.

41. The following data represent the leaching rates (percent of lead extracted vs. time in minutes) for lead in solutions of magnesium chloride ($MgCl_2$).

Time (x)	4	8	16	30	60	120
Percent Extracted (y)	1.2	1.6	2.3	2.8	3.6	4.4

- a) Compute the correlation coefficient.
- b) Compute the regression equation.
- c) Predict the percent extracted for 100 minutes.

42. Compute the regression equation for the predicted average high temperature ($^{\circ}F$) per month by the Farmer's Almanac (x) in Portland, Oregon and the actual high (y) temperature per month that occurred.

Farmer's Almanac	45	50	57	62	69	72	81	90	78	64	51	48
Actual High	46	52	60	61	72	78	82	95	85	68	52	49

Chapter 4

Probability



4.1 Introduction to Probability

4.2 Complement Rule

4.3 Union & Intersection

4.4 Independent Events & Conditional Probability

4.5 Counting Rules

4.1 Introduction to Probability

One story about how probability theory was developed is that a gambler wanted to know when to bet more and when to bet less. He talked to a couple of friends of his that happened to be mathematicians. Their names were Pierre de Fermat and Blaise Pascal. Since then, many other mathematicians have worked to develop probability theory.

Understanding probability is important in life. Examples of mundane questions that probability can answer for you are: do I need to carry an umbrella? Do you wear a heavy coat on a given day? More important questions that probability can help with are your chances that the car you are buying will need more maintenance, your chances of passing a class, your chances of winning the lottery, or your chances of catching a deadly virus. The chance of you winning the lottery is very small, yet many people will spend the money on lottery tickets. In general, events that have a low probability (under 5%) are unlikely to occur. Whereas if an event has a high probability of happening (over 80%), then there is a good chance that the event will happen. This chapter will present some of the theory that you need to help decide if an event is likely to happen or not.



First, some definitions:

Experiment: an activity or process that has specific results that can be repeated indefinitely which has a set of well-defined outcomes.

Outcomes: the results of an experiment.

Sample Space: collection of all possible outcomes of the experiment. Usually denoted as S .

Event: the set of outcomes that are a subset of the sample space. The symbol used for events is usually a capital letter often at the beginning of the alphabet like A , B or C .

Here are some examples of sample spaces and events.

Experiment	Sample Space	Example of Event
Toss a coin twice	{HH, HT, TH, TT}	$A =$ Getting exactly two heads = {HH}
Toss a coin twice	{HH, HT, TH, TT}	$B =$ Getting at least one head = {HH, HT, TH}
Roll a die	{1, 2, 3, 4, 5, 6}	$C =$ Roll an odd number = {1, 3, 5}
Roll a die	{1, 2, 3, 4, 5, 6}	$D =$ Roll a prime number = {2, 3, 5}
Roll a die	{1, 2, 3, 4, 5, 6}	$E =$ Roll an even number = {2, 4, 6}

Figure 4-1

A **tree diagram** is a graphical way of representing a random experiment with multiple steps.

Example 4-1: A bag contains 10 colored marbles: 7 red and 3 blue. A random experiment consists of drawing a marble from the bag, then drawing another marble without replacement (without putting the first marble back in the bag). Create the tree diagram for this experiment and write out the sample space.

Solution: The first marble that is drawn can be either red or blue and is represented with the first sideway V split. The next marble that is drawn is represented by the two sideway V splits on the right. The top sideways V assumes that a red marble was drawn on the first draw, and then the second marble drawn can be either red or blue. The bottom sideways V assumes that a blue marble was drawn on the first draw, and then the second marble drawn can be either red or blue. Then combine the colors as you trace up the four pathways from left to right.

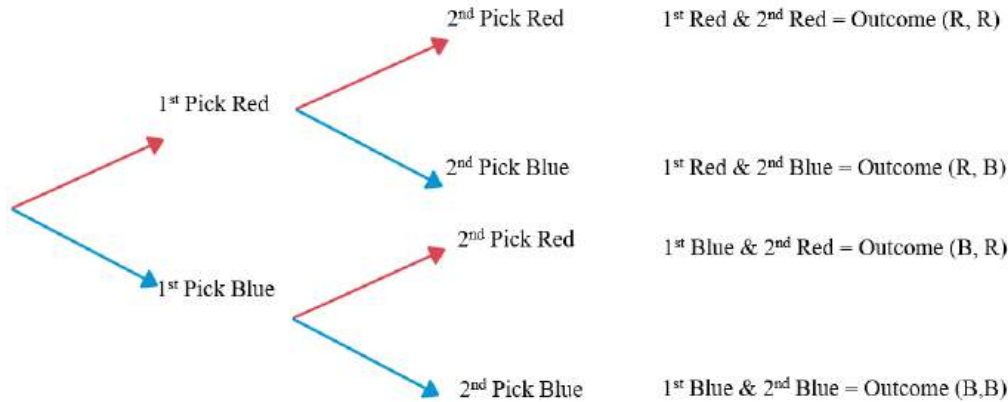


Figure 4-2

The sample space would be $S = \{RR, RB, BR, BB\}$. Note that the event RB and BR are considered different outcomes since they are picked in a different order and are considered distinct events.

Three Types of Probability

1. Classical Approach to Probability (Theoretical Probability)

$$P(A) = \frac{\text{Number of ways A can occur}}{\text{Number of different outcomes in S}}$$

The classical approach can only be used if each outcome has equal probability.

Example 4-2: If an experiment consists of flipping a coin twice, compute the probability of getting exactly two heads.

Solution: The event of getting exactly two heads is $A = \{HH\}$. The number of ways A can occur is 1. The number of different outcomes in $S = \{HH, HT, TH, TT\}$ is 4. Thus $P(A) = 1/4$.

Example 4-3: If a random experiment consists of rolling a six-sided die, compute the probability of rolling a 4.

Solution: The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. The event A is that you want is to get a 4, and the event space is $A = \{4\}$. Thus, in theory the probability of rolling a 4 would be $P(A) = 1/6 = 0.1667$.

Example 4-4: Suppose you have an iPhone with the following songs on it: 5 Rolling Stones songs, 7 Beatles songs, 9 Bob Dylan songs, 4 Johnny Cash songs, 2 Carrie Underwood songs, 7 U2 songs, 4 Mariah Carey songs, 7 Bob Marley songs, 6 Bunny Wailer songs, 7 Elton John songs, 5 Led Zeppelin songs, and 4 Dave Matthews Band songs. The different genre that you have are rock from the '60s which includes Rolling Stones, Beatles, and Bob Dylan; country includes Johnny Cash and Carrie Underwood; rock of the '90s includes U2 and Mariah Carey; Reggae includes Bob Marley and Bunny Wailer; rock of the '70s includes Elton John and Led Zeppelin; and bluegrass/rock includes Dave Matthews Band.

a) What is the probability that you will hear a Johnny Cash song?

Solution: The way an iPhone works, it randomly picks the next song so you have no idea what the next song will be. Now you would like to calculate the probability that you will hear the type of music or the artist that you are interested in. The sample set is too difficult to write out, but you can figure it from looking at the number in each set and the total number. The total number of songs you have is 67.

There are 4 Johnny Cash songs out of the 67 songs. $P(\text{Johnny Cash song}) = 4/67 = 0.0597$.

- b) What is the probability that you will hear a Bunny Wailer song?

Solution: There are 6 Bunny Wailer songs. $P(\text{Bunny Wailer}) = 6/67 = 0.0896$.

- c) What is the probability that you will hear a song from the '60s?

Solution: There are 5, 7, and 9 songs that are classified as rock from the '60s, which is a total of 21.

$P(\text{rock from the '60s}) = 21/67 = 0.3134$.

- d) What is the probability that you will hear a Reggae song?

Solution: There are total of 13 songs that are classified as Reggae. $P(\text{Reggae}) = 13/67 = 0.1940$.

- e) What is the probability that you will hear a song from the '90s or a bluegrass/rock song?

Solution: There are 7 and 4 songs that are songs from the '90s and 4 songs that are bluegrass/rock, for a total of 15. $P(\text{rock from the '90s or bluegrass/rock}) = 15/67 = 0.2239$.

- f) What is the probability that you will hear an Elton John or a Carrie Underwood song?

Solution: There are 7 Elton John songs and 2 Carrie Underwood songs, for a total of 9.

$P(\text{Elton John or Carrie Underwood song}) = 9/67 = 0.1343$.

- g) What is the probability that you will hear a country song or a U2 song?

Solution: There are 6 country songs and 7 U2 songs, for a total of 13. $P(\text{country or U2 song}) = 13/67 = 0.1940$.

2. **Empirical Probability** (Experimental or Relative Frequency Probability)

The experiment is performed many times and the number of times that event A occurs is recorded. Then the probability is approximated by finding the relative frequency.

$$P(A) = \frac{\text{Number of times } A \text{ occurred}}{\text{Number of times the experiment was repeated}}$$

Important: The probability of any event A satisfies $0 \leq P(A) \leq 1$, keep this in mind if the question is asking for a probability, and make sure your answer is a number between 0 and 1. A probability, relative frequency, percentage, and proportion are all different words for the same concept. Probability answers can be given as percentages, decimals, or reduced fractions.

Example 4-5: Suppose that the experiment is rolling a die. Compute the probability of rolling a 4.

Solution: The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. The event A is that you want to get a 4, and the event space is $A = \{4\}$. To do this experiment, start with rolling a die 10 times. Let pretend that you get a 4, two times. Based on this experiment, the probability of getting a 4 is 2 out of 10 or $1/10 = 1/5 = 0.2$. To get more accuracy, repeat the experiment more times. It is easiest to put this information in a table, where n represents the number of times the experiment is repeated. When you put the number of 4s found divisible by the number of times you repeat the experiment, this is the relative frequency. See the last column in Figure 4-3.

Trials for Die Experiment

n	Number of 4s	Relative Frequency
10	2	0.2
50	6	0.12
100	18	0.18
500	81	0.162
1,000	163	0.163

Figure 4-3

Notice that as n increased, the relative frequency seems to approach a number; it looks like it is approaching 0.163. You can say that the probability of getting a 4 is approximately 0.163. If you want more accuracy, then increase n even more by rolling the die more times.

These probabilities are called **experimental probabilities** since they are found by actually doing the experiment or simulation. They come about from the relative frequencies and give an approximation of the true probability.

The approximate probability of an event A , notated as $P(A)$, is

$$P(A) = \frac{\text{Number of times } A \text{ occurred}}{\text{Number of times the experiment was repeated}}$$

For the event of getting a 4, the probability would be $P(\text{Roll a 4}) = \frac{163}{1000} = 0.163$.

“‘What was that voice?’ shouted Arthur.
 ‘I don’t know,’ yelled Ford, ‘I don’t know. It sounded like a measurement of probability.’
 ‘Probability? What do you mean?’
 ‘Probability. You know, like two to one, three to one, five to four against. It said two to the power of one hundred thousand to one against. That’s pretty improbable you know.’”
 (Adams, 2002)

Law of Large Numbers: as n increases, the relative frequency tends towards the theoretical probability.

Figure 4-4 shows a graph of experimental probabilities as n gets larger and larger. The dashed yellow line is the theoretical probability of rolling a four of $1/6 \approx 0.1667$. Note the x-axis is in a log scale.

Note that the more times you roll the die, the closer the experimental probability gets to the theoretical probability.

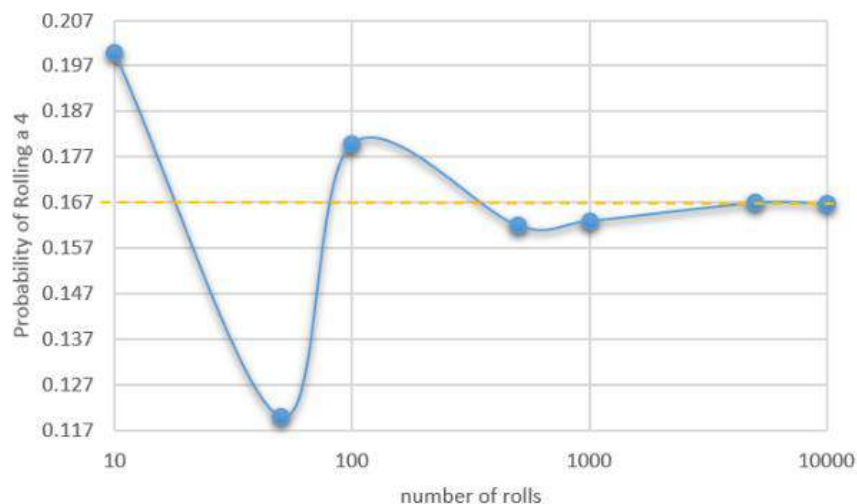


Figure 4-4

You can compute experimental probabilities whenever it is not possible to calculate probabilities using other means. An example is if you want to find the probability that a family has 5 children, you would have to actually look at many families, and count how many have 5 children. Then you could calculate the probability. Another example is if you want to figure out if a die is fair. You would have to roll the die many times and count how often each side comes up. Make sure you repeat an experiment many times, because otherwise you will not be able to estimate the true probability of 5 children. This is due to the law of large numbers, since the more times we repeat the experiment, the closer the experimental probabilities will get to the theoretical probabilities. For difficult theoretical probabilities, we can run computer simulations that can run an experiment repeatedly many times very quickly and come up with accurate estimates of the theoretical probability.

Example 4-6: A fitness center coach kept track of members over the last year. They recorded if the person stretched before they exercised, and whether they sustained an injury. The following contingency table shows their results. Select one member at random and find the following probabilities.

	Injury	No Injury
Stretched	52	270
Did Not Stretch	21	57

- a) Compute the probability that a member sustained an injury.

Solution: Find the totals for each row, column, and grand total.

	Injury	No Injury	Total
Stretched	52	270	322
Did Not Stretch	21	57	78
Total	73	327	400

Next, find the relative frequencies by dividing each number by the total of 400.

	Injury	No Injury	Total
Stretched	0.13	0.675	0.805
Did Not Stretch	0.0525	0.1425	0.195
Total	0.1825	0.8175	1

Using the definition of a probability we get $P(\text{Injury}) = \frac{\text{Number of injuries}}{\text{Total number of people}} = \frac{73}{400} = 0.1825$.

	Injury	No Injury	Total
Stretched	0.13	0.675	0.805
Did Not Stretch	0.0525	0.1425	0.195
Total	0.1825	0.8175	1

Using the table, we can get the same answer very quickly by just taking the column total under Injury to get 0.1825. As we get more complicated probability questions, these contingency tables will help organize your data.

- b) Compute the probability that a member did not stretch.

Solution: Using the relative frequency contingency table, take the total of the row for all the members that did not stretch and we get the $P(\text{Did Not Stretch}) = 0.195$.

	Injury	No Injury	Total
Stretched	0.13	0.675	0.805
Did Not Stretch	0.0525	0.1425	0.195
Total	0.1825	0.8175	1

c) Compute the probability that a member sustained an injury and did not stretch.

Solution: Using the relative frequency contingency table, take the intersection of the injury column with the did not stretch row and we get $P(\text{Injury and Did Not Stretch}) = 0.0525$.

	Injury	No Injury	Total
Stretched	0.13	0.675	0.805
Did Not Stretch	0.0525	0.1425	0.195
Total	0.1825	0.8175	1

3. Subjective Probability

The probability of event A is estimated using previous knowledge and is someone's opinion.

Example 4-7: Compute the probability of meeting Dolly Parton.

Solution: I estimate the probability of meeting Dolly Parton to be $1.2E-9 \approx 0.0000000012$ (i.e., very, very small).

Example 4-8: What is the probability it will rain tomorrow?

Solution: A weather reporter looks at several forecasts, uses their expert knowledge of the region, and reports the probability that it will rain in Portland, OR, is 80%.



4.2 Complement Rule

There is a faster way to computer these probabilities that will be important for more complicated probabilities called the **complement rule**. The complement of an event A represents all outcomes that are not in A . The complement of A is denoted as A^C . The events A and A^C are called complementary event and must be mutually exclusive. When two events cannot happen at the same time, they are called **mutually exclusive** or **disjoint** events.

Example 4-9: Suppose you know that the probability of it raining today is 80%. What is the probability of it not raining today?

Solution: Since not raining is the complement of raining, then $P(\text{not raining}) = 100\% - 80\% = 20\%$. Using probability notation, we change the percentages to proportions and say $P(\text{not raining}) = 1 - P(\text{raining}) = 1 - 0.8 = 0.2$.

If two events are **complementary events**, then to find the probability of one event just subtract the probability from 1. Notation used for complement of A also called “not A ” is A^C .

$$P(A) + P(A^C) = 1 \quad \text{or} \quad P(A) = 1 - P(A^C) \quad \text{or} \quad P(A^C) = 1 - P(A)$$

Some texts will use the notation for a complement as A' or \bar{A} , instead of A^C .

“On a small obscure world somewhere in the middle of nowhere in particular - nowhere, that is, that could ever be found, since it is protected by a vast field of improbability to which only six men in this galaxy have a key - it was raining.”
(Adams, 2002)

Example 4-10: A random sample of 500 records from the 2010 United States Census were downloaded to Excel and the following contingency table was found for biological sex and marital status. Select one member at random and find the following probabilities.

Count of Marital Status	Column Labels		
Row Labels	Female	Male	Grand Total
Divorced	21	17	38
Married/spouse absent	5	9	14
Married/spouse present	92	100	192
Never married/single	93	129	222
Separated	1	2	3
Widowed	20	11	31
Grand Total	232	268	500

- a) Compute the probability that a person is divorced.

Solution: Take the row total of all divorced which is 38 and then divide by the grand total of 500 to get $P(\text{Divorced}) = 38/500 = 0.076$.

- b) Compute the probability that a person is not divorced.

Solution: We can add up all the other category totals besides divorced $14 + 192 + 222 + 3 + 31 = 462$, then divide by the grand total to get $P(\text{Not Divorced}) = 462/500 = 0.924$.

The table contains 100% ($100\% = 1$ as a proportion) of our data so we can assume that the probability of the divorced is the opposite (complement) to the probability of not being divorced. Notice that the $P(\text{Divorced}) + P(\text{Not Divorced}) = 1$. This is because these two events have no outcomes in common, and together they make up the entire sample space.

Notice $P(\text{Not Divorced}) = 1 - P(\text{Divorced}) = 1 - 0.076 = 0.924$.

Venn Diagrams

Figure 4-5 is an example of a Venn diagram and is a visual way to represent sets and probability. The rectangle represents all the possible outcomes in the entire sample space (the population). The shapes inside the rectangle represent each event in the sample space. Usually these are ovals, but they can be any shape you want. If there are any shared elements between the events then the circles should overlap one another.

The field of statistics includes machine learning, data analysis, and data science. The field of computer science includes machine learning, data science and web development. The field of business and domain expertise includes data analysis, data science and web development. If you know machine learning, then you will need a background in both statistics and computer science. If you are a data scientist, then you will need a background in statistics, computer science, business and domain expertise.

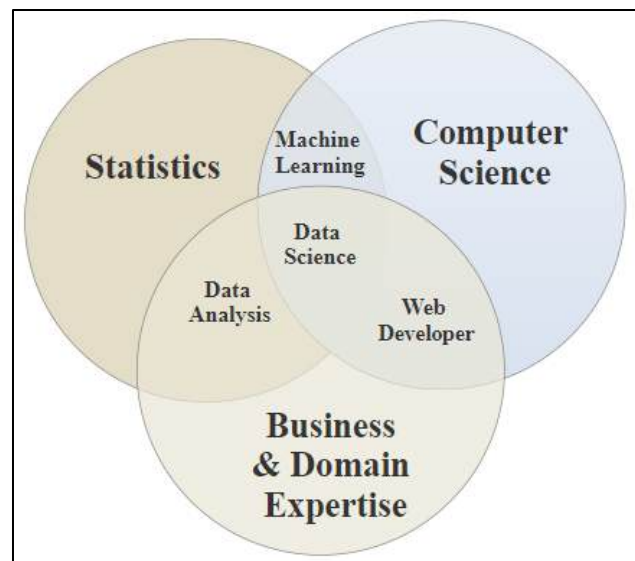


Figure 4-5

Example 4-11: Suppose you know the probability of getting the flu is 0.24. Draw a Venn diagram and find the probability of not getting the flu.

Solution: Since not getting the flu is the complement of getting the flu, the $P(\text{not getting the flu}) = 1 - P(\text{getting the flu}) = 1 - 0.24 = 0.76$.

Label each space as in Figure 4-6.

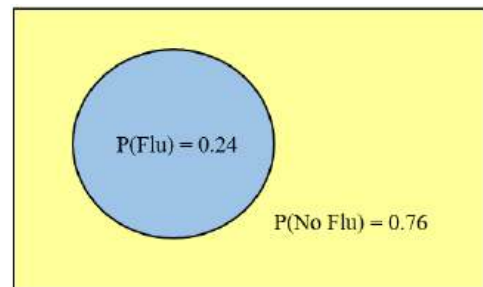


Figure 4-6

The complement is useful when you are trying to find the probability of an event that involves the words “at least” or an event that involves the words “at most.” As an example of an “at least” event is supposing you want to find the probability of making at least \$50,000 when you graduate from college. That means you want the probability of your salary being greater than or equal to \$50,000.

An example of an “at most” event is supposing you want to find the probability of rolling a die and getting at most a 4. That means that you want to get less than or equal to a 4 on the die, a 1, 2, 3, or 4.

The reason to use the complement is that sometimes it is easier to find the probability of the complement and then subtract from 1.

4.3 Union & Intersection

When two events A or B cannot happen at the same time, they are called **mutually exclusive** or **disjoint** events.

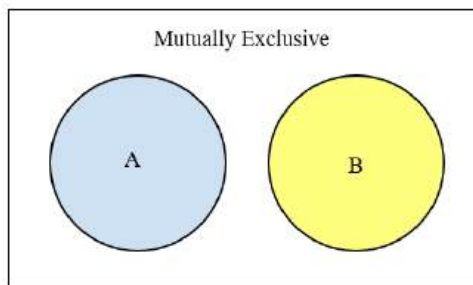


Figure 4-7

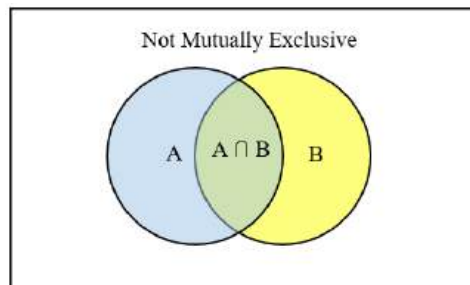


Figure 4-8

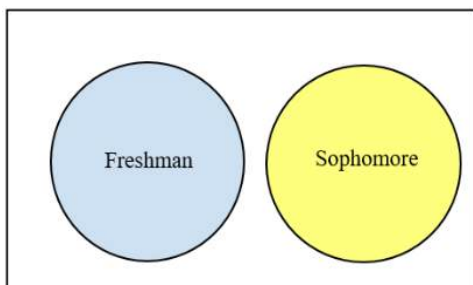


Figure 4-9

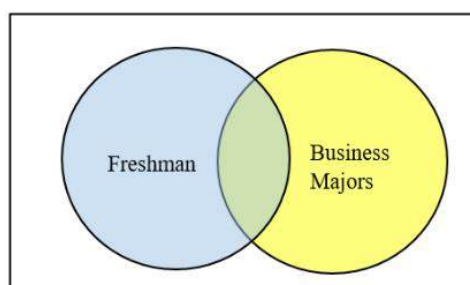


Figure 4-10

For example, a student cannot be a freshman and a sophomore at the same time, see Figure 4-9. These are mutually exclusive events. A student could be freshman and a business major at the same time so the event freshman and the event business major are not mutually exclusive see Figure 4-10.

Intersection

When we are finding the probability of both **A and B** happening at the same time, we denote this as $P(A \cap B)$. This overlap is called the intersection.

When two events, say A and B, occur at the same time, this is denoted as the **intersection** of A and B and is denoted as $(A \cap B)$. Think of the symbol \cap as an A in “and.”

If two events are mutually exclusive then $A \cap B = \{ \}$ the empty set (also denoted as \emptyset) and the $P(A \cap B) = 0$.

Union

When either event A, event B, or both occur then we call this the **union** of A or B, which is denoted as $(A \cup B)$. When finding the probability of A or B we denote this as $P(A \cup B)$. When we write “or” in statistics, we mean “and/or” unless we explicitly state otherwise. Thus, A or B occurs means A, B, or both A and B occur. Some texts will call this the addition rule.

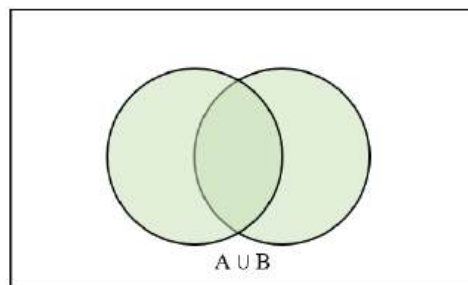


Figure 4-11 is a Venn diagram for the union rule.

Figure 4-11

The Union Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

If two events are mutually exclusive, then the probability of them occurring at the same time is $P(A \cap B) = 0$. So, if A and B are mutually exclusive then the $P(A \cup B) = P(A) + P(B)$ as shown in Figure 4-7. It is best to write out the rule with the intersection so that you do not forget to subtract any overlapping intersection.

Example 4-12: The family college data set contains a sample of 792 cases with two variables, teen and parents. The teen variable is either college or not, where the college label means the teen went to college immediately after high school. The parent’s variable takes the value degree if at least one parent of the teenager completed a college degree. Make a Venn Diagram for the data.

Example from [OpenIntroStatistics](#).

		Parent		Total
		Degree	No Degree	
Teen	College	231	214	445
	No College	49	298	347
	Total	280	512	792

Solution: Find the relative frequencies.

	Degree	No Degree	Total
College	0.29	0.27	0.56
No College	0.06	0.38	0.44
Total	0.35	0.65	1

Use shapes to show the overlaps between the groups, see Figure 4-12 for completed Venn Diagram. Note that you do not need to use circles to represent the sets.

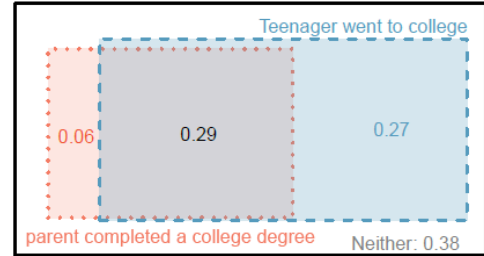


Figure 4-12

Example 4-13: A random sample of 500 people was taken from the 2010 United States Census. Their marital status and race were recorded in the following contingency table using the census labels. A person is randomly chosen from the census data. Find the following.

Marital Status	Race					Total
	American Indian	Black	Asian	White	Two Major Races	
Divorced	0	6	1	30	1	38
Married	1	25	23	156	4	209
Single	2	33	21	155	11	222
Widowed	0	7	2	22	0	31
Total	3	71	47	363	16	500

- a) $P(\text{Single} \cap \text{American Indian})$

Solution: The intersection for a contingency table is found by simply finding where the row or intersection meets. There are 2 Single American Indians, therefore the $P(\text{Single} \cap \text{American Indian}) = P(\text{Single and American Indian}) = 2/500 = 0.004$.

- b) $P(\text{Single} \cup \text{American Indian})$

Solution: There are 222 Single people, there are 3 American Indians, but we do not want to count the 2 Single American Indians twice, therefore the $P(\text{Single} \cup \text{American Indian}) = P(\text{Single or American Indian}) = 222/500 + 3/500 - 2/500 = 223/500 = 0.446$.

- c) Probability that the person is Asian or Married.

Solution: The union for a contingency table is found by either using the union formula or adding up all the numbers in the corresponding row and column. There are 47 Asian people, there are 209 Married people, but we do not want to count the 23 Married Asian people twice, therefore the $P(\text{Asian} \cup \text{Married}) = P(\text{Asian or Married}) = 47/500 + 209/500 - 23/500 = 233/500 = 0.466$.

- d) $P(\text{Single} \cap \text{Married})$

Solution: The events Single and Married are mutually exclusive so the $P(\text{Single} \cap \text{Married}) = 0$. Alternatively, there is no place in the table where the Single row and Married row meet.

- e) $P(\text{Single} \cup \text{Married})$

Solution: The events single and married are mutually exclusive so the $P(\text{Single} \cup \text{Married}) = P(\text{Single}) + P(\text{Married}) - P(\text{Single} \cap \text{Married}) = 222/500 + 209/500 - 0 = 431/500 = 0.862$.

Example 4-14: Use a random experiment consisting of rolling two dice and adding the numbers on the faces.

- a) Compute the probability of rolling a sum of 8.

Solution: There are 36 possible outcomes for rolling the two dice as shown in the following sum table. There are 5 pairs where the sum of the two dice is an 8, they are (2,6), (3,5), (4,4), (5,3) and (6,2). Note that the events (2,6) and (6,2) are different outcomes since they numbers come from different dice.

+		Second Die					
		1	2	3	4	5	6
First Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Thus, the $P(8) = 5/36 = 0.1389$.

- b) Compute the probability of rolling a sum of 8 or a sum of 5.

Solution: Highlight all the places where a sum of 5 or a sum of 8 occurs. There are 9 pairs where the sum of the two dice is a 5 or an 8.

+		Second Die					
		1	2	3	4	5	6
First Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Thus, the $P(5 \cup 8) = P(5) + P(8) - P(5 \cap 8) = 4/36 + 5/36 - 0 = 9/36 = 0.25$.

Note that rolling a sum of 5 is mutually exclusive from rolling a sum of 8 so the probability is zero for the intersection of the two events.

- c) Compute the probability of rolling a sum of 8 or a double (each die has the same number).

Solution: The events rolling an 8 and rolling doubles are not mutually exclusive since the pair of fours (4, 4) falls into both events.

An easy way is to highlight all the places a sum 8 or doubles occur, count the highlighted values, and divide by the total $10/36 = 0.2778$.

+		Second Die					
		1	2	3	4	5	6
First Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

When using the union rule, we subtract this overlap out one time to account for this. Using the union formula:

$$P(8 \cup \text{Doubles}) = P(8) + P(\text{Doubles}) - P(8 \cap \text{Doubles}) = 5/36 + 6/36 - 1/36 = 10/36 = 0.2778.$$

“It’s... well, it’s a long story,” he said, “but the Question I would like to know is the Ultimate Question of Life, the Universe and Everything. All we know is that the Answer is Forty-two, which is a little aggravating.”
 Prak nodded again.
 ‘Forty-two,’ he said. ‘Yes, that’s right.’
 He paused. Shadows of thought and memory crossed his face like the shadows of clouds crossing the land.
 ‘I’m afraid,’ he said at last, ‘that the Question and the Answer are mutually exclusive. Knowledge of one logically precludes knowledge of the other. It is impossible that both can ever be known about the same universe.’”
 (Adams, 2002)

Example 4-15: Randomly pick a card from a standard deck. A standard deck of cards, not including jokers consists of 4 suits called clubs = ♣, spades = ♠, hearts = ♥, and diamonds = ♦. The clubs and spades are called the black cards. The hearts and diamonds are called the red cards. Each suit has 13 cards. The numbered cards shown in Figure 4-13, are Ace = 1 or A, 2, 3, 4, 5, 6, 7, 8, 9, 10. The face cards are the Jack = J, Queen = Q, and King = K.

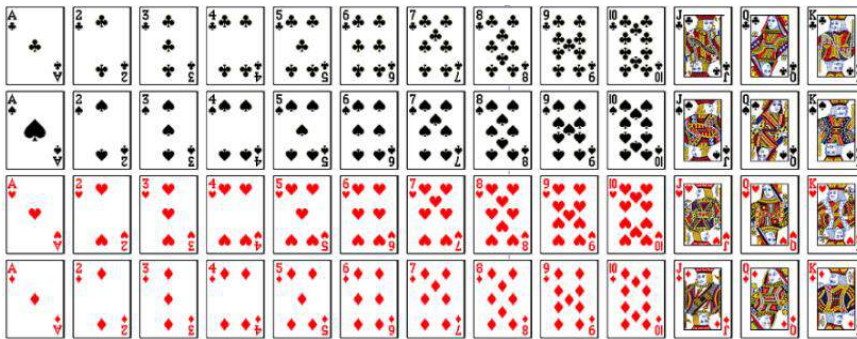


Figure 4-13

- a) Compute the probability of selecting a card that shows a club.

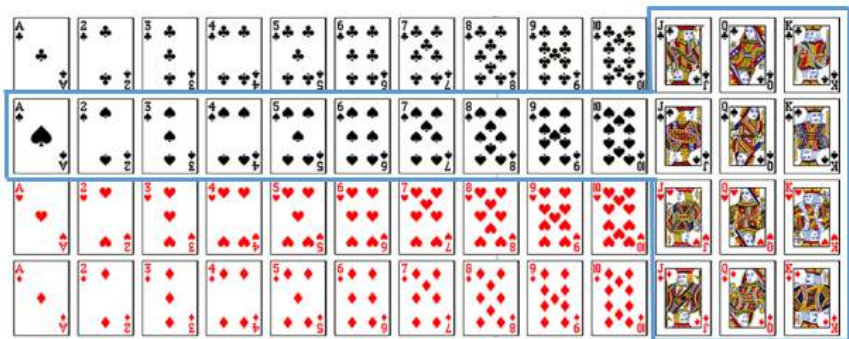
Solution: There are 52 cards in a standard deck. There are 13 cards of each suit. The $P(\clubsuit) = 13/52 = 0.25$.

- b) Compute the probability of selecting a heart or a spade card.

Solution: There are 52 cards in a standard deck. There are 13 cards of each suit. $P(\heartsuit \cup \spadesuit) = 26/52 = 0.5$.

- c) Compute the probability of selecting a spade or a face card.

Solution: There are 13 spades and 12 face cards. However, there are 3 cards that are both spades and face cards. $P(\spadesuit \cup \text{FC}) = P(\spadesuit) + P(\text{FC}) - P(\spadesuit \cap \text{FC}) = 13/52 + 12/52 - 3/52 = 22/52 = 0.4231$. Since the sample space is small, you could just count how many spades and face cards there are.



Words Are Important!

When working with probability, words such as “more than” or “less than” can drastically change the answer. Figure 4-14 shows some of the common phrases you may run into while reading a problem. It will be essential later in the course that you can correctly match these phrases with their correct symbol.

=	\leq	\geq
Is the same as	Is less than or equal to	Is greater than or equal to
Is equal to	Is at most	Is at least
Is exactly the same as	Is not greater than	Is not less than
Has not changed from	Within	
\neq	$>$	$<$
Is not	More than	Less than
Is not equal to	Greater than	Below
Is different from	Above	Lower than
Has changed from	Higher than	Shorter than
Is not the same as	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

Figure 4-14

Example 4-16: Use a random experiment consisting of rolling two dice and adding the numbers on the faces.

- a) Compute the probability of rolling a sum of less than 5.

Solution: Let X be the event rolling a sum less than 5. A sum less than 5 would not include the 5. For notation, we use $P(X < 5)$, which is read as the “probability that X is less than five.” Shade in all the sums that are less than 5.

+		Second Die					
		1	2	3	4	5	6
First Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Then the $P(X < 5) = 6/36 = 0.1667$.

- b) Compute the probability of rolling a sum of 5 or less.

Solution: Let X be the event rolling a sum of 5 or less. A sum of 5 or less includes the 5. For notation, we can use $P(X \leq 5)$, which is read as the “probability that X is less than or equal to five.” Shade in all the sums that are 5 or less.

+		Second Die					
		1	2	3	4	5	6
First Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Then the $P(X \leq 5) = 10/36 = 0.2778$.

4.4 Independent Events and Conditional Probability

Two trials (or events or results of a random experiment) are **independent** trials if the outcome of one trial does not influence the outcome of the second trial. If two events are not independent, they are dependent events. For instance, if two coins are flipped, they are independent since flipping one coin does not affect the outcome of the second coin.

Independent Events: If A and B are independent events, then $P(A \cap B) = P(A) \cdot P(B)$.

Be careful with this rule. You cannot just multiply probabilities to find an intersection unless you know they are independent. Also, do not confuse independent events with mutually exclusive events. Two events are mutually exclusive when the $P(A \cap B) = 0$.

Example 4-17: If a random experiment consists of flipping a coin twice, find the probability of getting heads twice in a row.

Solution: The event of getting a head on the first flip is independent of getting a head on the second flip since the probability does not change with each flip of the coin. Thus, using the multiplication rule of independent events, $P(\text{Both coins are heads}) = P(1^{\text{st}} \text{ coin is a head}) \cdot P(2^{\text{nd}} \text{ coin is a head}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25$.

Example 4-18: The probability of Apple stock rising is 0.3, the probability of Boeing stock rising is 0.4. Assume Apple and Boeing stocks are independent. What is the probability that neither stock rises?

Solution: Let A = Apple stock and B = Boeing stock. Since A and B are independent the probability of both stocks rising at the same time is $P(A \cap B) = 0.3 \cdot 0.4 = 0.12$. “Neither” is the complement to “either.” $P(\text{Not Either}) = 1 - P(A \cup B) = 1 - (P(A) + P(B) - P(A \cap B)) = 1 - (0.3 + 0.4 - 0.12) = 1 - 0.58 = 0.42$.

Example 4-19: The probability that a student has their own laptop is 0.78. If three students are randomly selected, what is the probability that at least one owns a laptop?

Solution: There is an assumption that the three students are not related and that the probability of one owning a laptop is independent of the other people owning a laptop. The probability of none owning a laptop is $(1 - 0.78)^3 = 0.0106$. The probability of at least one is the same as $1 - P(\text{None}) = 1 - 0.0106 = 0.9894$.

When two events are dependent, you cannot simply multiply their corresponding probabilities to find their intersection. You will need to use the General Multiplication Rule.

Conditional Probability

The probability of event B happening, given some information about a dependent event A, is called the **conditional probability**. The conditional probability is written as $P(B | A)$, and is read as “the probability of B, given A.” We can use the General Multiplication Rule when two events are dependent.

General Multiplication Rule: $P(A \cap B) = P(A) \cdot P(B | A)$

Example 4-20: A bag contains 10 colored marbles: 7 red and 3 blue. A random experiment consists of drawing a marble from the bag, then drawing another marble without replacement (without putting the first marble back in the bag). Find the probability of drawing a red marble on the first draw (event R_1), *and* drawing another red marble on the second draw (event R_2).

Solution: Drawing a red marble on the first draw and drawing a red marble on the second draw are dependent events because we do not place the marble back in the bag. The probability of drawing a red marble on the first draw is $P(R_1) = \frac{7}{10}$, but on the second draw, the probability of drawing a red marble given that a red marble was drawn on the first draw is $P(R_2|R_1) = \frac{6}{9}$.

Thus, by the general multiplication rule, $P(R_1 \text{ and } R_2) = P(R_1) \cdot P(R_2|R_1) = \left(\frac{7}{10}\right)\left(\frac{6}{9}\right) = 0.4667$.

Example 4-21: A bag contains 10 colored marbles: 7 red and 3 blue. A random experiment consists of drawing a marble from the bag, then drawing another marble without replacement. Create the tree diagram for this experiment and compute the probabilities of each outcome.

Solution:

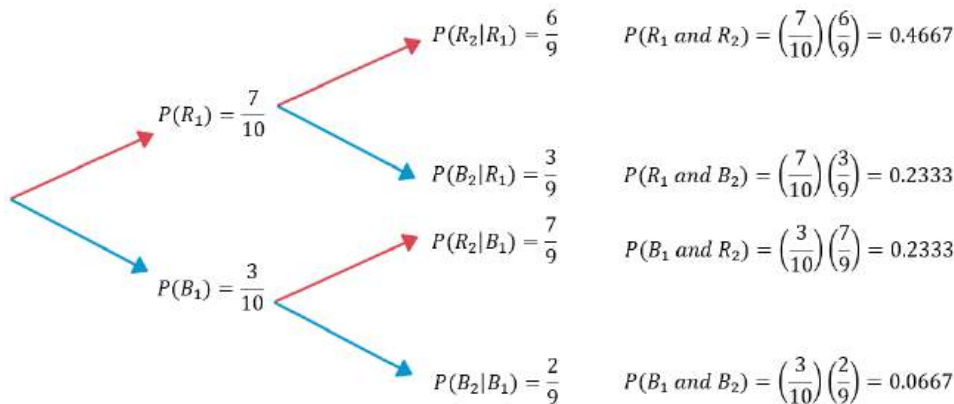


Figure 4-15

If we were to multiply the probabilities as we move from left to right up each set of tree branches as shown in Figure 4-15, we get the intersections. For example, by the general multiplication rule, $P(R_1 \text{ and } R_2) = P(R_1) \cdot P(R_2|R_1) = \left(\frac{7}{10}\right)\left(\frac{6}{9}\right) = 0.4667$.

Put the four intersection values into a contingency table and total the rows and columns. The table will help solve probability questions of other events.

	R_2	B_2	Total
R_1	0.4667	0.2333	0.7
B_1	0.2333	0.0667	0.3
Total	0.7	0.3	1

The grand total should add up to 1 since we have 100% of the sample space.

Conditional Probability Rule: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ or $P(B|A) = \frac{P(A \cap B)}{P(A)}$

Example 4-22: The following table shows the utility contract granted for a specific year. One contractor is randomly chosen.

	Corporation	Government	Individual	Total
United States	0.45	0.007	0.08	0.537
Foreign	0.41	0.003	0.05	0.463
Total	0.86	0.01	0.13	1

- a) Compute the probability the contractor is from the United States and is a corporation.

Solution: For the intersection in the contingency tables use where the row and column meet.
 $P(\text{U.S.} \cap \text{Corp}) = 0.45$.

- b) Compute the probability the contractor is from the United States given that they are a corporation.

Solution: $P(\text{U.S.} | \text{Corp}) = \frac{P(\text{U.S.} \cap \text{Corp})}{P(\text{Corp})} = \frac{0.45}{0.86} = 0.5233$.

- c) If the contractor is from a foreign country, what is the probability that it is from a government?

Solution: $P(\text{Gov} | \text{Foreign}) = \frac{P(\text{Gov} \cap \text{Foreign})}{P(\text{Foreign})} = \frac{0.003}{0.463} = 0.0065$.

- d) Are the events a “contractor is an individual” independent of a “contractor from the United States?”

Solution: Do not assume independence between two variables in a contingency table since the data may show relationships that you didn’t know were there.

Use the definition of independent events. If the two events are independent then we would have $P(\text{Individual} \cap \text{U.S.}) = P(\text{Individual}) \cdot P(\text{U.S.})$. First find the intersection using where the row and column meet to get $P(\text{Individual} \cap \text{U.S.}) = 0.08$. Then use the row and column totals to find $P(\text{Individual}) \cdot P(\text{U.S.}) = 0.13 \cdot 0.537 = 0.0698$. Since $P(\text{Individual} \cap \text{U.S.}) \neq P(\text{Individual}) \cdot P(\text{U.S.})$ these two events are dependent.

Example 4-23: A random sample of 500 people was taken from the 2010 United States Census. Their marital status and race were recorded in the following contingency table. A person is randomly chosen, find the following.

Marital Status	Race					Total
	American Indian	Black	Asian	White	Two Major Races	
Divorced	0	6	1	30	1	38
Married	1	25	23	156	4	209
Single	2	33	21	155	11	222
Widowed	0	7	2	22	0	31
Total	3	71	47	363	16	500

- a) $P(\text{Single and Asian})$

Solution: The intersection for a contingency table is found by simply finding where the row or intersection meets. There are 21 single Asians, therefore the $P(\text{Single} \cap \text{Asian}) = P(\text{Single and Asian}) = 21/500 = 0.042$. Do not multiply the row total times the column total since there is no indication that these are independent events.

- b) $P(\text{Single} | \text{Asian})$

Solution: In words we are trying to find the probability that the person is single given that we already know that their race is Asian. Using the conditional probability formula, we get

$$P(\text{Single} | \text{Asian}) = \frac{P(\text{Single} \cap \text{Asian})}{P(\text{Asian})} = \frac{21}{47} = 0.4468.$$

- c) Given that a person is single what is the probability their race is Asian?

Solution: This seems similar to the last question, however the part we know is that the person is single, but we do not know their race. In symbols we want to find the

$$P(\text{Asian} | \text{Single}) = \frac{P(\text{Asian} \cap \text{Single})}{P(\text{Single})} = \frac{21}{222} = 0.0946.$$

Keep in mind that usually $P(A | B) \neq P(B | A)$ since we would divide by a different total in the equation.

Example 4-24: A blood test correctly detects a certain disease 95% of the time (positive result), and correctly detects no disease present 90% of the time (negative result). It is estimated that 25% of the population have the disease. A person takes the blood test and they get a positive result. What is the probability that they have the disease?

Solution: Let D = Having the Disease, D^c = Not having the disease, + is a positive result, and - is a negative result. We are given in the problem the following: $P(+ | D) = 0.95$, $P(- | D^c) = 0.90$, $P(D) = 0.25$. We want to find $P(D | +) = \frac{P(D \cap +)}{P(+)}$.

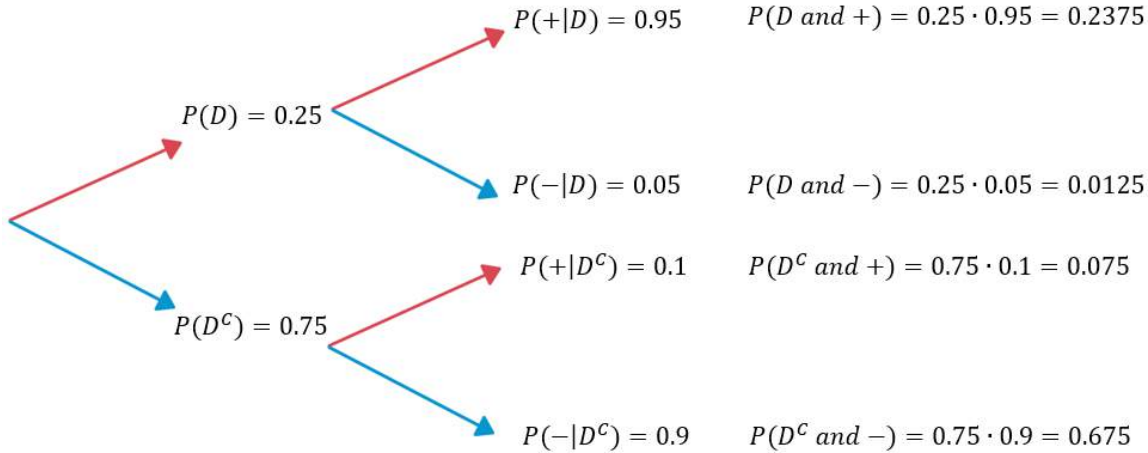


Figure 4-16

When you multiply up each pair of tree branches from left to right as shown in Figure 4-16, you are finding the intersection of the events. Place the multiplied values into a table. Note that the 0.2375 is not our answer. This is the people who have the disease and tested positive, but does not take into consideration the false positives. Since we know that the result was positive, we only divide by the proportion of positive results.

	D^c	D	Total
+	0.075	0.2375	0.3125
-	0.675	0.0125	0.6875
Total	0.75	0.25	1

$$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{0.2375}{0.3125} = 0.76$$

There is a 76% chance that they have the disease given that they tested positive. Many of the more difficult probability problems can be set up in a table, which makes the probabilities easier to find.

4.5 Counting Rules

There are times when the sample space is very large and is not feasible to write out. In that case, it helps to have mathematical tools for counting the size of the sample space. These tools are known as counting techniques or counting rules.

Fundamental Counting Rule: If task 1 can be done m_1 ways, task 2 can be done m_2 ways, and so forth to task n being done m_n ways. Then the number of ways to do task 1, 2, ..., n together would be to multiply the number of ways for each task $m_1 \cdot m_2 \cdot \dots \cdot m_n$.

Example 4-25: A menu offers a choice of 3 salads, 8 main dishes, and 5 desserts. How many different meals consisting of one salad, one main dish, and one dessert are possible?

Solution: There are three tasks, picking a salad, a main dish, and a dessert. The salad task can be done 3 ways, the main dish task can be done 8 ways, and the dessert task can be done 5 ways. The ways to pick a salad, main dish, and dessert are: $\frac{3}{\text{salad}} \cdot \frac{8}{\text{main}} \cdot \frac{5}{\text{dessert}} = 120$ different meals.

Example 4-26: How many 4-digit debit card personal identification numbers (PIN) can be made?

Solution: Four tasks must be done in this example. The tasks are to pick the first number, then the second number, then the third number, and then the fourth number. The first task can be done 10 ways since there are digits 1 through 9 or a zero. We can use the same numbers over again (repeats are allowed) to find that the second task can also be done 10 ways. The same with the third and fourth tasks, which also have 10 ways.

There are $\frac{10}{\text{first number}} \cdot \frac{10}{\text{second number}} \cdot \frac{10}{\text{third number}} \cdot \frac{10}{\text{fourth number}} = 10,000$ possible PINs.

Example 4-27: How many ways can the three letters a, b, and c be arranged with no letters repeating?

Solution: Three tasks must be done in this case. The tasks are to pick the first letter, then the second letter, and then the third letter. The first task can be done 3 ways since there are 3 letters. The second task can be done 2 ways, since the first task took one of the letters (repeats are not allowed). The third task can be done 1 way, since the first and second task took two of the letters.

There are $\frac{3}{1^{\text{st}} \text{ letter}} \cdot \frac{2}{2^{\text{nd}} \text{ letter}} \cdot \frac{1}{3^{\text{rd}} \text{ letter}} = 6$ ways.

You can also look at this example in a tree diagram, see Figure 4-17. There are 6 different arrangements of the letters. The solution was found by multiplying $3 \cdot 2 \cdot 1 = 6$.

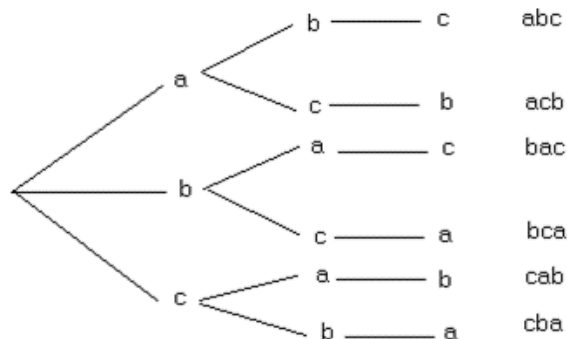


Figure 4-17

If we have 10 different letters for, say, a password, the tree diagram would be very time-consuming to make because of the length of options and tasks, so we have some shortcut formulas that help count these arrangements.

Example 4-28: How many possible automobile license plates are possible if the plate has 3 numbers followed by 3 letters?

Solution: Repeat numbers and letters are allowed on license plates. There are 10 possible numbers from 0 to 9. License plates only use the upper case letters. There are 26 letters in the alphabet. There are $10 \cdot 10 \cdot 10 \cdot 26 \cdot 26 \cdot 26 = 10^3 \cdot 26^3 = 17,576,000$ possible plates.

Many counting problems involve multiplying a list of numbers decreasing by one, which is called a **factorial**. The factorial is represented mathematically by the starting number followed by an exclamation point, in this case $3! = 3 \cdot 2 \cdot 1 = 6$. There is a special symbol for this and a function on your calculator or computer.

Factorial Rule: The number of different ways to arrange n objects is $n! = n \cdot (n - 1) \cdot (n - 2) \cdots 3 \cdot 2 \cdot 1$, where repetitions are not allowed.

Zero factorial is defined to be $0! = 1$, and 1 factorial is defined to be $1! = 1$.

TI-84: On the home screen, enter the number of which you would like to find the factorial. Press [MATH]. Use cursor keys to move to the PRB menu. Press 4 (4:!) Press [ENTER] to calculate.

TI-89: On the home screen, enter the number of which you would like to find the factorial. Press [2nd] [Math] > 7:Probability > 1:!. Press [ENTER] to calculate.

Excel: In an empty cell type in =FACT(n) where n is the number so $4!$ would be =FACT(4).

Example 4-29: How many ways can you arrange five people standing in line?

Solution: No repeats are allowed since you cannot reuse a person twice. Order is important since the first person is first in line and will be selected first. This meets the requirements for the factorial rule,
 $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ ways.

Sometimes we do not want to select the entire group but only select r objects from n total objects. The number of ways to do this depends on if the order you choose the r objects matters or if it does not matter. As an example, if you are trying to call a person on the phone, you have to have the digits of their number in the correct order. In this case, the order of the numbers matters. If you were picking random numbers for the lottery, it does not matter which number you pick first since they always arrange the numbers from the smallest to largest once the numbers are drawn. As long as you have the same numbers that the lottery officials pick, you win. In this case, the order does not matter.

A **permutation** is an arrangement of items with a specific order. You use permutations to count items when the order matters. Alternative notation for a permutation includes $P(n,k)$ or P_r^n .

Permutation Rule: The number of different ways of picking r objects from n total objects when repeats are not allowed and order matters ${}_n P_r = \frac{n!}{(n-r)!}$.

When the order does not matter, you use combinations. A **combination** is an arrangement of items when order is not important. When you do a counting problem, the first thing you should ask yourself is “are repeats allowed,” then ask yourself “does order matter?” Alternative notation for a combination includes $C(n,k)$, C_r^n or $\binom{n}{r}$.

Combination Rule: The number of ways to select r objects from n total objects when repeats are not allowed and order does not matter ${}_n C_r = \frac{n!}{r!(n-r)!}$.

The combination formula is sometimes called the binomial coefficient and has some interesting properties and worth an internet search. A quick way to help distinguish between permutations and combinations is to think of a permutation as a distinct sequence of outcomes with a specific order, and combinations are a collection of items in any order.

TI-84: Enter the number “trials” (n) on the home screen. Press [MATH]. Use cursor keys to move to the PRB menu. Press 2 for permutation ($2: {}_n P_r$), 3 for combination ($3: {}_n C_r$). Enter the number of “successes” (r). Press [ENTER] to calculate.

TI-89: Press [2nd] Math > 7:Probability > Press 2 for permutation ($2: {}_n P_r$), 3 for combination ($3: {}_n C_r$). Enter the sample size on the home screen, then a comma, then enter the number of “successes,” then end the parenthesis. Press [ENTER] to calculate.

Excel: In a blank cell type in the formula =COMBIN(n , r) or =PERMUT(n , r) where n is the total number of objects and r is the smaller number of objects that you are selecting out of n . For example =COMBIN(8, 3).

The following flow chart in Figure 4-18 may help with deciding which counting rule to use.

Start on the left; ask yourself if the same item can be repeated. For instance, a person on a committee cannot be counted as two distinct people; however, a number on a car license plate may be used twice. If repeats are not allowed, then ask, does the order in which the item is chosen matter? If it does not then we use the combinations, if it does then ask are you ordering the entire group, use factorial, or just some of the group, use permutation.

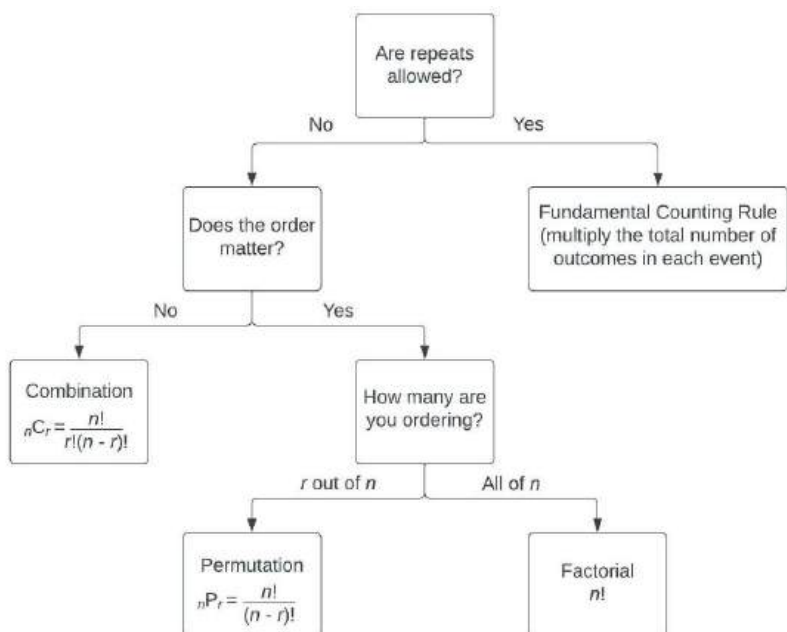


Image provided by Nancy Ikeda <https://math.fullcoll.edu/faculty/>
Figure 4-18

Example 4-30: Suppose you want to pick 7 people out of 20 people to take part in a survey. How many ways can you do this?

Solution: In this case, repeats are not allowed since we don't want to survey the same person more than once. The order in which you select each person does not matter, since you just want 7 people. This is a combination with $n = 20$ and $r = 7$.

$${}_{20}C_7 = \frac{20!}{(7!(20-7)!)} = \frac{20!}{(7! \cdot 13!)} = 77520 \quad \text{There are 77,520 possible groups of 7 people.}$$

Example 4-31: Critical Miss, PSU's Tabletop Gaming Club, has 15 members this term. How many ways can a slate of 3 officers consisting of a president, vice-president, and treasurer be chosen?

Solution: In this case, repeats are not allowed since we don't want the same member to hold more than one position. The order matters, since if you pick person 1 for president, person 2 for vice-president, and person 3 for treasurer, you would have different members in those positions than if you picked person 2 for president, person 1 for vice-president, and person 3 for treasurer. This is a permutation problem with $n = 15$ and $r = 3$.

$${}_{15}P_3 = \frac{15!}{(15-3)!} = \frac{15!}{12!} = 2730 \quad \text{There are 2,730 ways to elect these three positions.}$$

In general, if you were selecting items that involve rank, a position title, 1st, 2nd, or 3rd place or prize, etc. then the order in which the items are arranged is important and you would use permutation.

Example 4-32: Critical Miss, PSU's Tabletop Gaming Club, has 15 members this term. They need to select 3 members to have keys to the game office. How many ways can the 3 members be chosen?

Solution: In this case, repeats are not allowed, because we don't want one person to have more than one key. The order in which the keys are handed out does not matter. This is a combination problem with $n = 15$ and $r = 3$.

$${}_{15}C_3 = \frac{15!}{(3!(15-3)!)} = \frac{15!}{(3! \cdot 12!)} = 455 \quad \text{There are 455 ways to hand out the three keys.}$$

We can use these counting rules in finding probabilities. For instance, the probability of winning the lottery can be found using these counting rules.

Example 4-33: What is the probability of winning the jackpot in the Pick-4 Lottery? To play Pick-4, you choose 4 numbers from 0 to 9. This will give you a number between 0000 and 9999. You win the jackpot if you match your 4 numbers in the exact order they are drawn.

<https://www.oregonlottery.org/jackpot/pick-4/>

Solution: There is only one winning number, so the numerator for the probability will just be 1. The denominator will be all the ways to select the 4 numbers. Repeat numbers are allowed, for example you can have 4242 with repeating 4s and 2s. The order in which the balls are selected does matter by the rules of the game. Use the fundamental counting rule combined with the fundamental counting rule and we would get $10 \cdot 10 \cdot 10 \cdot 10 = 10,000$.

Thus, the probability of winning the jackpot would be $\frac{1}{10000} = 0.0001$.

Example 4-34: What is the probability of getting a full house if 5 cards are randomly dealt from a standard deck of cards?

Solution: A full house is a combined three of a kind and pair, for example, QQQ22. There are ${}_{13}C_1$ ways to choose a card between Ace, 2, 3, ..., King. Once a number is chosen, there are 4 cards with that rank and there are ${}_{4}C_3$ ways to choose a three of kind from that rank. Once we use up one of the ranks, such as the three queens, there are ${}_{12}C_1$ ways to choose the rank for the pair. Once the pair is chosen there are ${}_{4}C_2$ ways to choose a pair from that rank. All

together there are ${}_{52}C_5$ ways to randomly deal out 5 cards. The probability of getting a full house with then be $\frac{{}_{13}C_1 \cdot {}_4C_3 \cdot {}_{12}C_1 \cdot {}_4C_2}{{}_{52}C_5} = \frac{3744}{2598960} = 0.00144$.

Example 4-35: What is the probability of winning the Powerball jackpot? As of 2023, the Powerball lottery consists of drawing five white balls in any order numbered 1 through 69, and one red Powerball numbered 1 through 26. <https://www.oregonlottery.org/jackpot/powerball/>

Solution: There is only one winning number, so the numerator for the probability will just be 1. The denominator will be all the ways to select the 5 white balls and 1 red ball. The order in which the balls are selected does not matter and repeat numbers are not allowed. Using the combination rule combined with the fundamental counting rule we would get ${}_{69}C_5 \cdot {}_{26}C_1$.

Thus, the probability of winning the jackpot would be $\frac{1}{{}_{69}C_5 \cdot {}_{26}C_1} = 0.000000003707$.

Example 4-36: An iPhone has a 6 digit numerical password to unlock the phone. What is the probability of guessing the password on the first try?

Solution: There is only one correct password, so the numerator for the probability will just be 1. The denominator will be all the ways to select the 6 numbers. Repeat numbers are allowed. The order in which the numbers are entered into the phone matters. Using the fundamental counting rule, we would get $10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 10^6 = 1,000,000$ possible passwords.

Thus, the probability of guessing the password on the first try would be $\frac{1}{1000000} = 1 \text{ E-}6 = 0.000001$.

“The chances of this happening are more or less one to infinity against. Little is known of how this came about because none of the geophysicists, probability statisticians, meteoranalysts or bizzarrologists who are so keen to research it can afford to stay there.”
(Adams, 2002)

Chapter 4 Formulas

Complement Rules: $P(A) + P(A^c) = 1$ $P(A) = 1 - P(A^c)$ $P(A^c) = 1 - P(A)$	Mutually Exclusive Events: $P(A \cap B) = 0$
Union Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Independent Events: $P(A \cap B) = P(A) \cdot P(B)$
Intersection Rule: $P(A \cap B) = P(A) \cdot P(B A)$	Conditional Probability Rule: $P(A B) = \frac{P(A \cap B)}{P(B)}$
Fundamental Counting Rule: $m_1 \cdot m_2 \cdots m_n$	Factorial Rule: $n! = n \cdot (n - 1) \cdot (n - 2) \cdots 3 \cdot 2 \cdot 1$
Combination Rule: ${}_nC_r = \frac{n!}{(r!(n-r)!)}$	Permutation Rule: ${}_nP_r = \frac{n!}{(n-r)!}$

Chapter 4 Exercises

1. The number of M&M candies for each color found in a case were recorded in the table below.

Blue	Brown	Green	Orange	Red	Yellow	Total
481	371	483	544	372	369	2,620

What is the probability of selecting a red M&M?

- An experiment is to flip a fair coin three times. Write out the sample space for this experiment.
- An experiment is to flip a fair coin three times. What is the probability of getting exactly two heads?
- In the game of roulette, there is a wheel with spaces marked 0 through 36 and a space marked 00. Compute the probability of winning if you pick the number 30 and it comes up on the wheel.
- A raffle sells 1,000 tickets for \$35 each to win a new car. What is the probability of winning the car?
- Compute the probability of rolling doubles when two 6-sided dice are rolled.
- Compute the probability of rolling doubles when two 20-sided dice are rolled.
- Compute the probability of rolling a sum of two 6-sided dice that is more than 7.
- Compute the probability of rolling a sum of two 6-sided dice that is a 7 or a 12.
- A random sample of 500 people's marital status and biological sex from the 2010 United States Census are recorded in the following contingency table.

Count of Marital Status	Column Labels		Grand Total
	Female	Male	
Divorced	21	17	38
Married/spouse absent	5	9	14
Married/spouse present	92	100	192
Never married/single	93	129	222
Separated	1	2	3
Widowed	20	11	31
Grand Total	232	268	500

- Compute the probability that a randomly selected person is single.
 - Compute the probability that a randomly selected person is not single.
 - Compute the probability that a randomly selected person is single or male.
 - Compute the probability that a randomly selected person is divorced or widowed.
 - Given that randomly selected person is male, what is the probability they are single?
 - Are the events divorced and male mutually exclusive?
 - Are the events divorced and male independent? Verify using statistics.
11. The probability that a consumer entering a retail outlet for microcomputers and software packages will buy a computer of a certain type is 0.15. The probability that the consumer will buy a particular software package is 0.10. There is a 0.05 probability that the consumer will buy both the computer and the software package. What is the probability that the consumer will buy the computer or the software package?

12. A company conducted a survey among its employees and found that 80% of employees use smartphones, 45% of employees use tablets, and 30% of employees use both smartphones and tablets. If an employee is selected randomly, what is the probability of selecting an employee who uses either a smartphone or a tablet?
13. A poll showed that 48.7% of Americans say they believe that Marilyn Monroe had an affair with JFK. What is the probability of randomly selecting someone who does not believe that Marilyn Monroe had an affair with JFK?
14. The 2023 unemployment rate in Oregon was 4%. An employable adult Oregonian is randomly selected. What is the probability that they are employed?
15. Your favorite basketball player is an 81% free throw shooter. Find the probability that they do not make their next free throw shot.
16. The table below shows the percentage of market share held by each brand of smartphone. If a person owning a smartphone is randomly selected, what is the probability that they own a Samsung?

Smartphone Brand	Market Share (%)
Apple	40%
Samsung	30%
Huawei	15%
Xiaomi	10%
Others	5%

17. Randomly pick a card from a standard deck.
 - a) Compute the probability of selecting a card that shows a diamond.
 - b) Compute the probability of selecting a card that shows a diamond and an ace.
 - c) Compute the probability of selecting a card that shows a diamond or an ace.
 - d) Compute the probability of selecting a card that shows a 7 or an ace.
 - e) Compute the probability of selecting a card that shows a 7 and an ace.
 - f) Are the events, selecting a 7, and selecting an ace, mutually exclusive? Why?
18. Randomly pick a card from a standard deck.
 - a) Compute the probability of selecting a card that shows a queen.
 - b) Compute the probability of selecting a card that shows a 7.
 - c) Compute the probability of selecting a card that shows a number smaller than 7.
 - d) Compute the probability of selecting a card that shows a face card and a spade.
 - e) Compute the probability of selecting a card that shows a face card or a spade.
 - f) Are the events, selecting a face card, and selecting a spade, mutually exclusive? Why?
19. The following table shows the frequency of items sold over several month by each item category and day of the week for a local restaurant. One item is randomly selected. Find the following.

Item Category	Wednesday	Thursday	Friday	Saturday
Appetizers	250	189	364	205
Main Courses	483	407	672	591
Desserts	128	103	238	196

- a) $P(\text{Main Course})$
- b) $P(\text{Main Course} \cap \text{Friday})$
- c) Compute the probability that a randomly selected item is a dessert, given that it was a Friday.
- d) Given that the item is a dessert, compute the probability that the item was sold on Friday.
- e) $P(\text{Saturday} | \text{Appetizer})$

20. A fitness center owner kept track of members over the last year. They recorded if the person stretched before they exercised, and whether they sustained an injury. The following contingency table shows their results. Select one member at random and find the following.

	Injury	No Injury	Total
Stretched	52	270	322
Did Not Stretch	21	57	78
Total	73	327	400

- $P(\text{No Injury})$
 - $P(\text{Injury} \cap \text{Stretch})$
 - Compute the probability that a randomly selected member stretched or sustained an injury.
 - Compute the probability that a randomly selected member stretched given that they sustained an injury.
 - $P(\text{Injury} | \text{Did Not Stretch})$
21. Giving a test to a group of students, the grades and if they were business majors are summarized below. One student is chosen at random. Give your answer as a decimal out to at least 4 places.

	A	B	C	Total
Business Majors	4	5	13	22
Non-business Majors	18	10	19	47
Total	22	15	32	69

- Compute the probability that the student was a non-business major or got a grade of C.
 - Compute the probability that the student was a non-business major and got a grade of C.
 - Compute the probability that the student was a non-business major given they got a C grade.
 - Compute the probability that the student did not get a B grade.
 - Compute $P(B \cup \text{Business Major})$.
 - Compute $P(C | \text{Business Major})$.
22. A report for a school's computer web visits for the past month obtained the following information. Find the percentage that visited none of these three sites last month. Hint: Draw a Venn Diagram.
- 37% visited Facebook.
 - 42% visited LinkedIn.
 - 29% visited Google.
 - 27% visited Facebook and LinkedIn.
 - 19% visited Facebook and Google.
 - 19% visited LinkedIn and Google.
 - 14% visited all three sites.
23. The smallpox data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.

	Inoculated	Not Inoculated	Total
Lived	238	5136	5374
Died	6	844	850
Total	244	5980	6224

Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

- Compute the relative frequencies.

	Inoculated	Not Inoculated	Total
Lived			
Died			
Total			1

- b) Compute the probability that a person was inoculated.
- c) Compute the probability that a person lived.
- d) Compute the probability that a person died or was inoculated.
- e) Compute the probability that a person died if they were inoculated.
- f) Given that a person was not inoculated, what is the probability that they died?

24. A certain virus infects one in every 400 people. A test used to detect the virus in a person is positive 90% of the time if the person has the virus and 8% of the time if the person does not have the virus. (This 8% result is called a false positive.) Let A be the event "the person is infected" and B be the event "the person tests positive."
- a) Find the probability that a person has the virus given that they have tested positive, i.e., find $P(A|B)$.
 - b) Find the probability that a person does not have the virus given that they test negative, i.e., find $P(A^c|B^c)$.

25. A store purchases baseball hats from three different manufacturers. In manufacturer A's box there are 12 blue hats, 6 red hats, and 6 green hats. In manufacturer B's box there are 10 blue hats, 10 red hats, and 4 green hats. In manufacturer C's box, there are 8 blue hats, 8 red hats, and 8 green hats. A hat is randomly selected. Given that the hat selected is green, what is the probability that it came from manufacturer B's box? Hint: Make a table with the colors as the columns and the manufacturers as the rows.

26. The following table represents food purchase amounts and whether the customer used cash or a credit/debit card. One customer is chosen at random. Give your answer as a decimal out to at least 4 places.

	Less than \$10	\$10-\$49	\$50 or More	Total
Cash Purchase	11	10	18	39
Card Purchase	17	6	19	42
Total	28	16	37	81

- a) Compute the probability that the customer's purchasing method was a cash purchase or the customer spent \$10-\$49.
 - b) Compute the probability that the customer's purchasing method was cash purchase and the customer spent \$10-\$49.
 - c) Compute the probability that the customer's purchasing method was a cash purchase given they spent \$10-\$49.
 - d) Compute the probability that the customer spent less than \$50.
 - e) What percent of cash purchases were for \$50 or more?
27. The probability of stock A rising is 0.3; and of stock B rising is 0.4. What is the probability that neither of the stocks rise, assuming that these two stocks are independent?
28. The 2023 unemployment rate in Oregon was 4%. Four employable adult Oregonians are randomly selected. What is the probability that all four people are unemployed?
29. The following data show the all-time Olympic medals count by continent. A medal is randomly selected.

Continent	Gold	Silver	Bronze
Africa	126	146	169
America	1542	1390	1370
Asia	772	734	832
Europe	3949	4095	4393
Oceania	232	218	273

<https://www.olympiandatabase.com/index.php?id=21633&L=1>

- a) Compute the probability that the medal was gold.
- b) Compute the probability that the medal was gold and the athlete was from Africa.

- c) Compute the probability that the medal was gold or the athlete was from Africa.
 - d) Compute $P(\text{Gold}^c)$.
 - e) Compute $P(\text{Europe} \cap \text{Bronze})$.
 - f) Compute $P(\text{Europe} \cup \text{Bronze})$.
 - g) Compute $P(\text{Europe} | \text{Bronze})$.
30. You are going to a Humane Society benefit dinner, and need to decide before the dinner what you want for salad, main dish, and dessert. You have 2 salads to choose from, 3 main dishes, and 5 desserts. How many different meals are available?
 31. How many different phone numbers are possible in the area code 503, if the first number cannot start with a 0 or 1?
 32. You are opening a screen-printing business. You can have long sleeves or short sleeves, three different colors, five different designs, and four different sizes. How many different shirts can you make?
 33. The California license plate has one number followed by three letters followed by three numbers. How many different license plates are possible?
 34. Calculate the following.
 - a) ${}_9P_4$
 - b) ${}_{10}P_6$
 - c) ${}_{10}C_5$
 - d) ${}_{20}C_4$
 - e) $8!$
 - f) $5!$
 35. The PSU's Mixed Me club has 30 members. You need to pick a president, treasurer, and secretary from the 30. How many different ways can you do this?
 36. How many different 4-digit personal identification numbers (PIN) are there if repeats are not allowed?
 37. A baseball team has a 20-person roster. A batting order has nine people. How many different batting orders are there?
 38. How many ways can you choose 4 cookies from a cookie jar containing 25 cookies of all the same type?
 39. A computer generates a random password for your account (the password is not case sensitive). The password must consist of 8 characters, each of which can be any letter or number. How many different passwords could be generated?
 40. How many unique tests can be made from a test bank of 20 questions if the test consists of 8 questions, order does not matter?
 41. A typical PSU locker is opened with correct sequence of three numbers between 0 and 49 inclusive. A number can be used more than once, for example 8-8-8 is valid. How many possible locker combinations are there?
 42. In the game of Megabucks, you get six numbers from 48 possible numbers without replacement. Megabucks jackpots start at \$1 million and grow until someone wins. What is the probability of matching all 6 numbers in any order?

Chapter 5

Probability Distributions



- 5.1 Introduction to Probability Distributions
- 5.2 Discrete Probability Distributions
 - 5.2.1 Mean of a Discrete Probability Distribution
 - 5.2.2 Variance & Standard Deviation of a Discrete Probability Distribution
- 5.3 Binomial Distribution
- 5.4 Empirical Rule
- 5.5 Normal Distribution
 - 5.5.1 Standard Normal Distribution
 - 5.5.2 Applications of the Normal Distribution
 - 5.5.3 Normal Probability Plot
- 5.6 The Central Limit Theorem
- 5.7 Student's T-Distribution

5.1 Introduction to Probability Distributions

A variable or what will be called the random variable from now on, is represented by the letter X and it represents a quantitative (numerical) variable that is measured or observed in an experiment.

A **random variable** (usually X) is a numeric description of an event. Recall that a sample space is all the possible outcomes and an event is a subset of the sample space.

Usually, we use capital letters from the beginning of the alphabet to represent events, A, B, C , etc. We use capital letters from the end of the alphabet to represent random variables, X, Y , and Z . The possible outcomes of X are labeled with a corresponding lower-case letter x and subscripts like x_1, x_2 , etc.

For instance, if we roll two 6-sided dice the sample space is $S = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), \dots, (6,6)\}$ and the event E the sum of the two rolls is five, then $E = \{(1,4), (2,3), (3,2), (4,1)\}$.

Now, we could define the random variable X to denote the sum of the two rolls, then $X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ and event E corresponds to $x = 5$.

There are different types of quantitative variables, called discrete or continuous. **Discrete** random variables can only take on particular values in a range. **Continuous** random variables can take on any value in a range. Discrete random variables usually arise from counting while continuous random variables usually arise from measuring.

A **discrete random variable** is a variable that is finite or infinitely countable.

A **continuous random variable** is a variable that has an infinite number of possible values in an interval of numbers.

5.2 Discrete Probability Distributions

In many cases, the random variable is what you are measuring, but when it comes to discrete random variables, it is usually what you are counting. For the example of height, the random variable is the height of the child. For the example of how many fleas are on prairie dogs in a colony, the random variable is the number of fleas on a prairie dog in a colony.

Now suppose you put all the possible values of the random variable together with the probability that the random variable would occur. You could then have a distribution like before, but now it is called a probability distribution since it involves probabilities. A **probability distribution** is an assignment of probabilities to all the possible values of the random variable. The abbreviation of pdf is used for a probability density (distribution) function in your calculators.

The probability distribution of X lists all the possible values of x and their corresponding probabilities.

A valid **discrete probability distribution** has to satisfy two criteria:

1. The probability of x is between 0 and 1, $0 \leq P(x_i) \leq 1$.
2. The probability of all x values adds up to 1, $\sum P(x_i) = 1$.

Example 5-1: Two books are assigned for a statistics class: a textbook and its corresponding study guide. No students buy just the study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, 25% buy both books, and these percentages are relatively constant from one term to another. Is this a valid discrete probability distribution?

Solution: Each probability is a number between 0 and 1; $0 \leq 0.2 \leq 1$, $0 \leq 0.55 \leq 1$, and $0 \leq 0.25 \leq 1$. The sum of the probabilities adds up to 1; $\sum P(x_i) = 0.2 + 0.55 + 0.25 = 1$. Yes, this is a valid discrete probability distribution.

Example 5-2: A random experiment consists of flipping a fair coin three times. Let X = the number of heads that show up. Create the probability distribution of X .

Solution: When you flip a coin three times, you can get 0 heads, 1 head, 2 heads or 3 heads, the corresponding values of x are 0, 1, 2, 3. We can find the probabilities for each of these values by finding the sample space and using the classical method of computing probability.

The sample space is $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.

The event $x = 0$ can happen one way, namely $\{TTT\}$. Thus $P(X = x) = P(X = 0) = \frac{1}{8}$.

The event $x = 1$ can happen three ways, namely $\{HTT, THT, TTH\}$. Thus $P(X = 1) = \frac{3}{8}$.

The event $x = 2$ can happen three ways, namely $\{HHT, HTH, THH\}$. Thus $P(X = 2) = \frac{3}{8}$.

The event $x = 3$ can happen one way, namely $\{HHH\}$. Thus $P(X = 3) = \frac{1}{8}$.

Therefore, the probability distribution of X is:

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Figure 5-1

Note that Figure 5-1 is a valid discrete probability distribution because the probability of each x , $P(x)$, is between 0 and 1, and the probability of the sum of all x values from 0 to 3 is $\sum P(x) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$.

Figure 5-2 is a graph of the probability distribution using Example 5-2.



Figure 5-2

The height of each line corresponds to the probability of each x value. Sometimes you will see a bar graph instead.

Example 5-3: The following table shows the probability of winning an online video game, where X = the net dollar winnings for the video game. Is the following table a valid discrete probability distribution?

x	-5	-2.5	0	2.5	5
$P(X = x)$	0.55	$\frac{1}{4}$	0.15	0	5%

Solution: It is easier to have all the probabilities as proportions: $\frac{1}{4} = 0.25$ and $5\% = 0.05$. Each probability is a number between 0 and 1: $0 \leq 0.55 \leq 1$, $0 \leq 0.25 \leq 1$, $0 \leq 0.15 \leq 1$, $0 \leq 0 \leq 1$, $0 \leq 0.05 \leq 1$. The sum of the probabilities is equal to one: is $\sum P(x) = 0.55 + 0.25 + 0.15 + 0 + 0.05 = 1$. Yes, this is a valid discrete probability distribution since the table has the two properties that each probability is between 0 and 1, and the sum of the probabilities is one.

Example 5-4: Is the following table a valid discrete probability distribution?

x	0	1	2	3
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Solution: It is easier to have all the probabilities as proportions: $\frac{1}{2} = 0.5$ and $\frac{1}{4} = 0.25$. Each probability is a number between 0 and 1: $0 \leq 0.5 \leq 1$, $0 \leq 0.25 \leq 1$, $0 \leq 0.25 \leq 1$, $0 \leq 0.25 \leq 1$. The sum of the probabilities does not equal one: $\sum P(x) = 0.5 + 0.25 + 0.25 + 0.25 = 1.25$. No, this is not a valid discrete probability distribution since the sum of the probabilities is not equal to one.

Words Are Important!

When finding probabilities pay close attention to the following phrases.

$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
Is the same as	Is less than or equal to	Is greater than or equal to
Is equal to	Is at most	Is at least
Is exactly the same as	Is not greater than	Is not less than
Has not changed from	Within	

	$P(X > x)$	$P(X < x)$
	More than	Less than
	Greater than	Below
	Above	Lower than
	Higher than	Shorter than
	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

Figure 5-3

We can use these discrete probability distribution tables to find probabilities.

Example 5-5: The following is a valid discrete probability distribution of X = the net dollar winnings for an online video game. Find the probability of winning at most \$2.50.

x	-5	-2.5	0	2.5	5
$P(X = x)$	0.55	$\frac{1}{4}$	0.15	0	5%

Solution: It is easier to have all the probabilities as decimals: $\frac{1}{4} = 0.25$ and $5\% = 0.05$.

At most means the same as, less than or equal to. The probabilities can be found by adding all the probabilities that are less than or equal to 2.5. $P(X \leq 2.5) = 0.55 + 0.25 + 0.15 + 0 = 0.95$.

You can also use the complement rule since all the probabilities add to one.

$$P(X \leq 2.5) = 1 - P(X > 2.5) = 1 - 0.05 = 0.95.$$

Example 5-6: The 2010 United States Census found the chance of a household being a certain size.

Size of Household	1	2	3	4	5	6	7
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

a) Is this a valid probability distribution?

Solution: In this case, the random variable is X = number of people in a household. This is a discrete random variable, since you are counting the number of people in a household. This is a probability distribution since you

have the x value and the probabilities that go with it, all of the probabilities are between zero and one, and the sum of all of the probabilities is one.

- b) Compute the probability that a household has exactly 3 people.

Solution: Let X = the number of people per household. Trace along the Size of Household row until you get to 3, then take the corresponding probability in that column, $P(X = 3) = 0.158$.

- c) Compute the probability that a household has at most 3 people.

Solution: "At most" 3 would be 3 or less.

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.267 + 0.336 + 0.158 = 0.761.$$

- d) Compute the probability that a household has more than 3 people.

Solution: "More than" 3, does not include the 3, and would be all the probabilities for 4 or more added together. $P(X > 3) = P(X \geq 4) = 0.137 + 0.063 + 0.024 + 0.015 = 0.239$.

Note that this answer is the complement to part c so we could use the $P(X > 3) = 1 - P(X \leq 3) = 1 - 0.761 = 0.239$. This technique will be very useful when we have a large number of outcomes.

- e) Compute the probability that a household has at least 3 people.

Solution: "At least" 3, means 3 or more.

$$P(X \geq 3) = 0.158 + 0.137 + 0.063 + 0.024 + 0.015 = 0.397.$$

- f) Compute the probability that a household has 3 or more people.

Solution: This would be the same as the previous question just worded differently,

$$P(X \geq 3) = 0.158 + 0.137 + 0.063 + 0.024 + 0.015 = 0.397.$$

Rare Events

The reason probability is studied in statistics is to help in making decisions in inferential statistics.

To understand how making decisions is done, the concept of a rare event is needed.

Rare Event Rule for Inferential Statistics: If, under a given assumption, the probability of a particular observed event is extremely small, then you can conclude that the assumption is probably not correct.

An example of this is suppose you roll an assumed fair die 1,000 times and get six 600 times, when you should have only rolled a six around 160 times, then you should believe that your assumption about it being a fair die is untrue.

Determining if an event is unusual: If you are looking at a value of X for a discrete variable, and the $P(\text{the variable has a value of } x \text{ or more}) \leq 0.05$, then you can consider the x an unusually high value. Another way to think of this is if the probability of getting such a high value is less than or equal to 0.05, then the event of getting the value x is unusual.

Similarly, if the $P(\text{the variable has a value of } x \text{ or less}) \leq 0.05$, then you can consider this an unusually low value. Another way to think of this is if the probability of getting a value as small as x is less than or equal to 0.05, then the event x is considered unusual.

Why is it " x or more" or " x or less" instead of just " x " when you are determining if an event is unusual?

Consider this example: you and your friend go out to lunch every day. Instead of each paying for their own lunch, you decide to flip a coin, and the loser pays for both lunches. Your friend seems to be winning more often than you would expect, so you want to determine if this is unusual before you decide to change how you pay for lunch (or accuse your friend of cheating). The process for how to calculate these probabilities will be presented in a later section on the binomial distribution.

If your friend won 6 out of 10 lunches, the probability of that happening turns out to be about 20.5%, not unusual. The probability of winning 6 or more is about 37.7%, still not unusual.

However, what happens if your friend won 501 out of 1,000 lunches? That does not seem so unlikely! The probability of winning 501 or more lunches is about 47.8%, and that is consistent with your hunch that this probability is not so unusual. Nevertheless, the probability of winning exactly 501 lunches is much less, only about 2.5%.

That is why the probability of getting exactly that value is not the right question to ask: you should ask the probability of getting that value or more (or that value or less on the other side).

The value 0.05 will be explained later, and it is not the only value you can use for deciding what is unusual.

Example 5-7: The 2010 United States Census found the chance of a household being a certain size.

Size of Household	1	2	3	4	5	6	7
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

a) Is it unusual for a household to have six or more people in the family?

Solution: To determine this, you need to look at probabilities. However, you cannot just look at the probability of six people. You need to look at the probability of x being six or more people $P(X \geq 6) = 0.024 + 0.015 = 0.039$. Since this probability is less than 5%, then six or more family members is an unusually high value. It is unusual for a household to have six or more people in the family.

b) Is it unusual for a household to have four or more people in the family?

Solution: We need to look at the probability of x being four or more. $P(X \geq 4) = 0.137 + 0.063 + 0.024 + 0.015 = 0.239$, since this probability is more than 5%, four is not an unusually high value. Thus, four is not an unusual size of a family.

What was that voice?" shouted Arthur.
 "I don't know," yelled Ford, "I don't know. It sounded like a measurement of probability."
 "Probability? What do you mean?"
 "Probability. You know, like two to one, three to one, five to four against. It said two to the power of one hundred thousand to one against. That's pretty improbable you know."
 A million-gallon vat of custard upended itself over them without warning.
 "But what does it mean?" cried Arthur.
 "What, the custard?"
 "No, the measurement of probability!"
 "I don't know. I don't know at all. I think we're on some kind of spaceship."
 (Adams, 2002)

5.2.1 Mean of a Discrete Probability Distribution

The mean of a discrete random variable X is an average of the possible values of x , which considers the fact that not all outcomes are equally likely. The mean of a random variable is the value of x that one would expect to see after averaging a large number of trials. The mean of a random variable does not need to be a possible value of x .

Example 5-8: Two books are assigned for a statistics class: a textbook and its corresponding study guide. No students buy just the study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, 25% buy both books, and these percentages are relatively constant from one term to another. If there are 100 students enrolled, how many books should the bookstore expect to sell to this class?

Solution: It is expected that $0.20 \times 100 = 20$ students will not buy either book (0 books total), that $0.55 \times 100 = 55$ will buy just the textbook (55 books total), and $0.25 \times 100 = 25$ will buy both books (totaling 50 books for these 25 students). The bookstore should expect to sell about 105 total books for this class.

Example 5-9: The textbook costs \$137 and the study guide \$33. How much revenue should the bookstore expect from this class of 100 students? Use the results from the previous example.

Solution: It is expected that 55 students will just buy a textbook, providing revenue of $\$137 \times 55 = \$7,535$. The roughly 25 students who buy both the textbook and the study guide would pay a total of $(\$137 + \$33)25 = \$170 \times 25 = \$4,250$. Thus, the bookstore should expect to generate about $\$7,535 + \$4,250 = \$11,785$ from these 100 students for this one class. However, there might be some sampling variability so the actual amount may differ by a little bit.

Probability distribution for the bookstore's revenue from a single student. The distribution balances on a triangle representing the average revenue per student represented in Figure 5-4.

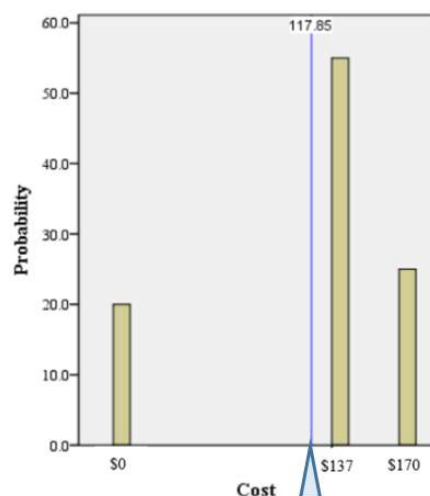


Figure 5-4

Example 5-10: What is the average revenue per student for this course?

Solution: The expected total revenue is \$11,785, and there are 100 students. Therefore, the expected revenue per student is $\$11,785/100 = \117.85 .

Mean of a Discrete Random Variable: Suppose that X is a discrete random variable with values x_1, x_2, \dots, x_k . Then the mean of X is, $\mu = \sum(x_i \cdot P(x_i)) = x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \dots + x_k \cdot P(x_k)$. The mean is also referred to as the expected value of x denoted as $E(X)$.

Example 5-11: Two books are assigned for a statistics class: a textbook and its corresponding study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, 25% buy both books, and these percentages are relatively constant from one term to another. Use the formula for the mean to find the average textbook cost.

Solution: Let X = the revenue from statistics students for the bookstore, then $x_1 = \$0$, $x_2 = \$137$, and $x_3 = \$170$, which occur with probabilities 0.20, 0.55, and 0.25. The distribution of X is summarized in the table below.

x	\$0	\$137	\$170
$P(X = x)$	0.2	0.55	0.25

Compute the average outcome of X as

$$\mu = \sum(x_i \cdot P(x_i)) = x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + x_3 \cdot P(x_3) = 0 \cdot 0.2 + 137 \cdot 0.55 + 170 \cdot 0.25 = 117.85.$$

We call this average the expected value of X , denoted by $E(X) = \$117.85$.

Note using this method we are not dividing the answer by the total number of students since the probabilities are found by dividing the frequency by the total so we would not divide again.

It may have been tempting to set up the table with the x values 0, 1 and 2 books. This would be fine if the question was asking for the average number of books sold. Since the books are different prices, we would not be able to get an average cost. Make sure X represents the variable that you are using to calculate the mean.

Example 5-12: The following is a valid discrete probability distribution of $X =$ the net dollar winnings for an online video game. Find the mean net earnings for playing the game.

x	-5	-2.5	0	2.5	5
$P(X = x)$	0.55	$\frac{1}{4}$	0.15	0	5%

Solution: It is easier to have all the probabilities as proportions: $\frac{1}{4} = 0.25$ and $5\% = 0.05$. Be careful with the negative x values.

$$\begin{aligned} \mu &= \sum(x_i \cdot P(x_i)) = x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + x_3 \cdot P(x_3) = (-5) \cdot 0.55 + (-2.5) \cdot 0.25 + 0 \cdot 0.15 + 5 \cdot 0.05 \\ &= (-2.75) + (-0.625) + 0 + 0.25 = -3.125. \end{aligned}$$

If you were to play the game many times, in the long run, you can expect to lose, on average, \$3.125.

Example 5-13: The 2010 United States Census found the chance of a household being a certain number of people (household size). Compute the mean household size.

Size of Household	1	2	3	4	5	6	7
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

Solution: To find the mean it is easier to use a table as shown below. The formula for the mean says to multiply the x value by the $P(x)$ value, so add a row into the table for this calculation.

Also, convert all $P(x)$ to decimal form.

x	1	2	3	4	5	6	7
$P(x)$	0.267	0.336	0.158	0.137	0.063	0.024	0.015
$x \cdot P(x)$	0.267	0.672	0.474	0.548	0.315	0.144	0.105

Add up the new row and you get the answer 2.525. This is the mean or the expected value, $\mu = 2.525$ people.

This means that you expect a household in the United States to have 2.525 people in it.

Keep your answer as a decimal. Now of course you cannot have part of a person, but what this tells you is that you expect a household to have either 2 or 3 people, with a little more 3-person households than 2-person households.

Just as with any data set, you can calculate the mean and standard deviation. In problems involving a table of probabilities, the probability distribution represents a population. The probability distribution in most cases comes from repeating an experiment many times. This is because you are using the data from repeated experiments to estimate the true probability.

Since a **probability distribution** represents a population, the mean and standard deviation that are calculated are actually the population parameters and not the sample statistics. The notation used is for the population mean μ “mu” and population standard deviation σ “sigma.”

Note, the mean can be thought of as the **expected value**. It is the value you expect to get on average, if the trials were repeated an infinite number of times. The mean or expected value does not need to be a whole number, even if the possible values of x are whole numbers.

Example 5-14: In the Oregon lottery called Pick-4, a player pays \$1 and then picks a four-digit number. If those four numbers are picked in that specific order, the person wins \$2,000. What is the expected value in this game?

Solution: To find the expected value, you need to first create the probability distribution. In this case, the random variable $x = \$$ net winnings. If you pick the right numbers in the right order, then you win \$2,000, but you paid \$1 to play, so you actually win a net amount of \$1,999. If you did not pick the right numbers, you lose the \$1, the x net value is $-\$1$.

You also need the probability of winning and losing. Since you are picking a four-digit number, and for each digit, there are 10 possible numbers to pick from, with each independent of the others, you can use the fundamental counting rule. To win, you have to pick the right numbers in the right order. The first digit, you pick 1 number out of 10, the second digit you pick 1 number out of 10, and the third digit you pick 1 number out of 10.

The probability of picking the right number in the right order is $\frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{10000} = 0.0001$.

The probability of losing (not winning) would be the complement rule $1 - 0.0001 = 0.9999$.

Putting this information into a table will help to calculate the expected value.

Game Outcome	Win	Lose
x	\$1,999	$-\$1$
$P(X = x)$	0.0001	0.9999

Find the mean which is the same thing as the expected value.

Game Outcome	Win	Lose	
x	\$1,999	$-\$1$	
$P(x)$	0.0001	0.9999	Total
$x \cdot P(x)$	\$0.1999	$-\$0.9999$	$-\$0.80$

Now sum the last row and you have the expected value of $\$0.1999 + (-\$0.9999) = -\$0.80$. If you kept playing this game, in the long run, you will expect to lose \$0.80 per game. Since the expected value is not 0, this game is not fair.

Most lottery and casino games such as craps, 21, roulette, etc. and insurance policies are built with negative expectations for the consumer. This is how casinos and insurance companies stay in business.

5.2.2 Variance & Standard Deviation of Discrete Probability Distributions

Suppose you ran the university bookstore. Besides how much revenue you expect to generate, you might also want to know the volatility (variability) in your revenue.

The variance and standard deviation can be used to describe the variability of a random variable. When we first introduced a method for finding the variance and standard deviation for a data set, we had sample data and found sample statistics. We first computed deviations from the mean $(x_i - \mu)$, squared those deviations $(x_i - \mu)^2$, and took an average to get the variance. In the case of a random variable, we again compute squared deviations. However, we take their sum weighted by their corresponding probabilities, just as we did for the expectation. This weighted sum of squared deviations equals the variance, and we calculate the standard deviation by taking the square root of the variance, just as we did for a sample variance. We also are using notation for the population parameters σ and σ^2 , instead of the sample statistics s and s^2 .

Variance of a Discrete Random Variable X :

Suppose that X is a discrete random variable with values x_1, x_2, \dots, x_k .

Then the variance of X is, $\sigma^2 = \sum(x_i - \mu)^2 \cdot P(x_i)$ or an easier to compute, algebraically equivalent formula $\sigma^2 = \sum(x_i^2 \cdot P(x_i)) - \mu^2 = (x_1^2 \cdot P(x_1) + x_2^2 \cdot P(x_2) + \dots + x_k^2 \cdot P(x_k)) - \mu^2$.

Standard Deviation of a Discrete Random Variable X :

The standard deviation of X is the positive square root of the variance, $\sigma = \sqrt{\sigma^2}$.

Example 5-15: The following is a valid discrete probability distribution of X = the net dollar winnings for an online video game. Find the standard deviation of the net earnings for playing the game.

x	-5	-2.5	0	2.5	5
$P(x)$	0.55	$\frac{1}{4}$	0.15	0	5%

Solution: It is easier to have all the probabilities as proportions: $\frac{1}{4} = 0.25$ and $5\% = 0.05$. Be careful with the negative x values. First, find the mean $\mu = \sum(x_i \cdot P(x_i)) = x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + x_3 \cdot P(x_3) = (-5) \cdot 0.55 + (-2.5) \cdot 0.25 + 0 \cdot 0.15 + 5 \cdot 0.05 = (-2.75) + (-0.625) + 0 + 0.25 = -3.125$.

Next find the variance, $\sigma^2 = \sum(x_i^2 \cdot P(x_i)) - \mu^2 = (x_1^2 \cdot P(x_1) + x_2^2 \cdot P(x_2) + x_3^2 \cdot P(x_3) + x_4^2 \cdot P(x_4) + x_5^2 \cdot P(x_5)) - \mu^2$
 $= ((-5)^2 \cdot 0.55 + (-2.5)^2 \cdot 0.25 + 0^2 \cdot 0.15 + 2.5^2 \cdot 0 + 5^2 \cdot 0.05) - (-3.125)^2$
 $= (13.75 + 1.5625 + 0 + 0 + 1.25) - 9.765625 = 16.5625 - 9.765625 = 6.796875$.

Now take the square root of the variance to get to the standard deviation, $\sqrt{6.796875} = 2.607$, or $\sigma = \$2.607$.

Example 5-16: The 2010 United States Census found the chance of a household being a certain size. Compute the variance and standard deviation.

Size of Household	1	2	3	4	5	6	7
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

Solution: Make a table similar to how we started the mean by changing the probabilities to decimals, but this time square each x value before multiplying by its corresponding probability.

x	1	2	3	4	5	6	7
$P(x)$	0.267	0.336	0.158	0.137	0.063	0.024	0.015
$x^2 \cdot P(x)$	0.267	1.344	1.422	2.192	1.575	0.864	0.735

Add this new row up to get the beginning of the variance formula $\sum(x_i^2 \cdot P(x_i)) = 8.399$.

In a previous example we found the mean household size to be $\mu = 2.525$ people.

To finish finding the variance we need to square the mean and subtract from the sum to get $\sigma^2 = \sum(x_i^2 \cdot P(x_i)) - \mu^2 = 8.399 - 2.525^2 = 2.023375$ people².

Having a measurement in squared units is not very helpful when trying to interpret so we take the square root to find the standard deviation, which is back in the original units.

The standard deviation of the number of people in a household is $\sigma = \sqrt{2.023375} = 1.4225$ people. This means that you can expect an average United States household to have 2.525 people in it, with an average spread or standard deviation of 1.42 people.

TI-84: Press [STAT], choose 1:Edit. For x and $P(x)$ data pairs, enter all x -values in one list. Enter all corresponding $P(x)$ values in a second list. Press [STAT]. Use cursor keys to highlight CALC. Select **1:1-Var Stats**.



Enter list 1 for List and list 2 for frequency list. Press Enter to calculate the statistics. For TI-83, you will just see 1-Var Stats on the screen. Enter each list separated with a comma by pressing [2nd], then press the number 1 key corresponding to your x list, then a comma, then [2nd] and the number 2 key corresponding to your $P(x)$ values. The home screen should look like this 1-Var Stats L_1, L_2 .

Where the calculator says \bar{x} is μ the population mean and σ_x is the population standard deviation (square the σ_x number to get the population variance).

TI-89: Go to the [Apps] **Stat/List Editor**, and type the x values into List 1 and $P(x)$ values into List2. Select F4 for the **Calc** menu. Use cursor keys to highlight **1:1-Var Stats**.

Under List, press [2nd] Var-Link, then select list1. Under Freq, press [2nd] Var-Link, then select list2. Press enter twice and the statistics will appear in a new window.

Use the cursor keys to arrow up and down to see all of the values. Note: \bar{x} is μ the population mean and σ_x is the population standard deviation; square the σ_x value to get the variance.



There are many situations that we can model a distribution with a formula. This will make finding probabilities with large sample sizes much easier than making a table. There are many types of probability distributions. We will just be covering a few of them.

5.3 Binomial Distribution

The binomial distribution is a discrete probability distribution used to find the probability of success when there are two outcomes to each trial, and there are a set number of independent trials with the same probability of occurrence.

Bernoulli trial

The focus of the previous section was on discrete probability distribution tables. To find the table for those situations, we usually need to actually conduct the experiment and collect data. Then one can calculate the experimental probabilities.

If certain conditions are met, we can instead use theoretical probabilities. One of these theoretical probabilities can be used when we have a **Bernoulli trial**.



Properties of a **Bernoulli trial** or binomial trial:

1. Trials are independent, which means that what happens on one trial does not influence the outcomes of other trials.
2. There are only two outcomes, which are called a success and a failure.
3. The probability of a success does not change from trial to trial, where p = probability of success and q = probability of failure, the complement of p , $q = 1 - p$.

If you know you have a Bernoulli trial, then you can calculate probabilities using some theoretical probability formulas. This is important because Bernoulli trials come up often in real life.

Examples of Bernoulli experiments are:

- Toss a fair coin, and find the probability of getting a head.
- Question a student, and look for the probability that they are a business major.
- A patient has a virus.
- What is the probability of getting a multiple-choice question correct on a test if you have not studied?

Bernoulli trials are used in the binomial distribution.

To develop the process for calculating the probabilities in a binomial experiment, consider the following example.

Example 5-17: Suppose you are given a three-question multiple-choice test. Each question has four responses and only one is correct. Suppose you want to find the probability that you can just guess at the answers and get two questions correct. To help with the idea that you are going to guess, suppose the test is on string theory.

- a) Is this a binomial experiment?

Solution:

The random variable is x = number of correct answers.

1. There are three questions, and each question is a trial, so there are a fixed number of trials. In this case, $n = 3$.
2. Getting the first question right has no effect on getting the second or third question correct, thus the trials are independent.
3. Either you get the question right or you get it wrong, so there are only two outcomes. In this case, the success is getting the question right.
4. The probability of getting a question right is one out of four. This is the same for every trial since each question has 4 responses. In this case, $p = \frac{1}{4}$ and $q = 1 - \frac{1}{4} = \frac{3}{4}$.

Since all of the properties are met, this is a binomial experiment.

- b) What is the probability of getting two questions right?

Solution: To answer this question, start with the sample space.

$SS = \{RRR, RRW, RWR, WRR, WWR, WRW, RWW, WWW\}$, where RRW means you get the first question right, the second question right, and the third question wrong.

Now the event space for getting 2 right is $\{RRW, RWR, WRR\}$. What you did in chapter four was just to find three divided by eight to compute the probability outcome. However, this would not be correct in this case, because the probability of getting a question right is different from getting a question wrong. What else can you do?

Look at just $P(RRW)$ for the moment. Again, that means $P(RRW) = P(R \text{ on 1st, R on 2nd, and W on 3rd})$. Since the trials are independent, then $P(RRW) = P(R \text{ on 1st, R on 2nd, and W on 3rd}) = P(R \text{ on 1st}) \cdot P(R \text{ on 2nd}) \cdot P(W \text{ on 3rd})$. Just multiply $p \cdot p \cdot q = P(RRW) = \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} = (\frac{1}{4})^2 \cdot (\frac{3}{4})^1$.

Similarly, you can compute the $P(RWR)$ and $P(WRR)$. To find the probability of 2 correct answers, just add these three probabilities together. $P(2 \text{ correct answers}) = P(RRW) + P(RWR) + P(WRR) = (\frac{1}{4})^2 \cdot (\frac{3}{4})^1 + (\frac{1}{4})^2 \cdot (\frac{3}{4})^1 + (\frac{1}{4})^2 \cdot (\frac{3}{4})^1 = 3 \cdot (\frac{1}{4})^2 \cdot (\frac{3}{4})^1$.

- c) What is the probability of getting zero right, one right, two right and all three right?

Solution: You could go through the same argument that you did above and come up with the following:

r right	P(r right)
0 right	$1 \cdot (\frac{1}{4})^0 \cdot (\frac{3}{4})^3$
1 right	$3 \cdot (\frac{1}{4})^1 \cdot (\frac{3}{4})^2$
2 right	$3 \cdot (\frac{1}{4})^2 \cdot (\frac{3}{4})^1$
3 right	$1 \cdot (\frac{1}{4})^3 \cdot (\frac{3}{4})^0$

Do you see the resulting pattern? You can now write the general formula for the probabilities for a binomial experiment.

First, the random variable in a binomial experiment is $x = \text{number of successes}$.

A **binomial probability distribution** results from a random experiment that meets all of the following requirements.

1. The procedure has a fixed number of trials (or steps), which is denoted by n .
2. The trials must be independent.
3. Each trial must have exactly two categories that can be labeled “success” and “failure.”
4. The probability of a “success,” denoted by p , remains the same in all trials. The probability of “failure” is often denoted by q , thus $q = 1 - p$.
5. Random Variable, X , counts the number of “successes.”

If a random experiment satisfies all of the above, the distribution of the random variable X , where X counts the number of successes, is called a binomial distribution. A binomial distribution is described by the population proportion p and the sample size n . If a discrete random variable X has a binomial distribution with population proportion p and sample size n , we write $X \sim B(n, p)$.

The binomial distribution is $P(X = x) = {}_n C_x \cdot p^x \cdot q^{(n-x)}$, $x = 0, 1, 2, \dots, n$. Where n is the number of trials (sample size), x is the number of successes out of n that you are trying to find the probability for, p is the probability of a success for one trial and $q = 1 - p$.

Be careful, a “success” is not always a “good” thing. Sometimes a success is something that is “bad,” like finding a defect or getting in a car crash. The success will be the event that you are trying to find the probability for in the question.

Excel Formula for Binomial Distribution: For exactly $P(X = x)$ use `=binom.dist(x,n,p,false)`.
For $P(X \leq x)$ use `=binom.dist(x,n,p,true)`.

TI-84: Press [2nd] [DISTR]. This will get you a menu of probability distributions. Press 0 or arrow down to **0:binompdf**(and press [ENTER]. This puts binompdf(on the home screen. Enter the values for n , p and x with a comma between each. Press [ENTER]. This is the probability density function and will return you the probability of exactly x successes. If you leave off the x value and just enter n and p , you will get all the probabilities for each x from 0 to n . Press [ALPHA] A or arrow down to **A:binomcdf**(and press [ENTER]. This puts binomcdf(on the home screen. Enter the values for n , p and x with a comma between each. If you have the newer operating system on the TI-84, the screen will prompt you for each value. Press [ENTER]. This is the cumulative distribution function and will return you the probability of at most (\leq) x successes. If you have at least x success (\geq), use the complement rule. If you have $<$ or $>$ adjust x to get \leq or \geq .

TI-89: Go to the [Apps] **Stat/List Editor**, then select F5 [DISTR]. This will get you a menu of probability distributions. Arrow down to **binomial Pdf** and press [ENTER]. Enter the values for n , p and x into each cell. Press [ENTER]. This is the probability density function and will return you the probability of exactly x successes. If you leave off the x value and just enter n and p , you will get all the probabilities for each x from 0 to n . Arrow down to **binomial Cdf** and press [ENTER]. Enter the values for n , p and lower and upper value of x into each cell. Press [ENTER]. This is the cumulative distribution function and will return you the probability between the lower and upper x -values, inclusive.

Example 5-18: When looking at a person's eye color, it turns out that only 2% of people in the world have green eyes (not to be confused with hazel colored eyes). Consider a randomly selected group of 20 people.

a) Is this a binomial experiment?

Solution: Yes, since all the requirements are met:

1. There are 20 people, and each person is a trial, so there are a fixed number of trials.
2. If you assume that each person in the group is chosen at random the eye color of one person does not affect the eye color of the next person, thus the trials are independent.
3. Either a person has green eyes or they do not have green eyes, so there are only two outcomes. In this case, the success is a person has green eyes.
4. The probability of a person having green eyes is 0.02. This is the same for every trial since each person has the same chance of having green eyes.

b) Compute the probability that none have green eyes.

Solution: You are looking for $P(X = 0)$, since this problem is asking for none $x = 0$. There are 20 people so $n = 20$. The success is selecting someone with green eyes, so the probability of a success $p = 0.02$. Then the probability of not selecting someone with green eyes is $q = 1 - p = 1 - 0.02 = 0.98$.

Using the formula: $P(X = 0) = {}_{20}C_0 \cdot 0.02^0 \cdot 0.98^{(20-0)} = 0.6676$.

Thus, there is a 66.76% chance that in a group of 20 people none of them will have green eyes.

TI-83/84, use `binompdf(20,0.02,0) = 0.6676`.

In Excel use the formula `=binom.dist(0,20,0.02,false) = 0.6676`.



c) Compute the probability that nine have green eyes.

Solution: $P(X = 9) = \text{binom.dist}(9,20,0.02,\text{false}) = 6.8859\text{E-}11 = 0.00000000006859$ which is zero rounded to four decimal places. The probability that out of 20 people, nine of them have green eyes is a very small chance and would be considered a rare event.

As you read through a problem look for some of the following key phrases in Figure 5-5. Once you find the phrase then match up to what sign you would use and then use the table to walk you through the computer or calculator formula.

$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
Is the same as	Is less than or equal to	Is greater than or equal to
Is equal to	Is at most	Is at least
Is exactly the same as	Is not greater than	Is not less than
Has not changed from	Within	Is more than or equal to
Excel =binom.dist($x,n,p,false$)	Excel =binom.dist($x,n,p,true$)	Excel =1-binom.dist($x-1,n,p,true$)
TI Calculator binompdf(n,p,x)	TI Calculator binomcdf(n,p,x)	TI Calculator 1-binomcdf($n,p,x-1$)
	$P(X > x)$	$P(X < x)$
	More than	Less than
	Greater than	Below
	Above	Lower than
	Higher than	Shorter than
	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced
	Excel =1-binom.dist($x,n,p,true$)	Excel =binom.dist($x-1,n,p,true$)
	TI Calculator 1-binomcdf(n,p,x)	TI Calculator binomcdf($n,p,x-1$)

Figure 5-5

Example 5-19: As of 2018, the Centers for Disease Control and Prevention (CDC) reported that about 1 in 88 children in the United States have been diagnosed with autism spectrum disorder (ASD). A researcher randomly selects 10 children. Compute the probability that 2 children have been diagnosed with ASD.

Solution: The random variable x = number of children with ASD. There are 10 children, and each child is a trial, so there are a fixed number of trials. In this case, $n = 10$. If you assume that each child in the group is chosen at random, then whether a child has ASD does not affect the chance that the next child has ASD. Thus, the trials are independent. Either a child has been diagnosed with ASD or they have not been diagnosed ASD, so there are two outcomes. In this case, the success is a child has ASD. The probability of a child having ASD is $1/88$. This is the same for every trial since each child has the same chance of having ASD, $p = \frac{1}{88}$ and $q = 1 - \frac{1}{88} = \frac{87}{88}$.

Using the formula: $P(X = 2) = {}_{10}C_2 \cdot \left(\frac{1}{88}\right)^2 \cdot \left(\frac{87}{88}\right)^{(10-2)} = 0.0053$.

Using the TI-83/84 Calculator: $P(X = 2) = \text{binompdf}(10, 1/88, 2) = 0.0053$.

Using Excel: $=\text{BINOM.DIST}(2, 10, 1/88, \text{FALSE}) = 0.0053$.

Example 5-20: Flip a fair coin exactly 10 times.

- a) What is the probability of getting exactly 8 tails?

Solution: There are only two outcomes to each trial, heads or tails. The coin flips are independent and the probability of a success, flipping a tail, $p = 1/2 = 0.5$ is the same for each trial. This is a binomial experiment with a sample size $n = 10$. Using the formula: $P(X = 8) = {}_{10}C_8 \cdot 0.5^8 \cdot 0.5^{(10-8)} = 0.0439$

- b) What is the probability of getting 8 or more tails?

Solution: We still have $p = 0.5$ and $n = 10$, however we need to look at the probability of 8 or more. The $P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10)$. We can stop at $x = 10$ since the coin was only flipped 10 times. $P(X \geq 8) = {}_{10}C_8 \cdot 0.5^8 \cdot 0.5^{(10-8)} + {}_{10}C_9 \cdot 0.5^9 \cdot 0.5^{(10-9)} + {}_{10}C_{10} \cdot 0.5^{10} \cdot 0.5^{(10-10)} = 0.0439 + 0.0098 + 0.0010 = 0.0547$.

So far, most of the examples for the binomial distribution were for exactly x successes. If we want to find the probability of accumulation of x values then we would use the cumulative distribution function (cdf) instead of the pdf. Phrases such as “at least,” “more than,” or “below” can drastically change the probability answers.

Example 5-21: Approximately 10.3% of American high school students drop out of school before graduation. Choose 10 students entering high school at random. Find the following probabilities.

- a) No more than two drop out.

Solution: There is a set sample size of independent trials. The person either has or has not dropped out of school before graduation. A “success” is what we are finding the probability for, so in this case a success is to drop out so $p = 0.103$ and $q = 1 - 0.103 = 0.897$.

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = {}_{10}C_0 \cdot 0.103^0 \cdot 0.897^{10} + {}_{10}C_1 \cdot 0.103^1 \cdot 0.897^9 + {}_{10}C_2 \cdot 0.103^2 \cdot 0.897^8 = 0.3372 + 0.3872 + 0.2001 = 0.9245.$$

Calculator shortcut use $\text{binompdf}(10,0.103,0) + \text{binompdf}(10,0.103,1) + \text{binompdf}(10,0.103,2) = 0.9245$ or the $\text{binomcdf}(10,0.103,2) = 0.9245$. On the TI-89 cdf enter the lower value of x as 0 and upper value of x as 2.

In Excel use the function $=\text{BINOM.DIST}(2,10,0.103,\text{TRUE})$. Note that if you choose False under cumulative this would return just the $P(X = 2)$ not $P(X \leq 2)$.

- b) At least 6 students graduate.

Solution: A success is to graduate so $p = 0.897$ and $q = 0.103$.

$$P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) = {}_{10}C_6 \cdot 0.897^6 \cdot 0.103^4 + {}_{10}C_7 \cdot 0.897^7 \cdot 0.103^3 + {}_{10}C_8 \cdot 0.897^8 \cdot 0.103^2 + {}_{10}C_9 \cdot 0.897^9 \cdot 0.103^1 + {}_{10}C_{10} \cdot 0.897^{10} \cdot 0.103^0 = 0.0123 + 0.0613 + 0.2001 + 0.3872 + 0.3372 = 0.9981.$$

This is a lot of work to do each one so we can use technology to find the answer.

Note that Excel and the older TI-84 programs only find the probability below x so you have to use the complement rule.

For the TI-84 calculator shortcut use $\text{binompdf}(10,0.897,6) + \text{binompdf}(10,0.897,7) + \text{binompdf}(10,0.897,8) + \text{binompdf}(10,0.897,9) + \text{binompdf}(10,0.897,10) = 0.9981$ or use the complement rule $P(X \geq 6) = 1 - P(X \leq 5) = 1 - \text{binomcdf}(10,0.897,5) = 0.9981$.

On the TI-89, just use the binomcdf with the lower x value as 6 and the upper x value as 10.

In Excel use $=1 - \text{BINOM.DIST}(5,10,0.897,\text{TRUE})$.

- c) Exactly 10 students stay in school and graduate.

Solution: A success is to graduate so $p = 0.897$ and $q = 0.103$. Find $P(X = 10) = {}_{10}C_{10} \cdot 0.897^{10} \cdot 0.103^0 = 0.3372$.

On the TI-84 use $\text{binompdf}(10,0.897,10) = 0.3372$.

In Excel $=\text{BINOM.DIST}(10,10,0.897,\text{FALSE}) = 0.3372$.

Example 5-22: When looking at a person's eye color, it turns out that only 2% of people in the world have green eyes (not to be confused with hazel colored eyes). Consider a randomly selected group of 20 people.

- a) Compute the probability that at most 3 have green eyes.

Solution: This fits a binomial experiment. $P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = {}_{20}C_0 \cdot 0.02^0 \cdot 0.80^{20} + {}_{20}C_1 \cdot 0.02^1 \cdot 0.80^{19} + {}_{20}C_2 \cdot 0.02^2 \cdot 0.80^{18} + {}_{20}C_3 \cdot 0.02^3 \cdot 0.80^{17} = 0.667608 + 0.272493 + 0.05283 + 0.006469 = 0.9994$.

On the TI-84 use $\text{binomcdf}(20,0.02,3) = 0.9994$. In Excel $=\text{BINOM.DIST}(3,20,0.02,\text{true}) = 0.9994$.

- b) Compute the probability that less than 3 have green eyes.

Solution: $P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = 0.667608 + 0.272493 + 0.05283 = 0.9929$.

On the TI-84 use $\text{binomcdf}(20,0.02,2) = 0.9929$. In Excel $=\text{BINOM.DIST}(2,20,0.02,\text{true}) = 0.9929$.

- c) Compute the probability that more than 3 have green eyes.

Solution: $P(X > 3) = 1 - P(X \leq 3) = 1 - 0.9994 = 0.0006$.

On the TI-84 use $1 - \text{binomcdf}(20,0.02,3) = 0.0006$. In Excel $=1 - \text{BINOM.DIST}(3,20,0.02,\text{true}) = 0.0006$.

- d) Compute the probability that 3 or more have green eyes.

Solution: $P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.9929 = 0.0071$

On the TI-84 use $1 - \text{binomcdf}(20,0.02,2) = 0.0071$. In Excel $=1 - \text{BINOM.DIST}(2,20,0.02,\text{true}) = 0.0071$.

Example 5-23: As of 2018, the Centers for Disease Control and Prevention (CDC) reported that about 1 in 88 children in the United States have been diagnosed with autism spectrum disorder (ASD). A researcher randomly selects 10 children. Compute the probability that at least 2 children have been diagnosed with ASD.

Solution: $P(X \geq 2) = 1 - P(X \leq 1) = 1 - {}_{10}C_0 \cdot \left(\frac{1}{88}\right)^0 \cdot \left(\frac{87}{88}\right)^{(10-0)} + {}_{10}C_1 \cdot \left(\frac{1}{88}\right)^1 \cdot \left(\frac{87}{88}\right)^{(10-1)} = 1 - (0.89202 + 0.102528) = 0.0055$.

On the TI-84 use $1 - \text{binomcdf}(10,1/88,1) = 0.0055$. In Excel $=1 - \text{BINOM.DIST}(1,10,1/88,\text{true}) = 0.0055$.

Mean, Variance & Standard Deviation of a Binomial Distribution

If you list all possible values of x in a Binomial distribution, you get the Binomial Probability Distribution (pdf). You can then find the mean, the variance, and the standard deviation using the general formulas $\mu = \sum(x_i \cdot P(x_i))$ and $\sigma^2 = \sum(x_i^2 \cdot P(x_i)) - \mu^2$. This, however, would take a lot of work if you had a large value for n . If you know the type of distribution, like binomial, then you can find the mean, variance and standard deviation using easier formulas. They are derived from the general formulas. For a Binomial distribution, μ , the expected number of successes, σ^2 , the variance, and σ , the standard deviation for the number of successes given by the following formulas, where p is the probability of success and $q = 1 - p$.

$$\begin{aligned}\text{Mean} &= \mu = n \cdot p \\ \text{Variance} &= \sigma^2 = n \cdot p \cdot q \\ \text{Standard Deviation} &= \sigma = \sqrt{n \cdot p \cdot q}\end{aligned}$$

Example 5-24: A random experiment consists of flipping a coin three times. Let X = the number of heads that show up. Compute the mean and standard deviation of X , that is, the mean and standard deviation for the number of heads that show up when a coin is flipped three times.

Solution: This experiment follows a binomial distribution; hence, we can use the mean and standard deviation formulas for a binomial. The mean of number of heads is $\mu = 3 \cdot 0.5 = 1.5$. The standard deviation of X is $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{(3 \cdot 0.5 \cdot (1 - 0.5))} = 0.8660$.

Example 5-25: When looking at a person's eye color, it turns out that only 2% of people in the world have green eyes (not to be confused with hazel colored eyes). Consider a randomly selected group of 20 people. Compute the mean, variance and standard deviation.

Solution: Since this is a binomial experiment, then you can use the formula $\mu = n \cdot p$. So $\mu = 20 \cdot 0.02 = 0.4$ people. You would expect on average that out of 20 people, less than 1 would have green eyes. The variance would be $\sigma^2 = n \cdot p \cdot q = 20(0.02)(0.98) = 0.392$ people². Once you have the variance, you just take the square root of the variance to find the standard deviation $\sigma = \sqrt{0.392} = 0.6261$ people. We would expect on average spread of the distribution to have 0.4 ± 0.6261 or 0 to 1 person out of 20 people to have green eyes.

Example 5-26: What is the expected grade on a 20-question multiple choice test with four possible choices for each question, if a person randomly guesses on each question?

Solution: A success would be getting a question correct. This is a binomial experiment with $n = 20$ and $p = 0.25$. The expected value is the mean, so $\mu = 20 \cdot 0.25 = 5$. A grade of 5 out of 20 = 25%, would be a failing grade.

5.4 Empirical Rule

Continuous Probability Distributions

A **continuous random variable** (usually denoted as X) is a variable that has an infinite number of random values in an interval of numbers. There are many different types of continuous distributions. To be a valid continuous distribution the total area under the curve has to be equal to one and the function's y-values need to be positive.

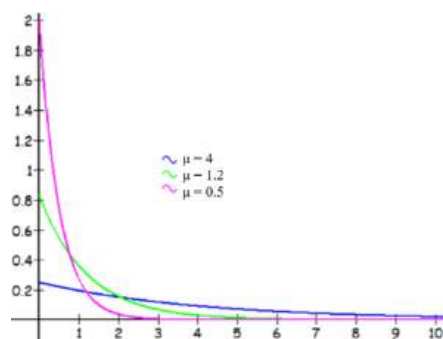


Figure 5-7

For example, we may have a random variable that is uniformly distributed so we could use the Uniform distribution that looks like a rectangle. See Figure 5-6.

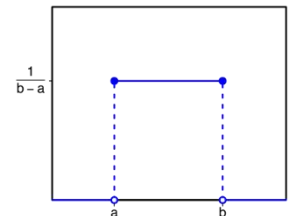


Figure 5-6

We may want to model the time it takes customer service to complete a call with the exponential distribution. See Figure 5-7.

We may have standardized test scores that follow a bell-shaped curve like the Gaussian (Normal) Distribution. See Figure 5-8.

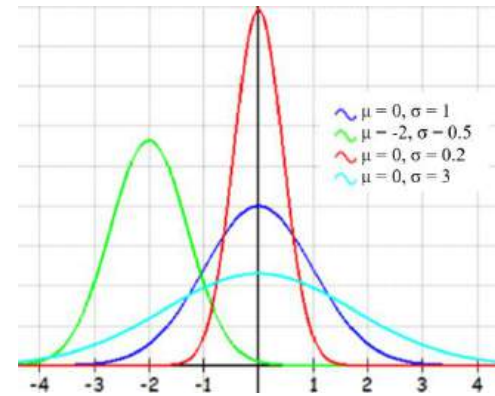


Figure 5-8

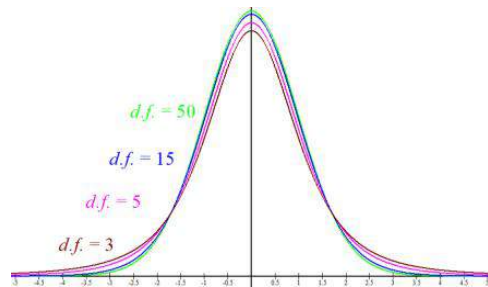


Figure 5-9

We may want to model the average time it takes for a component to be manufactured and use the bell-shaped Student's t-distribution. See Figure 5-9.

This is just an introductory course so we are only going to cover a few continuous distributions. If you want to explore more distributions, check out the chart by Larry Leemis at: <http://www.math.wm.edu/~leemis/chart/UDR/UDR.html>.

VERY IMPORTANT: The probability of an interval between two X values is equal to the area under the density curve between those two X values. For a discrete random variable, we can assign probabilities to each outcome. We cannot do this for a continuous random variable. The probability for a single X value for a continuous random variable is 0. Thus " $<$ " and " $>$ " are equivalent to " \leq " and " \geq ." In other words, $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$ since there is no area of a line.

Before looking at the process for computing probabilities for continuous distributions, it is somewhat useful to look at the **Empirical Rule** which gives the approximate proportion of data points under a bell-shaped curve between two points. The Empirical Rule, shown in Figure 5-10, is an approximation for probability under any bell-shaped distribution and will only be used in this section to give you an idea of the size of the probability for different shaded areas. A more precise method for finding probabilities will be demonstrated using technology for specific bell-shaped distributions.



The Empirical Rule should only be used with bell-shaped data.

The Empirical Rule (also called the 68-95-99.7 Rule)
 In a bell-shaped distribution with mean μ and standard deviation σ ,

- Approximately 68% of the observations fall within 1 standard deviation (σ) of the mean μ .
- Approximately 95% of the observations fall within 2 standard deviations (2σ) of the mean μ .
- Approximately 99.7% of the observations fall within 3 standard deviations (3σ) of the mean μ .

Note that we are using notation for the population mean μ and the population standard deviation σ , but the rule would also work using the sample mean and sample standard deviation.

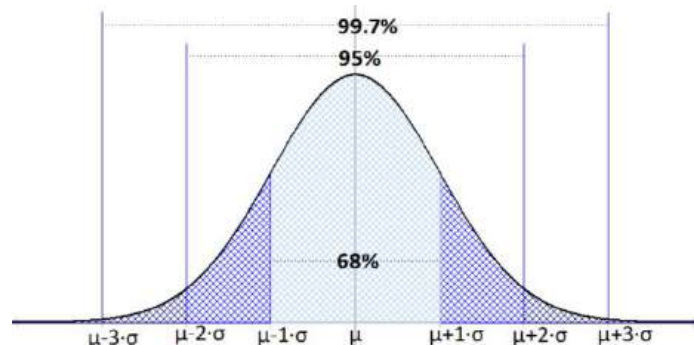


Figure 5-10

Example 5-27: In 2020, the average Scholastic Assessment Test (SAT) mathematics score in Colorado was 501, with a standard deviation of 116. Assume that SAT scores are bell-shaped.

- a) Approximately what proportion of students scored between 269 and 733 on the 2020 SAT mathematics test?

Solution: The key word is bell shaped so we can use the Empirical Rule. Start by finding the z -scores $z = \frac{x-\mu}{\sigma}$ for both endpoints given in the question.

A z -score by definition is the number of standard deviations a data value is from the mean.

$$z = \frac{269-501}{116} = -2$$

$$z = \frac{733-501}{116} = 2$$

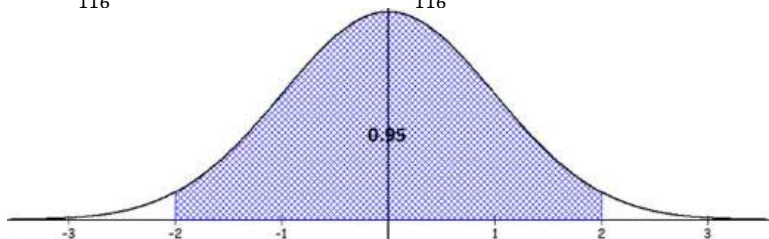


Figure 5-11

Draw and label the bell curve with the two z -scores and the area as shown in Figure 5-11. The two z -scores show that the test scores of 269 and 733 are two standard deviations from the mean. Using the second bulleted item in the Empirical Rule the answer would be approximately 95% of the math SAT scores will fall between 269 and 733.

- b) Approximately what proportion of students scored between 385 and 617 on the 2020 SAT mathematics test?

Solution: Take the z -scores of the endpoints to get: $z = \frac{385-501}{116} = -1$, $z = \frac{617-501}{116} = 1$.

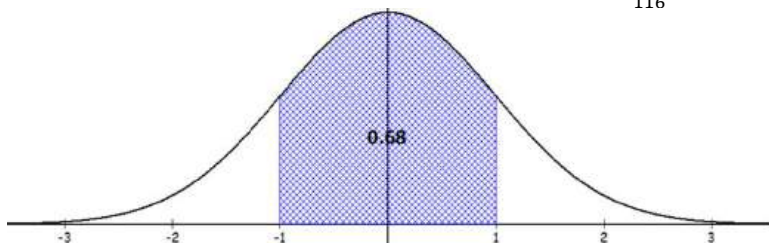


Figure 5-12

Draw and label the bell curve with the two z -scores and the area as shown in Figure 5-12. The two z -scores show that the test scores of 385 and 617 are one standard deviation from the mean. Using the first bulleted item in the Empirical Rule the answer would be approximately 68% of the math SAT scores will fall between 385 and 617.

- c) Approximately what proportion of students scored at least 617 on the 2020 SAT mathematics test?

Solution: Start by taking the z -score of 617, we get $z = \frac{617-501}{116} = 1$.

Since a bell-shaped curve is symmetric and we can assume that 100% of the population is represented then we can subtract the middle from the whole to get $100\% - 68\% = 32\%$. If we divide this outside area by two, $\frac{32\%}{2} = 16\%$, we would expect 16% in each tail area. See Figure 5-13.

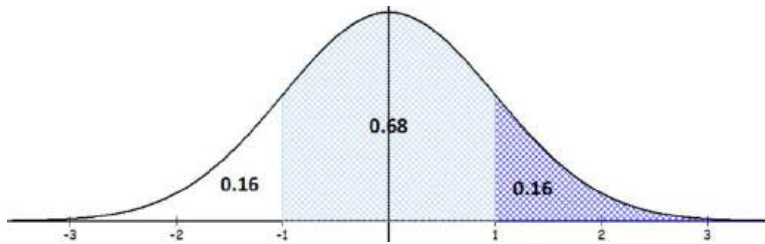


Figure 5-13

The answer would be approximately 16% of students scored at least 617 on the 2020 SAT mathematics test.

If you were to get a z-score that is not -3 , -2 , -1 , 1 , 2 or 3 then you would not be able to apply the Empirical Rule. We also need to ensure that our population has a bell-shaped curve before using the Empirical Rule.

5.5 Normal Distribution

For now, we will be working with the most common bell-shaped probability distribution known as the normal distribution, also called the Gaussian distribution, named after the mathematician Johann Carl Friedrich Gauss.



A **normal distribution** is a special type of distribution for a continuous random variable. Normal distributions are important in statistics because many situations in the real world have normal distributions.

Properties of the normal density curve:

1. Symmetric bell-shaped.
2. Unimodal (one mode).
3. Centered at the mean μ = median = mode.
4. The total area under the curve is equal to 1 or 100%.
5. The spread of a normal distribution is determined by the standard deviation σ . The larger σ is, the more spread out the normal curve is from the mean.
6. Follows the Empirical Rule.

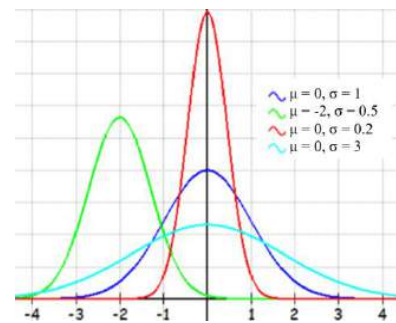


Figure 5-14

If a continuous random variable X has a normal distribution with mean μ and standard deviation σ then the distribution is denoted as $X \sim N(\mu, \sigma)$. Any x values from a normal distribution can be transformed or standardized into a standard Normal distribution by taking the z -score of x . Figure 5-14 shows several example normal distributions.

The formula for the normal probability density function (PDF) is: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. However, we will not be using this formula and instead be using technology to find the area under this curve. The probability is found by using integral calculus to find the area under the PDF curve. Prior to the handheld calculators and personal computers, there were probability tables made to look up these areas. This text does not use probability tables and will instead rely on technology to compute the area under the curve.

Every time the mean or standard deviation changes the shape of the normal distribution changes. The center of the normal curve will be the mean and the spread of the normal curve gets wider as the standard deviation gets larger.

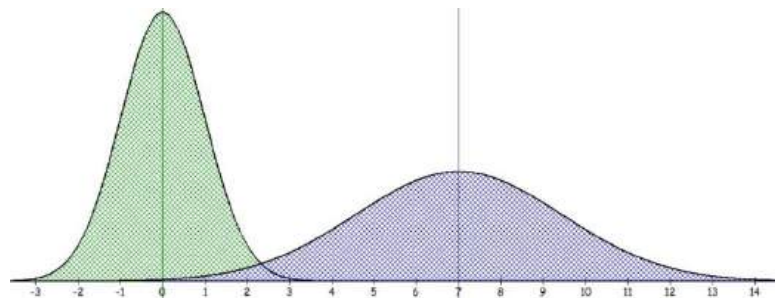


Figure 5-15

Figure 5-15 compares two normal distributions $N(0, 1)$ in green on the left and $N(7, 6)$ in blue on the right. When the standard deviation gets larger, the distribution is wider. The mean is the center of the distribution and shifts the curve left or right on the number line.

“So, what's odd about it?”
 “Nothing, it's Perfectly Normal.”
 (Adams, 2002)

VERY IMPORTANT: The probability of an interval between two X values is equal to the area under the density curve between those two X values. For a discrete random variable, we can assign probabilities to each outcome. We cannot do this for a continuous random variable. The probability for a single X value for a continuous random variable is 0. Thus “ $<$ ” and “ $>$ ” are equivalent to “ \leq ” and “ \geq .” In other words, $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$ since there is no area of a line.

5.5.1 Standard Normal Distribution

A normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal** distribution.

The letter Z is used exclusively to denote a variable that has a standard normal distribution and is written $Z \sim N(0, 1)$. A particular value of Z is denoted z (lower-case) and is referred to as a z -score. Recall that a z -score is the number of standard deviations x is from the mean. Anytime you are asked to find a probability of Z use the standard normal distribution.

Standardizing and Z-scores

A z -score is the number of standard deviations an observation x is above or below the mean μ . If the z -score is negative, x is below the mean. If the z -score is positive, x is above the mean.

If x is an observation from a distribution that has mean μ and standard deviation σ , the standardized value of x (or z -score) is $z = \frac{x - \mu}{\sigma}$.

To find the area under the probability density curve involves calculus so we will need to rely on technology to find the area. The area under the normal curve represents the probability.

Using Excel or TI-Calculator to Find Standard Normal Distribution

Use Figure 5-16 to help determine which signs match with the shaded bell curve.

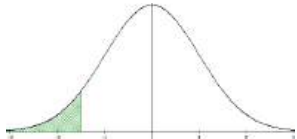
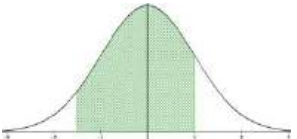
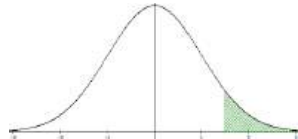
Standard Normal Distribution Finding a Probability		
$P(Z \leq z)$ or $P(Z < z)$	$P(z_1 < Z < z_2)$ or $P(z_1 \leq Z \leq z_2)$	$P(Z \geq z)$ or $P(Z > z)$
		
<code>=NORM.S.DIST(z,true)</code>	<code>=NORM.S.DIST(z2,true)- NORM.S.DIST(z1,true)</code>	<code>=1-NORM.S.DIST(z,true)</code>
<code>normalcdf(-1E99,z,0,1)</code>	<code>normalcdf(z1,z2,0,1)</code>	<code>normalcdf(z,1E99,0,1)</code>

Figure 5-16

Example 5-28: Compute the area under the standard normal distribution to the left of $z = 1.39$.

Solution: First, draw a bell-shaped distribution with 0 in the middle as shown in Figure 5-17. Mark 1.39 on the number line and shade to the left of $z = 1.39$.

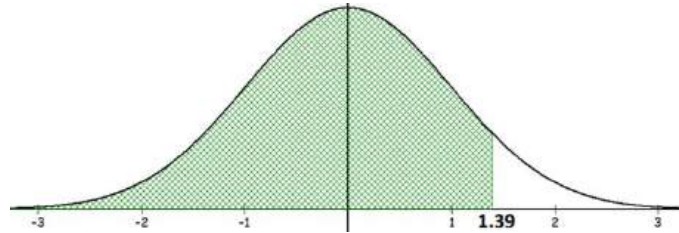
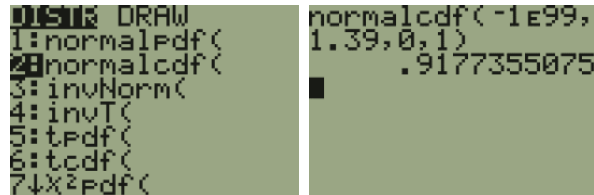


Figure 5-17

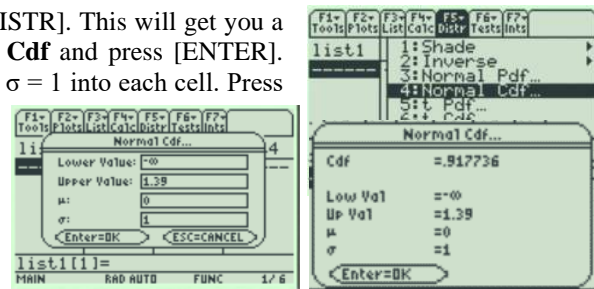
The normalcdf on the calculator needs the lower and upper value of the shaded area followed by the mean and standard deviation. The area to the left of $z = 1.39$ is the same as finding $P(Z < 1.39)$.

Note that the lower value of the shaded region is $-\infty$. The TI-84 does not have an infinity $-\infty$ symbol. Instead, we use a really small number in scientific notation $-1E99$ or -1×10^{99} (make sure you use the negative sign (-) not the minus - sign). Press [2nd] [EE] to get the scientific notation E (this will have a slightly smaller font size) and not the letter E. (The TI-89 uses $-\infty$ for the lower boundary instead of $-1E99$.) See technology directions below to calculate the area.

TI-84: Press [2nd] [DISTR] menu, select the normalcdf. Then type in the lower value, upper value, mean = 0, standard deviation = 1 to get $\text{normalcdf}(-1E99, 1.39, 0, 1) = 0.9177$, which is your answer. The area under the curve is equivalent to the probability of getting a z-score less than 1.39, or $P(Z < 1.39) = 0.9177$.



TI-89: Go to the [Apps] **Stat/List Editor**, then select F5 [DISTR]. This will get you a menu of probability distributions. Arrow down to **Normal Cdf** and press [ENTER]. Enter the values for the lower = $-\infty$, upper = 1.39, $\mu = 0$, and $\sigma = 1$ into each cell. Press [ENTER]. This is the cumulative distribution function and will return $P(Z < 1.39) = 0.9177$.



Excel: For Excel the program will only find the area to the left of a point. Therefore, if we want to find the area to the right of a point or between two points there will be one extra step. Use the formula $=\text{NORM.S.DIST}(1.39, \text{TRUE}) = 0.9177$.

Example 5-29: Compute the $P(-1.37 < Z < 1.68)$.

Solution: $P(-1.37 < Z < 1.68) = P(-1.37 \leq Z \leq 1.68)$ which is the same as finding the area under the curve between -1.37 and 1.68 . First, draw a bell-shaped distribution and identify the two points on the number line. Shade the area between the two points as shown in Figure 5-18.

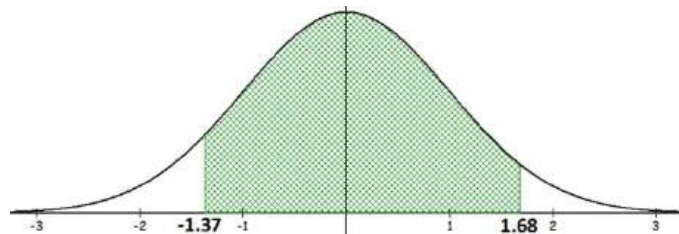


Figure 5-18

TI Calculator: $P(-1.37 \leq Z \leq 1.68)$
 $= \text{normalcdf}(-1.37, 1.68, 0, 1) = 0.8682$.

Excel: $P(-1.37 \leq Z \leq 1.68)$
 $= \text{NORM.S.DIST}(1.68, \text{TRUE}) - \text{NORM.S.DIST}(-1.37, \text{TRUE}) = 0.8682$.

Example 5-30: Compute the area to the right of $z = 1$.

Solution: Draw and shade the curve to find $P(Z > 1)$. See Figure 5-19.

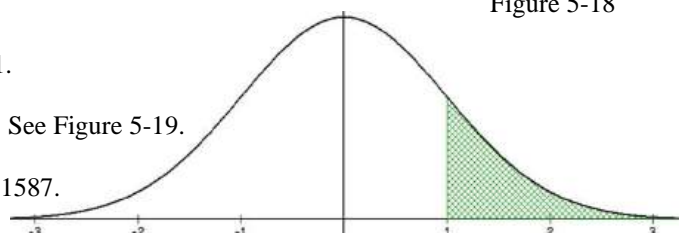


Figure 5-19

TI Calculator: $P(Z > 1) = \text{normalcdf}(1, 1E99, 0, 1) = 0.1587$.

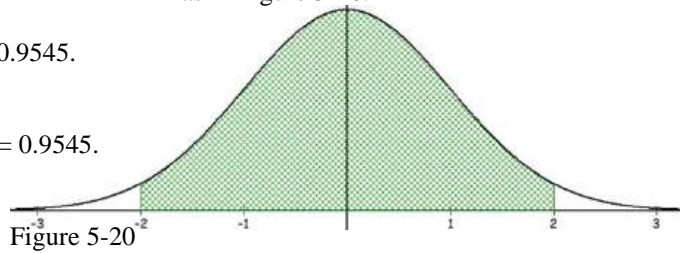
Excel: $P(Z > 1) = 1 - \text{NORM.S.DIST}(1, \text{TRUE}) = 0.1587$.

Example 5-31: Compute the area under the standard normal distribution that is 2 standard deviations from the mean.

Solution: A rough estimate using the Empirical Rule would be 0.95, however since this is not just any bell-shaped distribution, we will find the $P(-2 \leq Z \leq 2)$. Draw and shade the curve as in Figure 5-20.

TI Calculator: $P(-2 \leq Z \leq 2) = \text{normalcdf}(-2,2,0,1) = 0.9545$.

Excel: $P(-2 \leq Z \leq 2) = \text{NORM.S.DIST}(2, \text{TRUE}) - \text{NORM.S.DIST}(-2, \text{TRUE}) = 0.9545$.



Finding Percentiles for a Standard Normal Distribution

Sometimes you will be given an area or probability and have to find the associated with a z -score. For example, the probability below a point on the standard normal distribution is a percentile. We can use technology to find z -score given a percentile. Most technology has built in commands that will find the probability below a point. If you want to find the area above a point, or between two points, then find the area below a point by using the complement rule and keep in mind that the total area under the curve is 1 and the total area below the mean is 0.5.

If x is an observation from a distribution that has mean μ and standard deviation σ , the standardized value of x (or z -score) is $z = \frac{x - \mu}{\sigma}$.

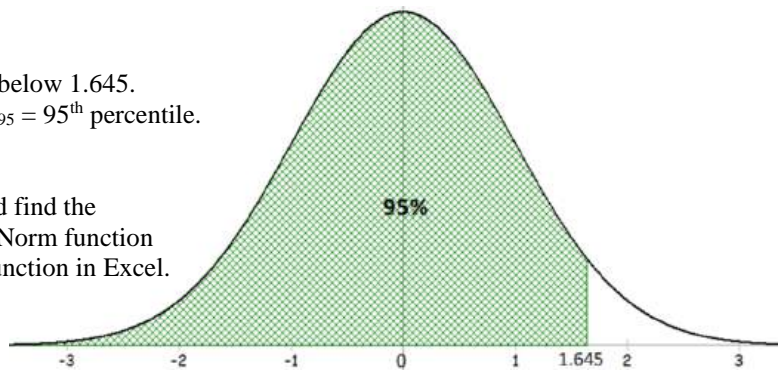
If you have a z -score and want to convert back to x , you can do so by solving the above equation for x , which yields $x = z\sigma + \mu$.

If we find that the $P(Z < 1.645)$
 $= \text{NORM.S.DIST}(1.645, \text{TRUE}) = 0.95$.

This tells us that about 95% of z -scores are below 1.645.
 In other words, the z -score of 1.645 is the $P_{95} = 95^{\text{th}}$ percentile.
 See Figure 5-21.

We can use technology to go backwards and find the z -score, given a percentile, by using the invNorm function in the TI calculator, or the NORM.S.INV function in Excel.

See technology directions below to calculate the z -score.



TI-84: Press [2nd] [DISTR]. This will get you a menu of probability distributions. Press 3 or arrow down to **3:invNorm**(and press [ENTER]. This puts $\text{invNorm}(\$ on the home screen. Enter the area to the left of the x value, μ , and σ with a comma between each. Press [ENTER]. This will return the percentile for the x value. For example, to find the 95^{th} percentile when the mean is 0 and the standard deviation is 1, you would have $\text{invNorm}(0.95,0,1)$. If you leave out the μ and σ , then the default is the **z -score** for the standard normal distribution. You get the z -score = 1.645.

TI-89: Go to the [Apps] **Stat/List Editor**, then select F5 [DISTR]. This will get you a menu of probability distributions. Arrow down to **Inverse Normal** and press [ENTER]. Enter the area to the left of the z -score = 0.95, $\mu = 0$, and $\sigma = 1$ into each cell. Press [ENTER]. This will return the percentile for the z -score. You get the z -score = 1.645.

Excel: $z = \text{NORM.S.INV}(0.95) = 1.645$



Using Excel or TI-Calculator for the Percentile of a Standard Normal Distribution

Use Figure 5-22 to help identify which function to use to find a z-score.

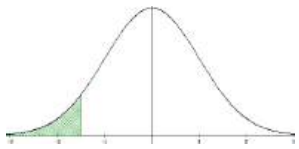
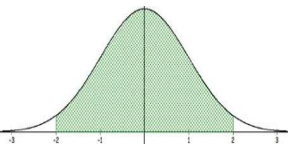
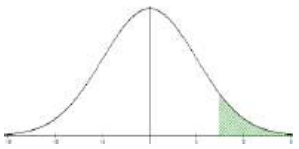
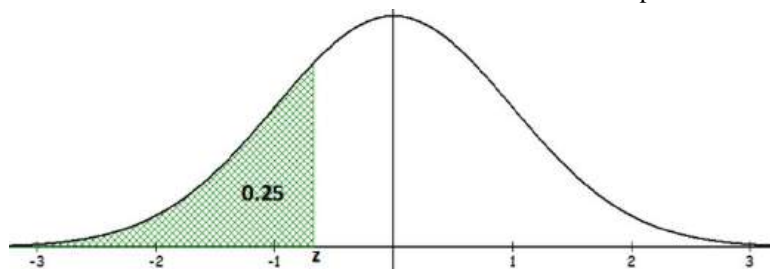
Standard Normal Distribution Finding an z-score Given an Area or Probability		
$P(Z \leq z)$ or $P(Z < z)$	$P(z_1 < Z < z_2)$ or $P(z_1 \leq Z \leq z_2)$	$P(Z \geq z)$ or $P(Z > z)$
		
Excel: $z = \text{NORM.S.INV}(\text{area})$	$z_1 = \text{NORM.S.INV}((1-\text{area})/2)$ $z_2 = \text{NORM.S.INV}(1-(1-\text{area})/2)$	$z = \text{NORM.S.INV}(1-\text{area})$
TI: $z = \text{invNorm}(\text{area}, 0, 1)$	$z_1 = \text{invNorm}((1-\text{area})/2, 0, 1)$ $z_2 = \text{invNorm}(1-(1-\text{area})/2, 0, 1)$	$z = \text{invNorm}(1-\text{area}, 0, 1)$

Figure 5-22

Example 5-32: Compute the z-score that corresponds to the 25th percentile.

Solution: First, draw the standard normal curve with zero in the middle as in Figure 5-23. The 25th percentile would have to be below the mean since the mean = median = 50th percentile for a bell-shaped distribution.



```

DISTR DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tcdf(
6:tcdf(
7:χ²pdf(

invNorm(.25,0,1)
-.6744897495
    
```

Figure 5-23

It is okay if you do not have this drawing to scale, but drawing a picture similar to Figure 5-23 will give you a good idea if your answer is correct. For instance, by just looking at this graph we can see that the answer for the z-score will need to be a negative number.

TI Calculator: $z = \text{invNorm}(0.25, 0, 1) = -0.6745$.

Excel: $z = \text{NORM.S.INV}(0.25) = -0.6745$.

The $z = -0.6745$ represents the 25th percentile.

Example 5-33: Compute the z-score that corresponds to the area of 0.4066 between zero and z shown in Figure 5-24.

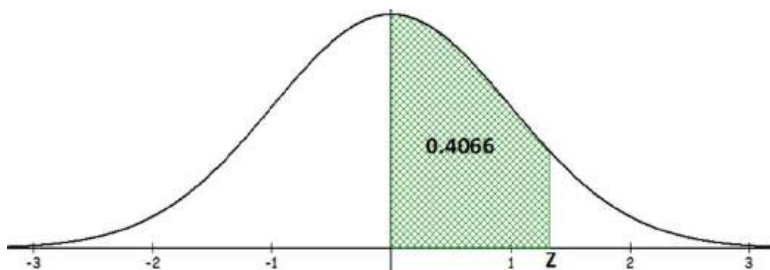


Figure 5-24

Solution: First notice that Figure 5-24 does not quite match the picture that the calculator or Excel use, which will only find for a left-tail area. The value of zero is the median on a standard normal distribution, so 50% of the area lies to the left of $z = 0$. This means that the total area to the left of the unknown z -score would be $0.5 + 0.4066 = 0.9066$.

TI Calculator: $z = \text{invNorm}(0.9066) = 1.32$.

Excel: $z = \text{NORM.S.INV}(0.9066) = 1.32$.

5.5.2 Applications of the Normal Distribution

Many variables are nearly normal, but rarely are exactly normal. Thus, the normal distribution, while not perfect for any single problem, is very useful for a variety of problems. Variables such as SAT scores and heights of United States adults closely follow the normal distribution. Note that the Excel function NORM.S.DIST is for a standard normal when $\mu = 0$ and $\sigma = 1$. When a question gives a mean and standard deviation other than zero and one, then use the NORM.DIST function. Use Figure 5-25 to help decide which functions to use based off of common phrases that may be used in the questions.

Using Excel or TI-Calculator to Find Normal Distribution Probabilities

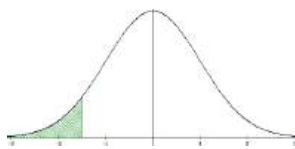
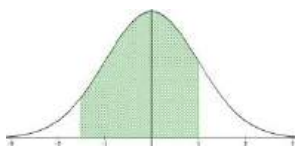
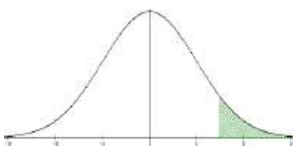
Normal Distribution Finding a Probability		
$P(X \leq x)$ or $P(X < x)$	$P(x_1 < X < x_2)$ or $P(x_1 \leq X \leq x_2)$	$P(X \geq x)$ or $P(X > x)$
Is less than or equal to	Between	Is greater than or equal to
Is at most		Is at least
Is not greater than		Is not less than
Within		More than
Less than		Greater than
Below		Above
Lower than		Higher than
Shorter than		Longer than
Smaller than		Bigger than
Decreased		Increased
Reduced		Larger
		
Excel: =NORM.DIST($x, \mu, \sigma, \text{true}$)	=NORM.DIST($x_2, \mu, \sigma, \text{true}$) – NORM.DIST($x_1, \mu, \sigma, \text{true}$)	=1 – NORM.DIST($x, \mu, \sigma, \text{true}$)
TI: =normalcdf(-1E99, x, μ, σ)	=normalcdf(x_1, x_2, μ, σ)	=normalcdf($x, 1E99, \mu, \sigma$)

Figure 5-25

TI-84: Press [2nd] [DISTR]. This will show a menu of probability distributions. Arrow down to **2:normalcdf**(and press [ENTER]. This puts normalcdf(on the home screen. Enter the values for the lower x value (x_1), upper x value (x_2), μ , and σ with a comma between each. Press [ENTER]. This is the cumulative distribution function and will return $P(x_1 < X < x_2)$. For example, to find $P(80 < X < 110)$ when the mean is 100 and the standard deviation is 20, you should have normalcdf(80,110,100,20). If you leave out the μ and σ , then the default is the standard normal distribution. For a left-tail area use a lower bound of $-1E99$ (negative infinity), (press [2nd] [EE] to get E) and for a right-tail area use an upper bound of $1E99$ (∞). For example, to find $P(Z < -1.37)$ you should have normalcdf(-1E99,-1.37,0,1).

TI-89: Go to the [Apps] **Stat/List Editor**, select F5 [DISTR]. This will show a menu of probability distributions. Arrow down to **Normal Cdf** and press [ENTER]. Enter the values for the lower x value (x_1), upper x value (x_2), μ , and σ into each cell. Press [ENTER]. This is the cumulative distribution function and will return $P(x_1 < X < x_2)$. For example, to find $P(80 < X < 110)$ when the mean is 100 and the standard deviation is 20, you should have in the following order 80, 110, 100, 20. If you have a z -score, use $\mu = 0$ and $\sigma = 1$, then you will get a standard normal distribution. For a left-tail area use a lower bound of negative infinity ($-\infty$), and for a right-tail area use an upper bound infinity (∞).

“The Hitchhiker's Guide to the Galaxy offers this definition of the word "Infinite." Infinite: Bigger than the biggest thing ever and then some. Much bigger than that in fact, really amazingly immense, a totally stunning size. "wow, that's big," time. Infinity is just so big that by comparison, bigness itself looks really titchy. Gigantic multiplied by colossal multiplied by staggeringly huge is the sort of concept we're trying to get across here.”
(Adams, 2002)

Example 5-34: Let X be the height of 15-year old boys in the United States. Studies show that the heights of 15-year old boys in the United States are normally distributed with average height 67 inches and a standard deviation of 2.5 inches. Compute the probability of randomly selecting one 15-year old boy who is 69.5 inches or taller.

Solution: Find $P(X \geq 69.5)$ where $X \sim N(67, 2.5)$. Draw the curve and label the mean, then shade the area to the right of 69.5 as shown in Figure 5-26.

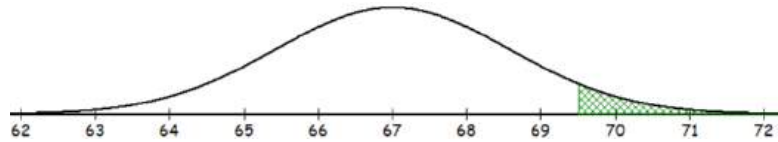


Figure 5-26

We could standardize the value of $x = 69.5$ using the z -score formula $z = \frac{x-\mu}{\sigma}$, where $\mu = 67$ and $\sigma = 2.5$. The standardized value of $x = 69.5$ is $z = \frac{69.5-67}{2.5} = 1$. See Figure 5-27.

Now using the standard normal distribution and shading the area to the right of $z = 1$ gives:

TI calculator: $P(Z \geq 1) = \text{normalcdf}(1, 1E99, 0, 1) = 0.158655$.

Excel: $P(Z \geq 1) = 1 - \text{NORM.S.DIST}(1, \text{TRUE}) = 0.158655$.

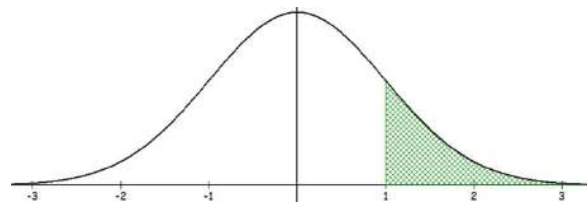


Figure 5-27

We could also use the Empirical rule to approximate $P(X \geq 69.5)$ because this is the same as being more than one standard deviation above the mean. Do you see why this makes sense? If we were to add a standard deviation to 67, we would get 69.5. Thus $P(X \geq 69.5) \approx 0.16$, which is close to our 15.87% using the standard normal distribution. The Empirical rule only gives an approximate value though.

The process of standardizing the X value was started so that we could use the standard normal distribution table to look up probabilities instead of using calculus. With technology, you no longer have to standardize first, we can just find $P(X \geq 69.5)$.

TI calculator: $P(X \geq 69.5) = \text{normalcdf}(69.5, 1E99, 67, 2.5) \approx 0.158655$ (The TI-89 use ∞ for the upper boundary instead of 1E99).

Excel: $P(X \geq 69.5) = 1 - \text{NORM.DIST}(69.5, 67, 2.5, \text{TRUE}) = 0.158655$.

Note that using the technology as in the last step will give you a more accurate answer than the Empirical Rule or finding a z -score first.

Example 5-35: In 2020, the average SAT mathematics score for Colorado was 501, with a standard deviation of 116.

a) If we randomly selected a student from that year who took the SAT, what is the probability of getting a SAT mathematics score between 400 and 600?

Solution: Draw the bell curve and label the mean $\mu = 501$ in the middle. Then label the points 400 and 600 and shade between the two x values. See Figure 5-28.

To calculate $P(400 \leq X \leq 600)$ using the TI calculator, use $\text{normalcdf}(400,600,501,116) = 0.6113$.

In Excel we can use the following formula $=\text{NORM.DIST}(600,501,116,\text{TRUE}) - \text{NORM.DIST}(400,501,116,\text{TRUE})$. Note that the right-hand endpoint goes in first. If you put the 400 in first you would get a negative answer, and probabilities are never negative. $P(400 \leq X \leq 600) = 0.6113$.

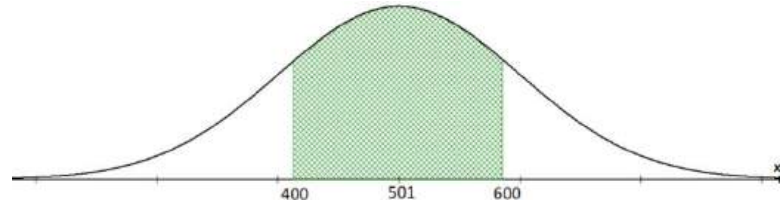


Figure 5-28

Note that for a continuous distribution there is no area on the line so the $P(400 \leq X \leq 600) = P(400 < X < 600)$.

b) What is the probability that a randomly selected student will have a score above 501?

Solution: Find $P(X > 501)$. Since $\mu = 501$ is in the middle of the distribution at the half way point the mean = median, so there is no need to use a calculator since $P(X > 501) = 50\% = 0.5$.

c) Compute the probability that a randomly selected student will have a score above 750.

Solution: Find $P(X > 750)$. Draw the bell curve and label the mean $\mu = 501$ in the middle. Then label the point 750 and shade to the right of the x value. See Figure 5-29.

```
normalcdf(750,1E99,501,116)
.0159144362
```

To calculate $P(X > 750)$ using the TI calculator, use $\text{normalcdf}(750,1E99,501,116) = 0.0159$.

In Excel we can use the following formula $=1 - \text{NORM.DIST}(750,501,116,\text{TRUE}) = 0.0159$.

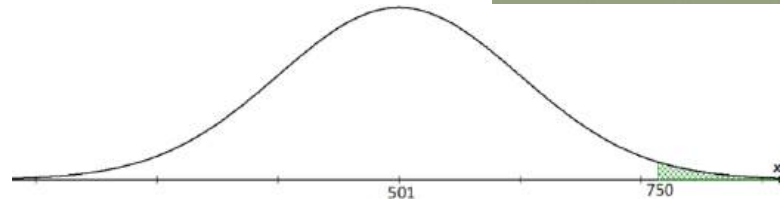


Figure 5-29

d) Compute the probability that a randomly selected student will have a score of at most 450.

Solution: Find $P(X \leq 450)$. Draw the bell curve and label the mean $\mu = 501$ in the middle. Then label the point 450 and shade to the left of the x value. See Figure 5-30.

To calculate $P(X \leq 450)$ using the TI calculator, use $\text{normalcdf}(-1E99,450,501,116) = 0.3301$.

In Excel we can use the following formula $=\text{NORM.DIST}(450,501,116,\text{TRUE}) = 0.3301$.

```
normalcdf(-1E99,450,501,116)
.3300934631
```

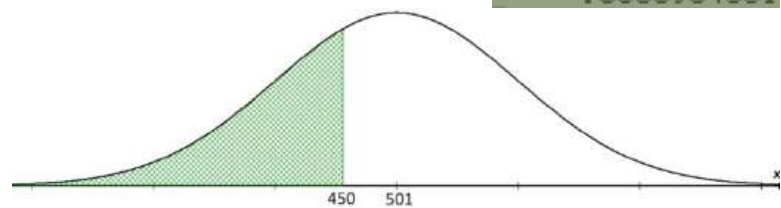


Figure 5-30

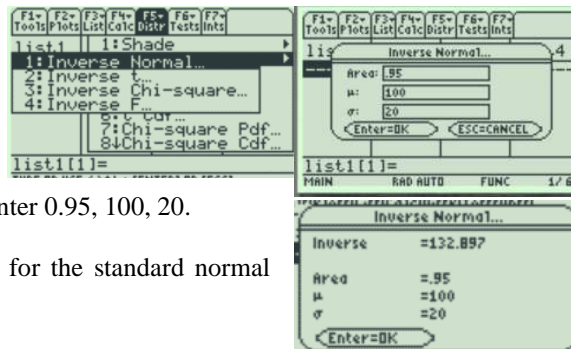
Finding Percentiles for a Normal Distribution

Sometimes you will be given an area or probability and have to find the associated random variable x . For example, the probability below a point on the normal distribution is a percentile. We can use technology to find x given a percentile. Most technology has built in commands that will find the probability below a point. If you want to find the area above a point, or between two points, then find the area below a point by using the complement rule and keep in mind that the total area under the curve is 1 and the total area below or above the mean is 0.5.

TI-84: Press [2nd] [DISTR]. This will get you a menu of probability distributions. Press 3 or arrow down to **3:invNorm**(and press [ENTER]. This puts $\text{invNorm}(\text{ on the home screen. Enter the area to the left of the } x \text{ value, } \mu, \text{ and } \sigma \text{ with a comma between each. Press [ENTER]. This will return the percentile for the } x \text{ value. For example, to$

find the 95th percentile when the mean is 100 and the standard deviation is 20, you should have $\text{invNorm}(0.95,100,20)$. If you leave out the μ and σ , then the default is the **z-score** for the standard normal distribution.

TI-89: Go to the [Apps] **Stat/List Editor**, then select F5 [DISTR]. This will get you a menu of probability distributions. Arrow down to **Inverse Normal** and press [ENTER]. Enter the area to the left of the x value, μ , and σ into each cell. Press [ENTER]. This will return the percentile for the x value. For example, to find the 95th percentile when the mean is 100 and the standard deviation is 20, you should enter 0.95, 100, 20.



If you use $\mu = 0$ and $\sigma = 1$, then the default is the **z-score** for the standard normal distribution.

Using Excel or TI-Calculator for the Percentile of a Normal Distribution

Figure 5-31 shows common phrases in questions that may help identify which function to use to find x value or z -score. Note that if use $\mu = 0$ and $\sigma = 1$, you would get the same answers as using the standard normal distribution function.

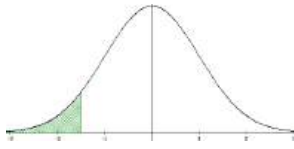
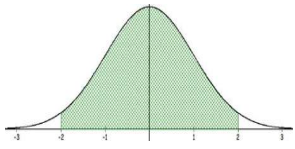
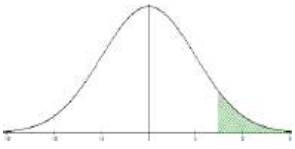
Normal Distribution Finding an x -value Given an Area or Probability		
$P(X \leq x)$ or $P(X < x)$	$P(x_1 < X < x_2)$ or $P(x_1 \leq X \leq x_2)$	$P(X \geq x)$ or $P(X > x)$
Lower	Between	Upper
Bottom		Top
Below		Above
Reduced		More than
Less than		Greater than
Lower than		Larger
Shorter than		Higher than
Smaller than		Longer than
Decreased		Bigger than
		Increased
		
Excel: $x = \text{NORM.INV}(\text{area}, \mu, \sigma)$	$x_1 = \text{NORM.INV}((1-\text{area})/2, \mu, \sigma)$ $x_2 = \text{NORM.INV}(1-(1-\text{area})/2, \mu, \sigma)$	$x = \text{NORM.INV}(1-\text{area}, \mu, \sigma)$
TI: $x = \text{invNorm}(\text{area}, \mu, \sigma)$	$x_1 = \text{invNorm}((1-\text{area})/2, \mu, \sigma)$ $x_2 = \text{invNorm}(1-(1-\text{area})/2, \mu, \sigma)$	$x = \text{invNorm}(1-\text{area}, \mu, \sigma)$

Figure 5-31

Example 5-36: In 2020, the average SAT mathematics score for Colorado was 501, with a standard deviation of 116. Find the SAT score that is for the top 10% of students taking the exam that year.

Solution: Use the phrases in Figure 5-31 to help you draw the distribution curve and shade the top 10% as shown in Figure 5-32. The “top” of the bell curve is the area to the right. However, most calculators use the area below the unknown x value. If you have 10% above the unknown x value, then there must be $100\% - 10\% = 90\%$ below.

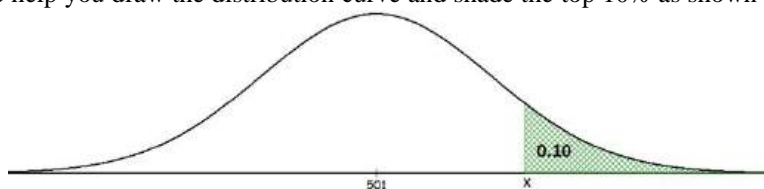


Figure 5-32

The area below the unknown x value in Figure 5-32 is $1 - 0.10 = 0.90$.

TI Calculator: $\text{invNorm}(0.9,501,116) = 649.66$. Excel: $=\text{NORM.INV}(0.9,501,116) = 649.66$.

A student that scored above 649.66 would be in the top 10%, also known as the 90th percentile.

We could have also found the z -score that corresponds to the top 10%. Then use the z -score formula to find the x -value. Using Excel $=\text{NORM.S.INV}(0.9) = 1.2816$. Then use the formula $x = z\sigma + \mu = 1.2816 \cdot 116 + 501 = 649.67$. This is using a rounded z -score so the answer is slightly off, but close.

Note: It is common practice to round z -scores to two decimal places. This is left over from using probability tables that only went out to two decimal places. If you use a rounded z -score in other calculations then keep in mind that you will get a larger rounding error.

Example 5-37: If the average price of a new one family home is \$246,300 with a standard deviation of \$15,000, compute the minimum and maximum prices of the houses that a contractor will build to satisfy the middle 98% of the market. Assume that the variable is normally distributed.

Solution: First, draw the curve and shade the middle area of 0.98, see Figure 5-33.

We need to get the area to the left of the x_1 value. Take the complement $1 - 0.98 = 0.02$, then split this area between both tails.

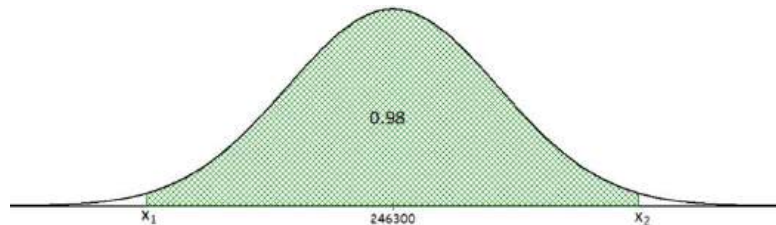


Figure 5-33

The lower tail area for x_1 would have $0.02/2 = 0.01$. The upper value of x_2 will have a left tail area of 0.99. On the calculator use, $\text{invNorm}(0.01,246300,15000)$ and you get a minimum price of \$211404.78 and use $\text{invNorm}(0.99,246300,15000)$ and you get a maximum price of \$281195.22.

In Excel you would have to do this in two separate cells $=\text{NORM.INV}(0.01,246300,15000) = \$211,404.78$ and $=\text{NORM.INV}(0.99,246300,15000) = \$281,195.22$.

A nice feature of this section is that the problems will say that the distribution is normally distributed, unlike the discrete distributions where you have to look for certain characteristics. However, when handling real data, you may have to know how to detect whether the data is normally distributed. One way to see if your variable is approximately normally distributed is by looking at a histogram, or we can use a normal probability plot.

5.5.3 Normal Probability Plot

A normal quantile plot, also called a normal probability plot, is a graph that is useful in assessing normality. A normal quantile plot plots the variable x against each of the x values corresponding z -score. It is not practical to make a normal quantile plot by hand.

Interpreting a normal quantile plot to see if a distribution is approximately normally distributed.

1. All of the points should lie roughly on a straight line $y = x$.
2. There should be no S pattern present.
3. Outliers appear as points that are far away from the overall pattern of the plot.

Here are two examples of histograms with their corresponding quantile plots. Note that as the distribution becomes closer to a normal distribution the dots on the quantile plot will be in a straighter line. Figure 5-34 is the histogram

and Figure 5-35 is the corresponding normal probability plot. Note the histogram is skewed to the left and dots do not line up on the $y = x$ line. Figures 5-36 and Figure 5-37 represent a sample that is approximately normally distributed. Note that the dots still do not line up perfectly on the line $y = x$, but they are close to the line.

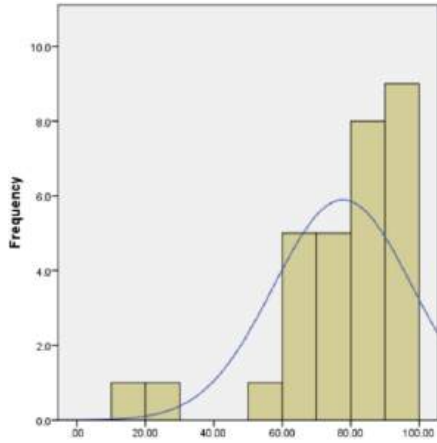


Figure 5-34

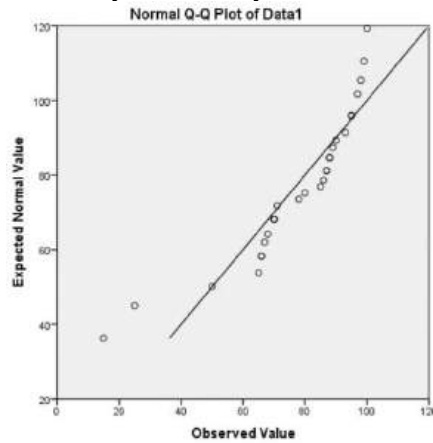


Figure 5-35

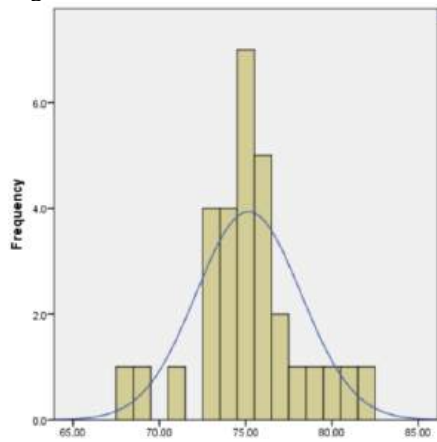


Figure 5-36

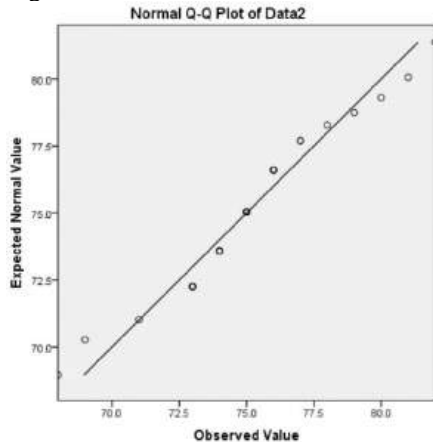


Figure 5-37

5.6 The Central Limit Theorem

The sample mean, denoted \bar{x} , is the average of a sample of a variable X . The sample mean is an estimate of the population mean μ . Every sample has a sample mean and these sample means differ (depending on the sample). Thus, before a sample is selected \bar{X} is a variable, in fact, if the sample is a random sample, then \bar{X} is a random variable. For this reason, we can think of the “distribution of \bar{X} ,” called the “Sampling Distribution of \bar{X} ,” as the theoretical histogram constructed from the sample averages of all possible samples of size n .

Mean and Standard Deviation of a Sample Mean

Let \bar{x} be the mean of a random sample of size n from a population having mean μ and standard deviation σ , then:

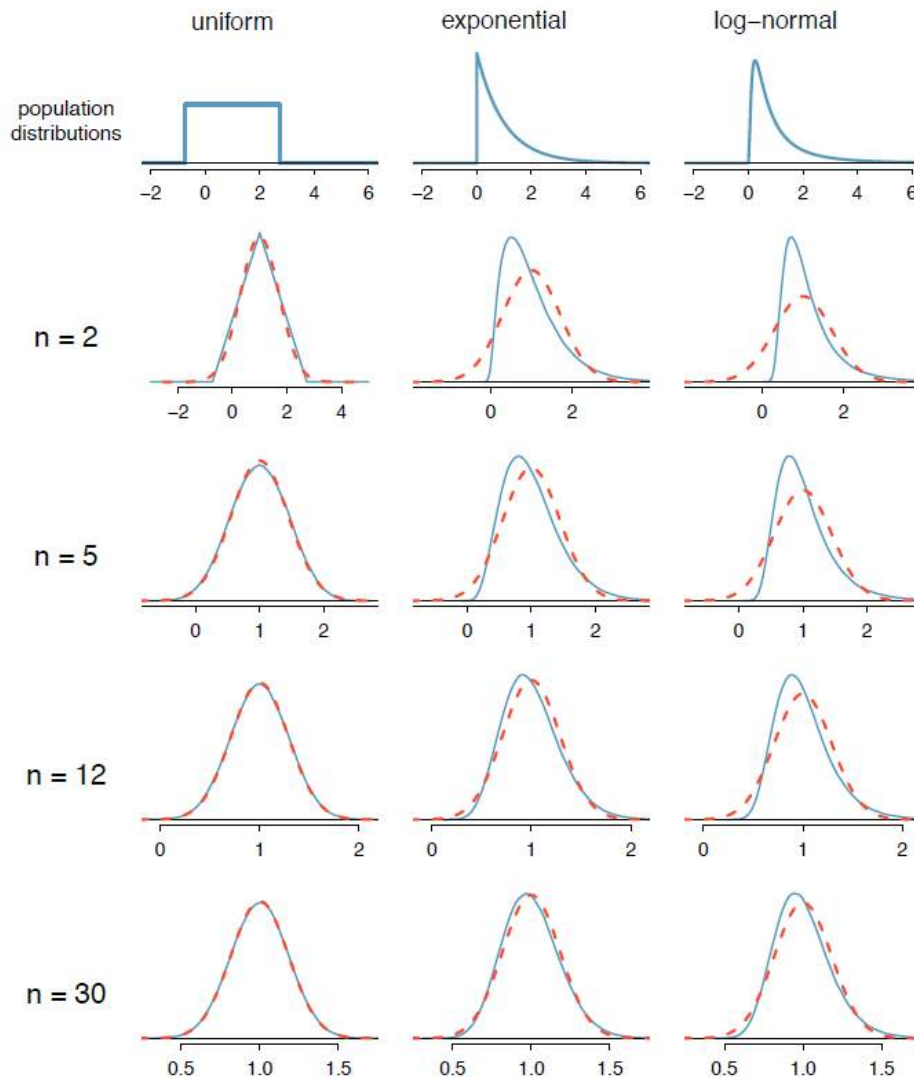
- the mean of the sample means = $\mu_{\bar{x}} = \mu$,
- and the standard deviation (standard error) of the sample means = $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

This says that the mean of the sample means is the same as the population mean. The standard deviation of the sample means is the population standard deviation divided by the square root of the sample size. This is called the **sampling distribution** of the mean. The mean of the sampling distribution of the mean is denoted as $\mu_{\bar{x}}$. The standard deviation of the sampling distribution of the mean is denoted as $\frac{\sigma}{\sqrt{n}}$, which is also called the standard error.

Sampling Distribution of a Sample Mean

If a population is normally distributed $N(\mu, \sigma)$, then the sample mean \bar{X} of n independent observations is normally distributed as $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Figure 5-38 shows three population distributions and the corresponding sampling distributions for sample sizes of 2, 5, 12 and 30. Notice as the sample size gets larger, the sampling distribution gets closer to the dashed red line of the normal distribution. **Video explanation of this process:** https://youtu.be/lsCc_pS3O28.



Retrieved from [OpenIntroStatistics](https://openintrostatistics.com).

Figure 5-38

The Central Limit Theorem establishes that in some situations the distribution of the sample statistic will take on a normal distribution, even when the population is not normally distributed. This allows us to use the normal distribution to make inferences from samples to populations.

Example 5-38: Let X be the height of 15-year old boys in the United States. Studies show that the heights of 15-year old boys in the United States are normally distributed with average height 67 inches and a standard deviation of 2.5 inches. A random experiment consists of choosing 16, 15-year old boys at random. Compute the mean and standard deviation of \bar{X} , that is, the mean and standard deviation for the average height of a random sample of 16 boys.

Solution: The mean of the sample means is the same as the population mean $\mu_{\bar{x}} = 67$.

The standard deviation in the sample means is the population standard deviation divided by the square root of the sample size, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{16}} = 0.625$.

Notice that the mean of a sample means is always the same as the mean of the population, but the standard deviation is smaller. See Figure 5-39.

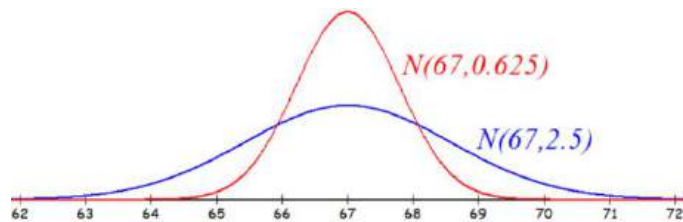


Figure 5-39

Central Limit Theorem

Draw a random sample of size n from any population having population mean μ and finite standard deviation σ .

When n is large ($n \geq 30$), the sampling distribution of \bar{x} is approximately Normal or $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

This means that if we are trying to find the probability of a sample mean for a group that is 30 or more, you do not need to assume that the population is normally distributed, since the sampling distribution will be approximately normal.

The Central Limit Theorem guarantees that the distribution of the sample mean will be normally distributed when the sample size is large (usually 30 or higher) no matter what shape the population distribution is.

Finding Probabilities Using the Central Limit Theorem (CLT)

If we are finding the probability of a sample mean and have a sample size of 30 or more, or the population was normally distributed, then we can use the normal distribution to find the probability that the sample mean is below, above or between two values using the CLT.

Watch this video on using this applet for the Central Limit Theorem, and then take some time to play with the applet to get a sense of the difference between the distribution of the population, the distribution of a sample and the sampling distribution.

Watch the video on how to use the applet: <https://youtu.be/aIPvgiXyBMI>.

Try the applet on your own. Applet: http://onlinestatbook.com/stat_sim/sampling_dist/index.html.

Example 5-39: The population of midterm scores for all students taking a PSU Business Statistics course has a known standard deviation of 5.27. The mean of the population is 18.07 and the median of the population is 19. A sample of 25 was taken and the sample mean was 18.07 and we want to know what the sampling distribution for the mean looks like. Figure 5-40 shows 3 graphs using the [Sampling Distribution Applet](#), but the order has been mixed up.

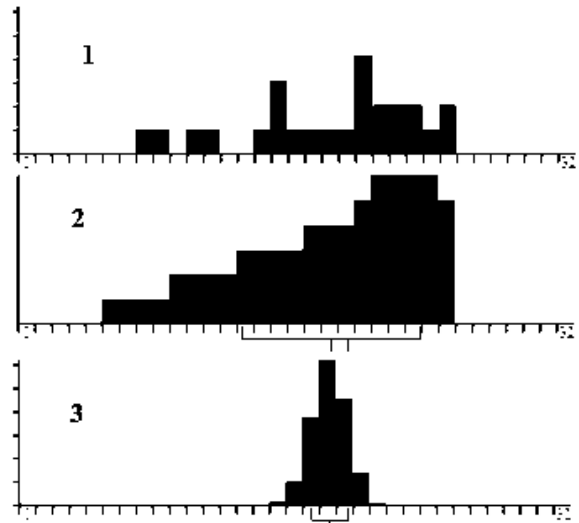


Figure 5-40

- a) What is the mean and standard deviation of the sampling distribution?

Solution: By the Central Limit Theorem (CLT) the mean of the sampling distribution $\mu_{\bar{x}}$ equals the mean of the population which was given as $\mu=18.07$. The standard deviation of the sampling distribution by the CLT would be the population standard deviation divided by the square root of the sample size $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5.27}{\sqrt{25}} = 1.054$.

- b) Would you expect midterm exam scores to be skewed or bell-shaped?

Solution: The population mean = 18.07 is smaller than the median = 19 therefore the distribution is negatively skewed, the mean is pulled in the direction of the outliers.

- c) Which of these graphs in Figure 5-40 correspond to the distribution of the population, distribution of a single sample and the sampling distribution of the mean?

Solution: Using the [Sampling Distribution Applet](#) and the CLT, the sampling distribution will be bell-shaped therefore, graph 3 has to be the sampling distribution.

Graphs 1 & 2 in Figure 5-40 are both negatively skewed. A single sample of 25 should look similar to the entire population, but we would expect only 25 items and not every score possible would be received from the 25 students. Graph 1 in Figure 5-40 fits this description and therefore the graph of the distribution of a single sample (which is not the same thing as the sampling distribution) is graph 1.

This leaves graph 2 as the distribution of the population (population distribution).

Figure 5-41 is a picture of the applet modeling the exam scores. Note the top picture is the population distribution, the second graph is simulating a single sample drawn and the bottom picture is a graph of all the sample means for each sample. This last graph is the sampling distribution of the means.

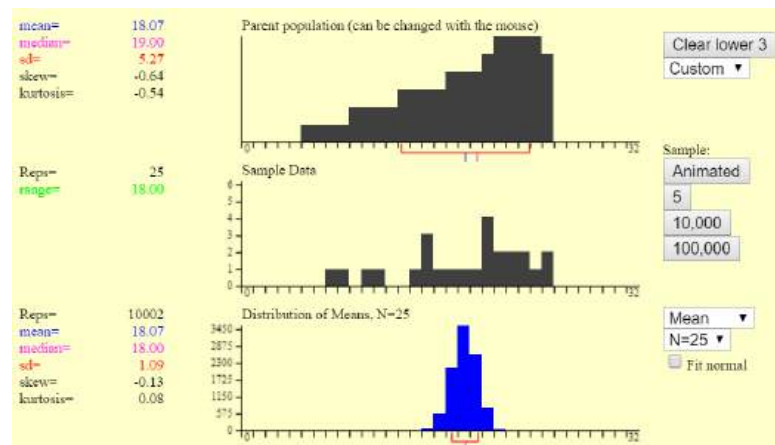


Figure 5-41

- d) Compute the probability that for next term's class they have a sample mean of more than 20.

Solution: The $P(\bar{X} > 20)$ would be normally distributed with a mean $\mu_{\bar{x}} = 18.07$ with a standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5.27}{\sqrt{25}} = 1.054$.

Draw and shade the sampling distribution curve.

This calculator can be used to draw and shade the sampling distribution:

<http://homepage.divms.uiowa.edu/~mbognar/applets/normal.html>, filling in the mean μ , standard deviation $\frac{\sigma}{\sqrt{n}}$ and x-value (in this case the sample mean) will find the probability. See Figure 5-42.

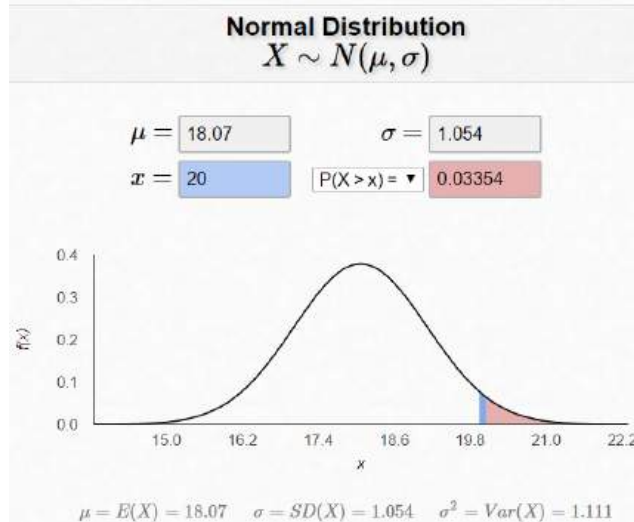


Figure 5-42

TI Calculator: $P(\bar{X} > 20) = \text{normalcdf}(20, 1E99, 18.07, 5.27/\sqrt{(25)}) = 0.0335$.

Excel: $P(\bar{X} > 20) = 1 - \text{NORM.DIST}(20, 18.07, 5.27/\text{SQRT}(25), \text{TRUE}) = 0.0335$.

Example 5-40: Let X be the height of 15-year old boys in the United States. Studies show that the heights of 15-year old boys in the United States are normally distributed with average height of 67 inches and a standard deviation of 2.5 inches. A random experiment consists of randomly choosing sixteen 15-year old boys. Compute the probability that the mean height of those sampled is 69.5 inches or taller.

Solution: The sample mean \bar{X} is approximately $N(67, 0.625)$. $P(\bar{X} \geq 69.5) = P\left(\frac{\bar{X} - 67}{0.625} \geq \frac{69.5 - 67}{0.625}\right) = P(Z \geq 4) \approx 0.00003$, using the calculator, be careful with the scientific notation. This is a very small probability. This should make sense because one would think that the likelihood of randomly selecting 16 boys that have an average height of 5'9.5" would be slim.

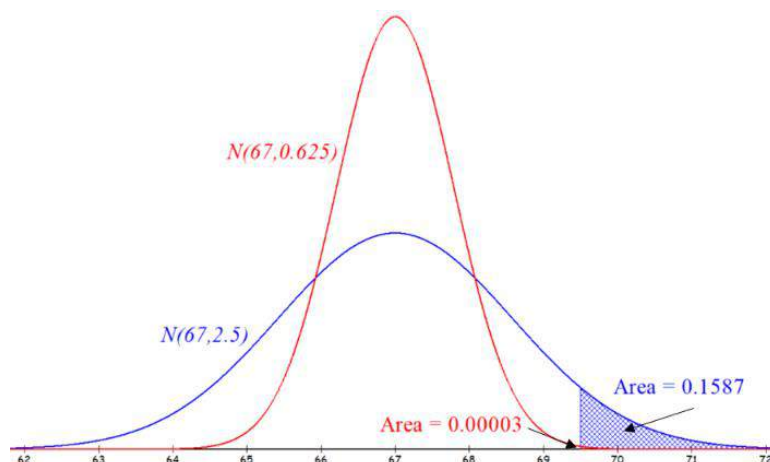
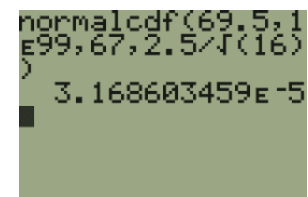


Figure 5-43

Figure 5-43 shows the density curves showing the shaded areas of $P(X \geq 69.5)$ on the population distribution in blue and $P(\bar{X} \geq 69.5)$ on the sampling distribution in red.

The sampling distribution has a much smaller spread (standard deviation) and hence less area to the right of 69.5.

In general, the Central Limit Theorem questions will use the same method as previous sections, however you will use a standard deviation of $\frac{\sigma}{\sqrt{n}}$ and a z-score of $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$.

Note that the question does not have to state that the population is normally distributed as long as the sample size is 30 or more. If the sample size is less than 30, then the population distribution needs to be approximately normally distributed. The questions using the Central Limit Theorem will be asking to find a probability of a sample mean, not an individual item.

Example 5-41: The average teacher's salary in Connecticut (ranked first among states) is \$57,337. Suppose that the distribution of salaries is normally distributed with a standard deviation of \$7,500.

- a) What is the probability that a randomly selected teacher makes less than \$55,000 per year?

Solution: Find $P(X < 55000)$, since we are only looking at one person use $z = \frac{x - \mu}{\sigma}$. If we were asked to standardize the salary, $z = \frac{55000 - 57337}{7500} = -0.3116$, however we can use technology and skip this step.

Use the `normalcdf(-1E99, 55000, 57337, 7500)` (the TI-89 use `-∞` for the lower boundary instead of `-1E99`) and you get the probability of 0.3777. The $P(X < 55000) = P(Z < -0.3116) = 0.3777$.

```
normalcdf(-1E99,
55000, 57337, 7500
)
.3776723318
```

Note that we are not using the CLT since we are not finding the probability of an average for a group of people, just the probability for one person.

- b) If we sample 10 teachers' salaries, what is the probability that the sample mean is less than \$55,000?

Solution: Find $P(\bar{X} < 55000)$, but we are looking at the probability of a mean for 10 teachers, so use $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$. Standardize the salary, $z = \frac{55000 - 57337}{\left(\frac{7500}{\sqrt{10}}\right)} = -0.9853657189$, use your calculator to get $P(\bar{X} < 55000) = P(Z < -0.9853657189) = 0.1622$.

```
(55000 - 57337) / (7
500 / sqrt(10))
- .9853657189
normalcdf(-1E99,
-.9853657189, 0, 1
)
.162222231
```

You do not need the extra step of finding the z-score first. Instead, you can use `normalcdf(-1E99, 55000, 57337, 7500/sqrt(10)) = 0.1622`.

- c) If we sample 100 teachers' salaries, what is the probability that the sample mean is less than \$55,000?

Solution: Find $P(\bar{X} < 55000)$, but we are looking at the probability of a mean for 100 teachers, so use $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$. Standardize the salary, $z = \frac{55000 - 57337}{\left(\frac{7500}{\sqrt{100}}\right)} = -3.116$ use your calculator `normalcdf(-1E99, 55000, 57337, 7500/sqrt(100))` to get $P(\bar{X} < 55000) = P(Z < -3.116) = 0.0009167$.

```
(55000 - 57337) / (7
500 / sqrt(100))
-3.116
normalcdf(-1E99,
-3.116, 0, 1)
9.166789234E-4
```

As the sample size increase, the probability of seeing a sample mean of less than \$55,000 is getting smaller.

When you have a z-score that is less than -3 or greater than 3 we would call this a rare event or outlier. We will use a similar process in inferential statistics.

5.7 Student's T-Distribution

A t-distribution is another symmetric distribution for a continuous random variable.

William Gosset was a statistician employed at Guinness and performed statistics to find the best yield of barley for their beer. Guinness prohibited its employees to publish papers so Gosset published under the name Student. Gosset's distribution is called the Student's t-distribution.



A **t-distribution** is another special type of distribution for a continuous random variable.

Properties of the t-distribution density curve:

1. Symmetric, Unimodal (one mode)
Bell-shaped.
2. Centered at the mean $\mu = \text{median} = \text{mode} = 0$.
3. The spread of a t-distribution is determined by the degrees of freedom which are determined by the sample size.
4. As the degrees of freedom increase, the t-distribution approaches the standard normal curve.
5. The total area under the curve is equal to 1 or 100%.

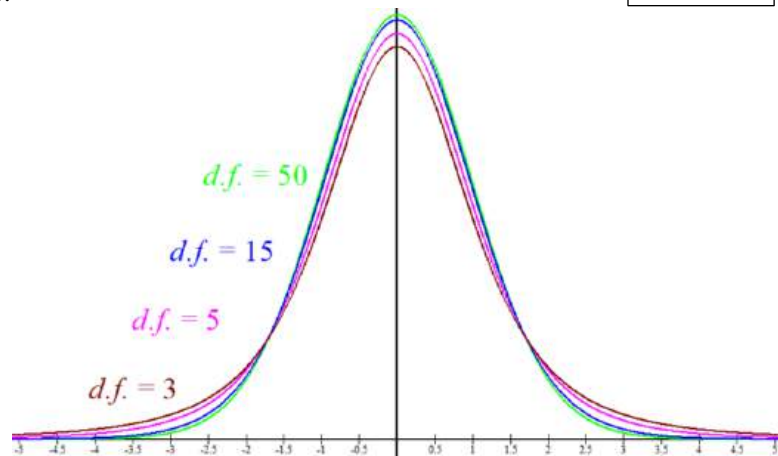


Figure 5-44

Figure 5-44 shows examples of four different t-distributions with degrees of freedom of 3, 5, 15 and 50. Note that as the degrees of freedom (df) increase the distribution has a smaller standard deviation and will get closer in shape to the normal distribution.

Example 5-42: Calculate $P(t \leq -2.492)$ for $n = 25$.

Solution: Use a t-distribution with $df = n - 1 = 25 - 1 = 24$, and find the area below $t = -2.492$. Draw the curve and shade the area to the left of -2.492 as shown in Figure 5-45.

```
tcdf(-1E99, -2.492, 24)
.010003546
```

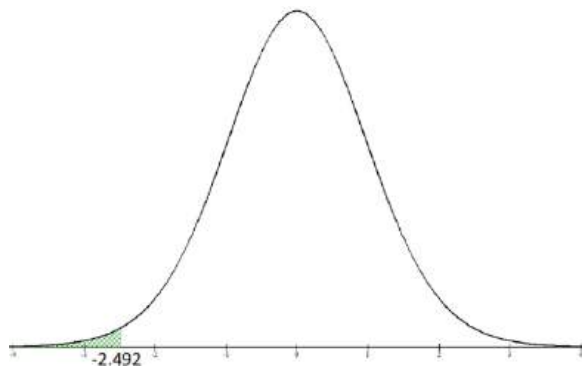


Figure 5-45

On the TI calculator, use the DISTR menu tcdf option. The calculator wants the lower and upper t scores for the area, and df . For this example, we would use $\text{tcdf}(-\infty, -2.492, 24)$. $P(t \leq -2.492) = 0.01$.

Excel: $t = \text{T.DIST}(-2.492, 24, \text{TRUE}) = 0.01$.

Example 5-43: Calculate $P(t > 1.584)$ for $df = 30$.

Solution: Use a t-distribution with $df = 30$, and find the area above $t = 1.584$. Draw the curve and shade the area to the right of 1.584 as shown in Figure 5-46.

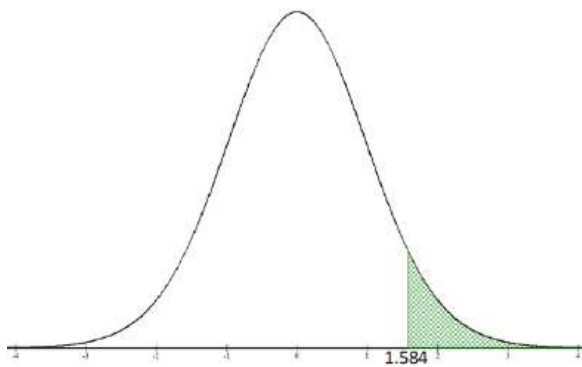


Figure 5-46

On the TI calculator, use the DISTR menu tcdf option. The calculator wants the lower and upper t scores for the area, and df . For this example, we would use $\text{tcdf}(1.584, \infty, 30)$. $P(t > 1.584) = 0.0618$.

Excel only finds the cumulative area below a t-score. This means that you need to use the complement rule when finding the area above a t-score. Use the Excel function $=1-\text{T.DIST}(1.584, 30, \text{TRUE}) = 0.0618$.

Example 5-44: Compute the probability of getting a t-score larger than 1.8399 with a sample size of 13.

Solution: Draw the bell curve and shade the area to the left of $t = 1.8399$ as shown in Figure 5-47.

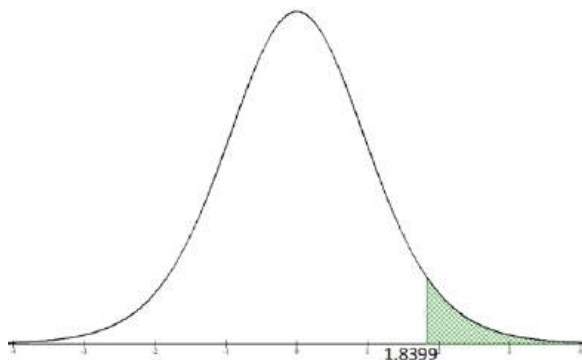
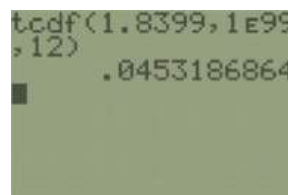


Figure 5-47

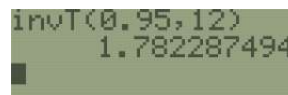
To find the $P(t > 1.8399)$ on the TI calculator, go to DISTR use $\text{tcdf}(\text{lower}, \text{upper}, df)$. For this example, we would have $\text{tcdf}(1.8399, \infty, 12)$. Note that the TI-83 and 84 calculators do not have an infinity ∞ symbol so use the EE button for scientific notation to get 1E99 for a very large number. $P(t > 1.8399) = \text{tcdf}(1.8399, 1\text{E}99, 12) = 0.0453$.



In Excel use $=1-\text{T.DIST}(1.8399, 12, \text{TRUE}) = 0.0453$.

Example 5-45: Calculate the t-score that has 5% of the area in the upper tail for $n = 13$.

Solution: Use a t-distribution with the degrees of freedom, $df = n - 1 = 13 - 1 = 12$. Draw and shade the upper tail area as in Figure 5-48.



Label the areas, let $\alpha = 0.05$, then the area below the unknown t-score would use the complement $1 - \alpha = 1 - 0.05 = 0.95$.

In Excel, use the function $=\text{T.INV}(0.95, 12) = 1.7823$.

In the TI calculator, use the DISTR menu invT option.

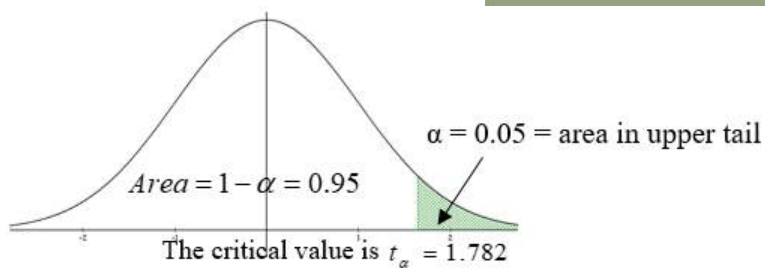


Figure 5-48

For the invT function, you always use the area to the left of the point. If want 5% in the upper tail, then that means there is 95% in the bottom tail area. $t_\alpha = \text{invT}(\text{area } \alpha \text{ below t-score}, df) = \text{invT}(0.95, 12) = 1.782$.

Note that if you have an older TI-84 or a TI-83 calculator you need to have the program INVT installed. You can download the INVT program to your calculator from <http://MostlyHarmlessStatistics.com> or use Excel.

Example 5-46: Calculate the t-scores that have the middle 90% of the area in the between them for $n = 10$.

Solution: Use a t-distribution with the degrees of freedom, $df = n - 1 = 10 - 1 = 9$. Draw a bell curve and label 0 in the middle. Draw two lines equal distance from the mean such that approximately 90% of the distribution is shaded. Let the area in both tails, called α , be the complement of the middle area: $\alpha = 1 - 0.90 = 0.10$. Divide α by 2, to get: $\alpha/2 = 0.10/2 = 0.05$. Both the lower and upper tail areas shaded in Figure 5-49 have an area of 0.05. Use the TI calculator's DISTR menu, then the invT option. The function wants the area below a the t-score, so we would need to do the function separately for both sides. The left hand t-score would have area = 0.05 below it, so use $\text{invT}(\text{lower tail area}, df) = \text{invT}(0.05, 9) = -1.8331$. The right hand t-score would have the area = $0.05 + 0.90 = 0.95$ below the t-score, to get $\text{invT}(\text{lower tail area}, df) = \text{invT}(0.95, 9) = 1.8331$.

In Excel use $=\text{T.INV}(\text{lower area}, df) = \text{T.INV}(0.05, 9) = -1.8331$ and $=\text{T.INV}(\text{lower area}, df) = \text{T.INV}(0.95, 9) = 1.8331$.

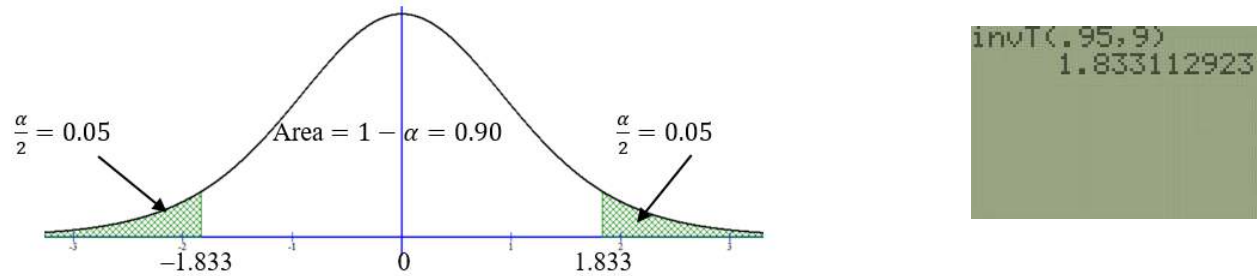
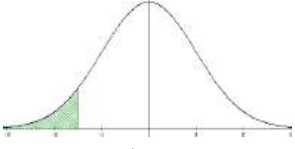
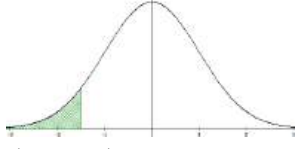
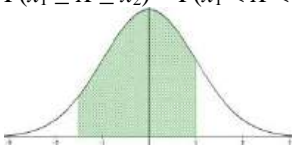
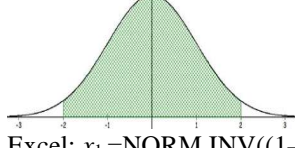
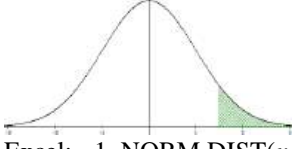
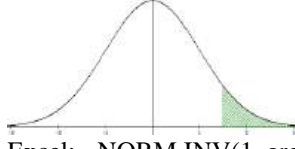
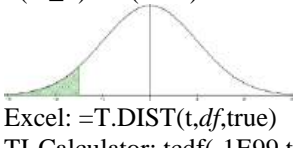
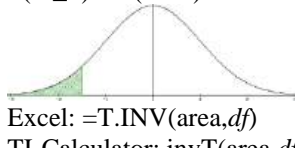
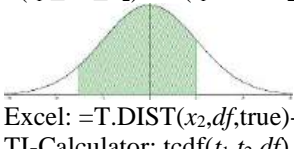
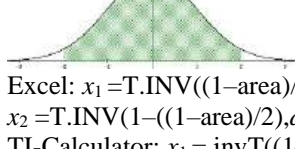
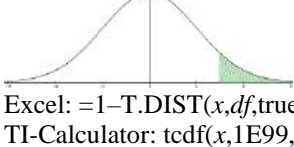
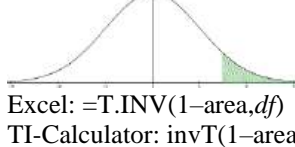


Figure 5-49

The two t-scores, $t = \pm 1.8331$, would give the middle 90% of the t-distribution.

We call these t-scores critical values in the next chapter. Sometimes we use the $\alpha/2$ label of the tail area to distinguish these t-scores from other t-scores that we calculate from sample data and are written as $t_{\alpha/2} = -1.8331$.

Chapter 5 Formulas

<p>Discrete Distribution Table: $0 \leq P(x_i) \leq 1$ $\sum P(x_i) = 1$ Mean: $\mu = \sum(x_i \cdot P(x_i))$ Variance: $\sigma^2 = \sum(x_i^2 \cdot P(x_i)) - \mu^2$ Standard Deviation: $\sigma = \sqrt{\sigma^2}$</p>	<p>Binomial Distribution: $P(X = x) = {}_n C_x \cdot p^x \cdot q^{(n-x)}$, $x = 0, 1, 2, \dots, n$ Mean: $\mu = n \cdot p$ Variance: $\sigma^2 = n \cdot p \cdot q$ Standard Deviation: $\sigma = \sqrt{n \cdot p \cdot q}$</p>
<p>Normal Distribution Probabilities:</p>  <p>$P(X \leq x) = P(X < x)$ Excel: =NORM.DIST(x,μ,σ,true) TI-Calculator: normalcdf(-1E99,x,μ,σ)</p>	<p>Percentiles for Normal Distribution:</p>  <p>$P(X \leq x) = P(X < x)$ Excel: =NORM.INV(area,μ,σ) TI-Calculator: invNorm(area,μ,σ)</p>
<p>$P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2) =$</p>  <p>Excel: =NORM.DIST(x2,μ,σ,true)-NORM.DIST(x1,μ,σ,true) TI-Calculator: normalcdf(x1,x2,μ,σ)</p>	<p>$P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2) =$</p>  <p>Excel: $x_1 = \text{NORM.INV}((1-\text{area})/2, \mu, \sigma)$ $x_2 = \text{NORM.INV}(1-((1-\text{area})/2), \mu, \sigma)$ TI-Calculator: $x_1 = \text{invNorm}((1-\text{area})/2, \mu, \sigma)$ $x_2 = \text{invNorm}(1-((1-\text{area})/2), \mu, \sigma)$</p>
<p>$P(X \geq x) = P(X > x)$</p>  <p>Excel: =1-NORM.DIST(x,μ,σ,true) TI-Calculator: normalcdf(x,1E99,μ,σ)</p>	<p>$P(X \geq x) = P(X > x)$</p>  <p>Excel: =NORM.INV(1-area,μ,σ) TI-Calculator: invNorm(1-area,μ,σ)</p>
<p>T-Distribution Probabilities: $P(T \leq t) = P(T < t)$</p>  <p>Excel: =T.DIST(t,df,true) TI-Calculator: tcdf(-1E99,t,df)</p>	<p>Percentiles for t-Distribution: $P(T \leq t) = P(T < t)$</p>  <p>Excel: =T.INV(area,df) TI-Calculator: invT(area,df)</p>
<p>$P(t_1 \leq T \leq t_2) = P(t_1 < T < t_2) =$</p>  <p>Excel: =T.DIST(x2,df,true)-T.DIST(x1,df,true) TI-Calculator: tcdf(t1,t2,df)</p>	<p>$P(t_1 \leq T \leq t_2) = P(t_1 < T < t_2) =$</p>  <p>Excel: $x_1 = \text{T.INV}((1-\text{area})/2, df)$, $x_2 = \text{T.INV}(1-((1-\text{area})/2), df)$ TI-Calculator: $x_1 = \text{invT}((1-\text{area})/2, df)$, $x_2 = \text{invT}(1-((1-\text{area})/2), df)$</p>
<p>$P(T \geq t) = P(T > t)$</p>  <p>Excel: =1-T.DIST(x,df,true) TI-Calculator: tcdf(x,1E99,df)</p>	<p>$P(T \geq t) = P(T > t)$</p>  <p>Excel: =T.INV(1-area,df) TI-Calculator: invT(1-area,df)</p>

Chapter 5 Exercises

1. Determine if the following tables are valid discrete probability distributions. If they are not state why.

a)

x	-5	-2.5	0	2.5	5
$P(X = x)$	0.15	0.25	0.32	0.18	0.1

b)

x	0	1	2	3	4
$P(X = x)$	0.111	0.214	0.312	0.163	0.159

c)

x	0	1	2	3	4
$P(X = x)$	0.2	-0.3	0.5	0.4	0.2

2. The random variable X = the number of vehicles owned.

x	0	1	2	3	4
$P(X = x)$	0.1	0.35	0.25	0.2	0.1

- Compute the probability that a person owns at least 2 vehicles.
- Compute the $P(X > 2)$.
- Compute the probability that a person owns less than 2 vehicles.
- Compute the expected number of vehicles owned.
- Compute the standard deviation of the number of vehicles owned.
- Compute σ^2 .

3. The following discrete probability distribution represents the amount of money won for a raffle game.

x	-5	-2.5	0	2.5	5
$P(X = x)$	0.15	0.25	0.32	0.18	0.1

- Compute μ .
- Compute σ .

4. Keke's Kookies sells mini cookies in packs of 5 and has determined a probability distribution for the number of cookies that they sell in a given day.

$x = \#$ sold	0	5	10	15	20
$P(X = x)$	0.22	0.38	0.14	...?	0.07

- What is the probability of selling 15 mini cookies in a given day?
- Find the expected number of mini cookies sold in a day using the discrete probability distribution.
- Find the variance of the number of mini cookies sold in a day using the discrete probability distribution.

- The bookstore also offers a chemistry textbook for \$159 and a book supplement for \$41. From experience, they know about 25% of chemistry students just buy the textbook while 60% buy both the textbook and supplement, the remaining 15% of students do not buy either book. Compute the standard deviation of the bookstore revenue.
- A \$100,000 life insurance policy for a 50-year-old woman has an annual cost of \$335. The probability that a 50-year-old woman will die is 0.003118. What is the expected value of the policy for the woman's estate?
- An LG Dishwasher, which costs \$1000, has a 24% chance of needing to be replaced in the first 2 years of purchase. If the company has to replace the dishwasher within the two-year extended warranty, it will cost the company \$112.10 to replace the dishwasher.

- Fill out the probability distribution for the value of the extended warranty from the perspective of the company.

x		
$P(X = x)$		

- What is the expected value of the extended warranty?

c) Write a sentence interpreting the expected value of the warranty.

8. The Oregon lottery has a game called Pick 4 where a player pays \$1 and picks a four-digit number. If the four numbers come up in the order you picked, then you win \$2000.

a) Fill out the probability distribution for a player's winnings.

x		
$P(X = x)$		

b) What are your expected winnings?

c) Write a sentence interpreting the expected winnings.

9. The following table represents the probability of the number of pets owned by a college student.

x	0	1	2	3
$P(X = x)$	0.46	0.35	0.12	0.07

a) Is this a valid discrete probability distribution? Explain your answer.

b) Find the mean number of pets owned.

c) Find the standard deviation of the number of cars owned.

d) Find σ^2 .

10. Suppose a random variable, x , arises from a binomial experiment. If $n = 14$, and $p = 0.13$, compute the following.

a) $P(X = 3)$

b) $P(X \leq 3)$

c) $P(X < 3)$

d) $P(X > 3)$

e) $P(X \geq 3)$

f) Mean

g) Variance

h) Standard Deviation

11. Suppose a random variable, X , arises from a binomial experiment. If $n = 25$, and $p = 0.85$, compute the following.

a) $P(X = 15)$

b) $P(X \leq 15)$

c) $P(X < 15)$

d) $P(X > 15)$

e) $P(X \geq 15)$

f) μ

g) σ

h) σ^2

12. An unprepared student takes a 10 question TRUE/FALSE quiz and ended up guessing each answer.

a) What is the probability that the student got 7 questions correct?

b) What is the probability that the student got 7 or more questions correct?

13. A local county has an unemployment rate of 7.3%. A random sample of 20 employable people are picked at random from the county and are asked if they are employed. The distribution is a binomial. Round answers to 4 decimal places.

a) Find the probability that exactly 3 in the sample are unemployed.

b) Find the probability that there are fewer than 4 in the sample are unemployed.

c) Find the probability that there are more than 2 in the sample are unemployed.

d) Find the probability that there are at most 4 in the sample are unemployed.

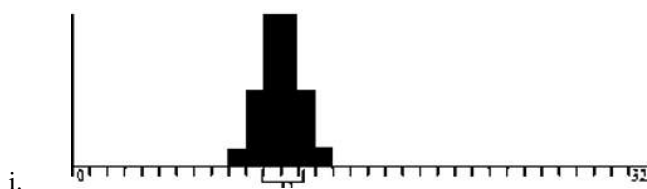
14. A fair coin is flipped 30 times.
- What is the probability of getting exactly 15 heads?
 - What is the probability of getting 15 or more heads?
 - What is the probability of getting at most 15 heads?
 - How many times would you expect to get heads?
 - What is the standard deviation of the number of heads?
15. Approximately 10% of all people are left-handed. Out of a random sample of 15 people, find the following.
- What is the probability that 4 of them are left-handed?
 - What is the probability that less than 4 of them are left-handed?
 - What is the probability that at most 4 of them are left-handed?
 - What is the probability that at least 4 of them are left-handed?
 - What is the probability that more than 4 of them are left-handed?
 - Compute μ .
 - Compute σ .
 - Compute σ^2 .
16. Approximately 8% of all people have blue eyes. Out of a random sample of 20 people, find the following.
- What is the probability that 2 of them have blue eyes?
 - What is the probability that at most 2 of them have blue eyes?
 - What is the probability that less than 2 of them have blue eyes?
 - What is the probability that at least 2 of them have blue eyes?
 - What is the probability that more than 2 of them have blue eyes?
 - Compute μ .
 - Compute σ .
 - Compute σ^2 .
17. About 1% of the population has a particular genetic mutation. Find the standard deviation for the number of people with the genetic mutation in a group of 100 randomly selected people from the population.
18. You really struggle remembering to bring your lunch to work. Each day seems to be independent as to whether you remember to bring your lunch or not. The chance that you forget your lunch each day is 25.6%. Consider the next 48 days. Let X be the number of days that you forget your lunch out of the 48 days. Compute $P(10 \leq X \leq 14)$.
19. A poll is given, showing 72% are in favor of a new building project. Let X be the number of people who favor the new building project when 37 people are chosen at random. What is the probability that between 10 and 16 (including 10 and 16) people out of 37 favor the new building project?
20. A flu vaccine has a 90% effective rate. If a random sample of 200 people are given the vaccine, what is the probability that at most 180 people did not get the flu?
21. The Lee family had 6 children. Assuming that the probability of a child being a girl is 0.5, find the probability that the Lee family had at least 4 girls?
22. If a seed is planted, it has a 70% chance of growing into a healthy plant. If 142 randomly selected seeds are planted, answer the following.
- What is the probability that exactly 100 of them grow into a healthy plant?
 - What is the probability that less than 100 of them grow into a healthy plant?
 - What is the probability that more than 100 of them grow into a healthy plant?
 - What is the probability that exactly 103 of them grow into a healthy plant?
 - What is the probability that at least 103 of them grow into a healthy plant?
 - What is the probability that at most 103 of them grow into a healthy plant?

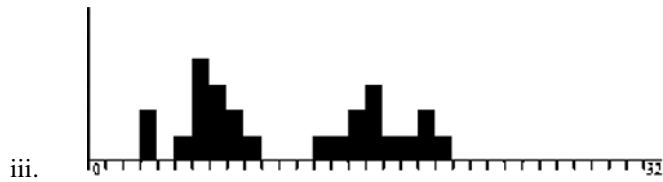
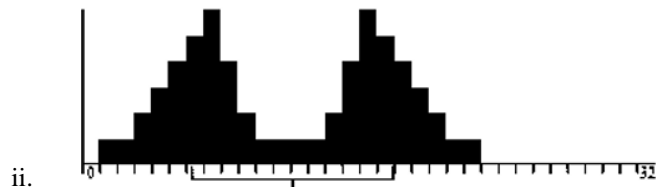
23. A manufacturing machine has a 6% defect rate. An inspector chooses 4 items at random.
- What is the probability that at least one will have a defect?
 - What is the probability that exactly two will have a defect?
 - What is the probability that less than two will have a defect?
 - What is the probability that more than one will have a defect?
24. A large fast-food restaurant is having a promotional game where game pieces can be found on various products. Customers can win food or cash prizes. According to the company, the probability of winning a prize (large or small) with any eligible purchase is 0.162. Consider your next 33 purchases that produce a game piece. Calculate the following:
- What is the probability that you win 5 prizes?
 - What is the probability that you win more than 8 prizes?
 - What is the probability that you win between 3 and 7 (inclusive) prizes?
 - What is the probability that you win 3 prizes or fewer?
25. A small regional carrier accepted 20 reservations for a particular flight with 17 seats. 15 reservations went to regular customers who will arrive for the flight. Each of the remaining passengers will arrive for the flight with a 60% chance, independently of each other.
- Find the probability that overbooking occurs.
 - Find the probability that the flight has empty seats.
26. Scores on the SAT for a certain year were bell-shaped with a mean of 1511 and a standard deviation of 194. Use the empirical rule.
- What two SAT scores that separated the middle 68% of SAT scores for that year?
 - What two SAT scores that separated the middle 95% of SAT scores for that year?
 - How high did a student need to score that year to be in the top 2.5%?
27. In a mid-size company, the distribution of the number of phone calls answered each day by each of the 12 employees is bell-shaped and has a mean of 59 and a standard deviation of 10. Using the empirical rule, what is the approximate percentage of daily phone calls numbering between 29 and 89?
28. The number of potholes in any given 1 mile stretch of pavement in Portland has a bell-shaped distribution. This distribution has a mean of 54 and a standard deviation of 5. Using the empirical rule, what is the approximate percentage of 1-mile long roadways with potholes numbering between 44 and 59?
29. A company has a policy of retiring company cars; this policy looks at number of miles driven, purpose of trips, style of car and other features. The distribution of the number of months in service for the fleet of cars is bell-shaped and has a mean of 42 months and a standard deviation of 3 months. Using the empirical rule, what is the approximate percentage of cars that remain in service between 48 and 51 months?
30. Compute the following probabilities where $Z \sim N(0,1)$.
- $P(Z < 1.57)$
 - $P(Z > -1.24)$
 - $P(-1.96 \leq Z \leq 1.96)$
 - $P(Z \leq 3)$
 - $P(1.31 < Z < 2.15)$
 - $P(Z \geq 1.8)$

31. For a standard normal distribution, find the following probabilities.
- $P(Z > -2.06)$
 - $P(-2.83 < Z < 0.21)$
 - $P(Z < 1.58)$
 - $P(Z \geq 1.69)$
 - $P(Z < -2.82)$
 - $P(Z > 2.14)$
 - $P(1.97 \leq Z \leq 2.93)$
 - $P(Z \leq -0.51)$
32. Use the area under the standard normal distribution for the following.
- Compute the area to the left of $z = -1.05$.
 - Compute the area to the left of $z = -0.69$.
 - Compute the area to the right of $z = 2.08$.
 - Compute the area to the right $z = 1.22$.
 - Compute the area between $z = -0.29$ and $z = 0.14$.
 - Compute the area between $z = -2.97$ and $z = -2.14$.
33. Compute the following probabilities where $Z \sim N(0,1)$.
- $P(Z \leq -2.03)$
 - $P(Z > 1.58)$
 - $P(-1.645 \leq Z \leq 1.645)$
 - $P(Z < 2)$
 - $P(-2.38 < Z < -1.12)$
 - $P(Z \geq -1.75)$
34. Use the area under the standard normal distribution for the following.
- Compute the area to the right of $z = 2.5758$.
 - Compute the area to the right of $z = -1.3671$.
 - Compute the area to the left of $z = 0.85$.
 - Compute the area to the left of $z = -1.645$.
 - Compute the area between $z = -1.96$ and $z = 1.96$.
 - Compute the area between $z = 1.734$ and $z = 2.583$.
35. Use the standard normal distribution for the following.
- Compute the z -score that gives the 29th percentile.
 - Compute the z -score that gives the 75th percentile.
 - Compute the two z -scores that give the middle 99% area.
36. Use the standard normal distribution for the following.
- Compute the z -score that gives the 85th percentile.
 - Compute the two z -scores that give the middle 95% area.
 - Compute the IQR.
37. Arm span is the physical measurement of the length of an individual's arms from fingertip to fingertip. A man's arm span is approximately normally distributed with mean of 70 inches with a standard deviation of 4.5 inches.
- Compute the probability that a randomly selected man has an arm span below 65 inches.
 - Compute the probability that a randomly selected man has an arm span between 60 and 72 inches.
 - Compute the length in inches of the 99th percentile for a man's arm span.

38. The size of fish is very important to commercial fishing. A study conducted in 2012 found the length of Atlantic cod caught in nets in Karlskrona to have a mean of 49.9 cm and a standard deviation of 3.74 cm (Ovegard, Berndt & Lunneryd, 2012). Assume the length of fish is normally distributed.
- Compute the probability that a cod is longer than 55 cm.
 - What is the length in cm of the longest 15% of Atlantic cod in this area?
39. A dishwasher has a mean life of 12 years with an estimated standard deviation of 1.25 years ("Appliance life expectancy," 2013). Assume the life of a dishwasher is normally distributed.
- Compute the probability that a dishwasher will last less than 10 years.
 - Compute the probability that a dishwasher will last between 8 and 10 years.
 - Compute the number of years that the bottom 25% of dishwashers would last.
40. The price of a computer is normally distributed with a mean of \$1400 and a standard deviation of \$60.
- What is the probability that a buyer paid less than \$1220?
 - What is the probability that a buyer paid between \$1400 and \$1580?
 - What is the probability that a buyer paid more than \$1520?
 - What is the probability that a buyer paid between \$1340 and \$1460?
 - What is the probability that a buyer paid between \$1400 and \$1520?
 - What is the probability that a buyer paid between \$1400 and \$1460?
41. Heights of 10-year-old children, regardless of sex, closely follow a normal distribution with mean 55.7 inches and standard deviation 6.8 inches.
- Compute the probability that a randomly chosen 10-year-old child is less than 50.4 inches.
 - Compute the probability that a randomly chosen 10-year-old child is more than 59.2 inches.
 - What proportion of 10-year-old children are between 50.4 and 61.5 inches tall?
 - Compute the 85th percentile for 10-year-old children.
42. The mean yearly rainfall in Sydney, Australia, is about 137 mm and the standard deviation is about 69 mm ("Annual maximums of," 2013). Assume rainfall is normally distributed. How many yearly mm of rainfall would there be in the top 25%?
43. The mean daily milk production of a herd of cows is assumed to be normally distributed with a mean of 33 liters, and standard deviation of 10.3 liters. Compute the probability that daily production is more than 40.9 liters?
44. The amount of time to complete a physical activity in a PE class is normally distributed with a mean of 33.2 seconds and a standard deviation of 5.8 seconds. Round answers to 4 decimal places.
- What is the probability that a randomly chosen student completes the activity in less than 28.9 seconds?
 - What is the probability that a randomly chosen student completes the activity in more than 37.2 seconds?
 - What proportion of students take between 28.5 and 37.3 seconds to complete the activity?
 - 70% of all students finish the activity in less than _____ seconds.
45. A study was conducted on students from a particular high school over the last 8 years. The following information was found regarding standardized tests used for college admittance. Scores on the SAT test are normally distributed with a mean of 1023 and a standard deviation of 204. Scores on the American College Testing (ACT) test are normally distributed with a mean of 19.3 and a standard deviation of 5.2. It is assumed that the two tests measure the same aptitude, but use different scales.
- Compute the SAT score that is the 50-percentile.
 - Compute the ACT score that is the 50-percentile.

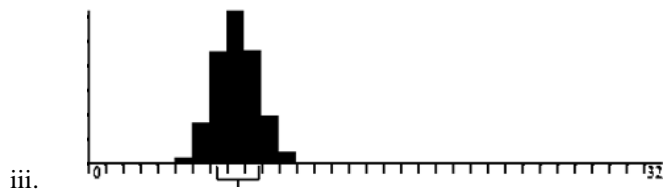
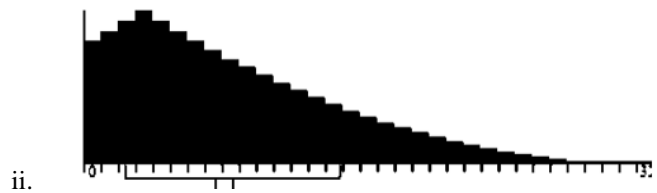
- c) If a student gets an SAT score of 1288, find their equivalent ACT score. Go out at least 5 decimal places between steps.
46. The length of a human pregnancy is normally distributed with a mean of 272 days with a standard deviation of 9 days (Bhat & Kushtagi, 2006).
- Compute the probability that a pregnancy lasts longer than 281 days.
 - Compute the probability that a pregnancy lasts less than 250 days.
 - How many days would a pregnancy last for the shortest 20%?
47. The MAX light rail in Portland, OR has a waiting time that is uniformly distributed with a mean waiting time of 5 minutes with a standard deviation of 2.9 minutes. A random sample of 40 wait times was selected. What is the probability the sample mean wait time is under 4 minutes?
48. The average credit card debt back in 2016 was \$16,061 with a standard deviation of \$4100. What is the probability that a sample of 35 people owe a mean of more than \$18,000?
49. A certain brand of electric bulbs has an average life of 300 hours with a standard deviation of 45. A random sample of 100 bulbs is tested. What is the probability that the sample mean will be less than 295?
50. Assume that the birth weights of babies are normally distributed with a mean of 3363 grams and a standard deviation of 563 grams.
- Compute the probability that a randomly selected baby weighs between 3200 grams and 3600 grams.
 - Compute the probability that the average weight of 30 randomly selected babies is between 3200 grams and 3600 grams.
 - Why did the probability increase?
51. If the Central Limit Theorem is applicable, this means that the sampling distribution of a _____ population can be treated as normal since the _____ is _____.
- symmetrical; variance; large
 - positively skewed; sample size; small
 - negatively skewed; standard deviation; large
 - non-normal; mean; large
 - negatively skewed; sample size; large
52. Delivery times for shipments from a central warehouse are exponentially distributed with a mean of 1.73 days (note that times are measured continuously, not just in number of days). The standard deviation of an exponential distribution is equal to the mean of 1.73 days. A random sample of 108 shipments are selected and their shipping times are observed. Use the Central Limit Theorem to find the probability that the mean shipping time for the 108 shipments is less than 1.53 days.
53. Match the following 3 graphs with the distribution of the population, the distribution of the sample, and the sampling distribution.
- Distribution of the Population
 - Distribution of the Sample
 - Sampling Distribution





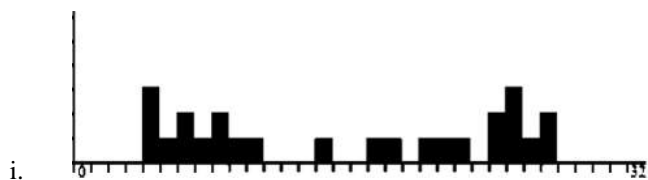
54. Match the following 3 graphs with the distribution of the population, the distribution of the sample, and the sampling distribution.

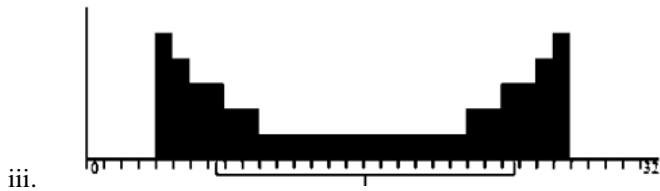
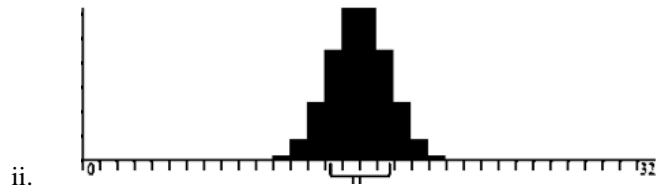
- a) Distribution of the Population
- b) Distribution of the Sample
- c) Sampling Distribution



55. Match the following 3 graphs with the distribution of the population, the distribution of the sample, and the sampling distribution.

- a) Distribution of the Population
- b) Distribution of the Sample
- c) Sampling Distribution





56. For a t-distribution, find the following probabilities.

- a) $P(t > -3.06), df = 35$
- b) $P(-2.17 < t < 2.17), df = 15$
- c) $P(t < 1.58), n = 8$
- d) $P(t \geq 1.69), n = 14$

57. For a t-distribution, find the following probabilities.

- a) $P(t \leq -2.83), n = 42$
- b) $P(t > 1.587), n = 11$
- c) $P(-1.833 \leq t \leq 1.833), df = 25$
- d) $P(t < 2), df = 18$

58. Use the area under the t-distribution for the following.

- a) Compute the area to the left of $t = -1.07$, when $df = 30$.
- b) Compute the area to the left of $t = -0.42$, when $n = 20$.
- c) Compute the area to the right of $t = 2.081$, when $df = 15$.
- d) Compute the area to the right of $t = 3.462$, when $n = 35$.

59. Use the area under the t-distribution for the following.

- a) Compute the area to the left of $t = 2.8563$, when $n = 12$.
- b) Compute the area to the left of $t = -1.8709$, when $df = 16$.
- c) Compute the area to the right of $t = 3.0173$, when $n = 30$.
- d) Compute the area to the right of $t = -1.4327$, when $df = 10$.

60. Compute the t-scores that give the middle 90% of the t-distribution for a sample size of 14.

61. Compute the t-scores that give the middle 99% of the t-distribution for $df = 28$.

62. Using a t-distribution with $df = 25$, find the $P(t \geq 2.185)$.

Chapter 6

Confidence Intervals for One Population



- 6.1 Introduction to Interval Estimates
- 6.2 Confidence Interval for a Proportion
- 6.3 Confidence Interval for a Mean
- 6.4 Interpreting a Confidence Interval
- 6.5 Determining Sample Size

6.1 Introduction to Interval Estimates

Confidence intervals are an essential tool in statistics for estimating population parameters based on sample data. Essentially, a confidence interval is a range of values that is likely to contain the true population parameter with a certain degree of confidence. For example, a 95% confidence interval for a population proportion is a range of values that we can be 95% confident contains the true population proportion. Confidence intervals are widely used in many fields, including business, medicine, and social sciences. Understanding how to calculate and interpret confidence intervals is crucial for making informed decisions based on statistical data.

Statistical inference is a method used to draw conclusions about a population based on a sample. We do this by using probability distributions and the Central Limit Theorem to understand what is happening in the population. However, measuring the entire population is often difficult, so we take a sample and use descriptive statistics from the sample to infer back to what is happening in our population.

Although there are many types of statistical inference tools, we will only cover confidence intervals for proportions and means in this text.

In inferential statistics, it is important to distinguish between a population parameter and a sample statistic. We frequently use a representative sample to generalize a population.

A **statistic** is any characteristic or measure from a sample, such as the sample mean \bar{x} . On the other hand, a **parameter** is any characteristic or measure from a population, such as the population mean μ . A **point estimate** for a parameter is a statistic. For example, the point estimate for the population mean μ is the sample mean \bar{x} . The point estimate for the population standard deviation σ is the sample standard deviation s , and so on.

A $100(1 - \alpha)\%$ **confidence interval** for a population parameter (μ , σ , etc.) represents that the proportion $100(1 - \alpha)\%$ of times that the true value of the population parameter is contained within the interval. The **confidence level** (or level of confidence) is $1 - \alpha$. Common percentages used for confidence interval levels are 90%, 95%, and 99%. Some corresponding confidence values of alpha are: 90% would be $\alpha = 0.10 = 10\%$, 95% would be $\alpha = 0.05 = 5\%$, and 99% would be $\alpha = 0.01 = 1\%$. In this context, α , "alpha," represents the complement of the confidence level, and its definition will be explained in the next chapter.

When a symmetric distribution, such as a normal distribution, is used, confidence intervals are always of the form: point estimate \pm margin of error. The **margin of error** defines the "radius" of the interval necessary to obtain the desired confidence level. The margin of error depends on the desired confidence level. Higher levels of confidence come at a cost, namely larger margins of error, which means our estimate will be less accurate.

The margin of error formula will usually include a value from a sampling distribution called the **critical value**. The critical value measures the number of standard errors to be added and subtracted in order to achieve your desired confidence level based on the α level chosen.

For example, if we use the standard normal distribution values that would give the middle 95% of the distribution when $\alpha = 0.05$ since $100(1 - 0.05)\% = 95\%$. The two critical values $-z_{\alpha/2}$ and $+z_{\alpha/2}$, as shown Figure 6-1. Note: in the notation $z_{\alpha/2}$, the $\alpha/2$ represents the area in each of the tails.

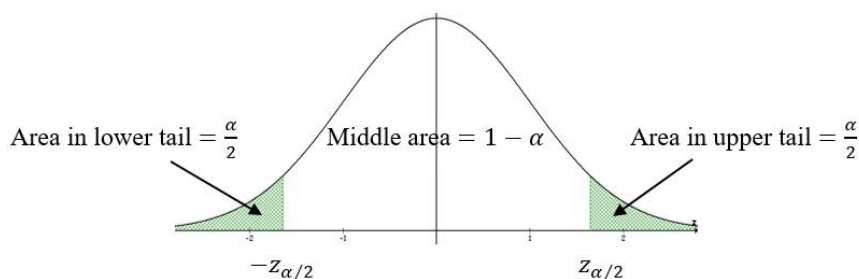


Figure 6-1

Example 6-1: Use Excel or your calculator to find the critical values $z_{\alpha/2}$ for a 95% confidence interval.

Solution: We are trying to find the two z -scores that give a lower tail area of 2.5%, notated as $z_{\alpha/2} = z_{0.025}$ and the upper tail area of 2.5% (lower tail of 97.5%) notated as $z_{1-\alpha/2} = z_{0.975}$ values. In Excel use =NORM.INV(lower tail area, mean, standard deviation). It is easier to deal with the positive z -score so use the z to the right of the mean, which would have $1 - \alpha/2 = 0.975$ area. In Excel use =NORM.INV(0.975,0,1) or the calculator invNorm(0.975,0,1) which gives $z_{1-\alpha/2} = 1.96$. For the lower value, use symmetry to get $z_{\alpha/2} = -1.96$ or use the previous function with area 0.025. The critical values $z = \pm 1.96$ are shown in Figure 6-2.

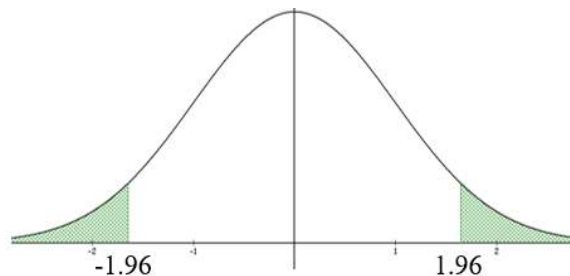


Figure 6-2

6.2 Confidence Interval for a Proportion

Suppose you want to estimate the population proportion, p . As an example, an administrator may want to know what proportion of students at your school smoke. An insurance company may want to know what proportion of accidents are caused by teenage drivers who do not have a drivers' education class. Every time we collect data from a new sample, we would expect the estimate of the proportion to change slightly. If you were to find a range of values over an interval this would give a better estimate of where the population proportion falls. This range of values that would better predict the true population parameter is called an interval estimate or confidence interval.

The sample proportion \hat{p} is the point estimate for p . The standard error (the standard deviation of the sampling distribution) of \hat{p} is $\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$. The $z_{\alpha/2}$ is the critical value using the standard normal distribution, and the margin of error $E = z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$. Some textbooks use π instead of p for the population proportion, and \bar{p} (pronounced "p-bar") or p^* instead of \hat{p} for sample proportion.

Choose a simple random sample of size n from a population having *unknown* population proportion p .

The $100(1 - \alpha)\%$ confidence interval estimate for p is given by $\hat{p} \pm z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$.

Where $\hat{p} = \frac{x}{n} = \frac{\text{\# of successes}}{\text{\# of trials}}$ (read as "p hat") is the sample proportion, and $\hat{q} = 1 - \hat{p} = \frac{\text{\# of failures}}{\text{\# of trials}}$.

The above confidence interval can be expressed as an inequality or an interval of values.

$$\hat{p} - z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)} < p < \hat{p} + z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)} \quad \text{or} \quad \left(\hat{p} - z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}, \hat{p} + z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}\right)$$

Assumption: $n \cdot \hat{p} \geq 10$ and $n \cdot \hat{q} \geq 10$.

*This assumption **must be** addressed before using these statistical inferences.

This formula is derived from the normal approximation of the binomial distribution, therefore the same conditions for a binomial need to be met, namely a set sample size of independent trials, two outcomes that have the same probability for each trial.

Steps for Calculating a Confidence Interval

1. State the random variable and the parameter in words.
 x = number of successes

- p = proportion of successes
- State and check the assumptions for confidence interval.
 - A simple random sample of size n is taken.
 - The conditions for the binomial distribution are satisfied.
 - To determine the sampling distribution of \hat{p} , you need to show that $n \cdot \hat{p} \geq 10$ and $n \cdot \hat{q} \geq 10$, where $\hat{q} = 1 - \hat{p}$. If this requirement is true, then the sampling distribution of \hat{p} is well approximated by a normal curve. (In reality, this is not really true, since the correct assumption deals with p . However, in a confidence interval you do not know p , so you must use \hat{p} . This means you just need to show that $x \geq 10$ and $n - x \geq 10$.)

- Compute the sample statistic $\hat{p} = \frac{x}{n}$ and the confidence interval $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$.

- Statistical Interpretation: In general, this looks like:

“We can be $(1 - \alpha) \cdot 100\%$ confident that the interval $\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)} < p < \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$ contains the true proportion.”

Real World Interpretation: Restate using the context from the question.

Example 6-2: A concern was raised in Australia that the percentage of deaths of indigenous Australian prisoners was higher than the percent of deaths of nonindigenous Australian prisoners, which is 0.27%. A sample of six years (1990-1995) of data was collected, and it was found that out of 14,495 indigenous Australian prisoners, 51 died (“Indigenous deaths in,” 1996). Find a 95% confidence interval for the proportion of indigenous Australian prisoners who died.

Solution:

- State the random variable and the parameter in words.

x = number of indigenous Australian prisoners who die
 p = proportion of indigenous Australian prisoners who die
- State and check the assumptions for a confidence interval.
 - A simple random sample of 14,495 indigenous Australian prisoners was taken. However, the sample was not a random sample, since it was data from six years. It is the numbers for all prisoners in these six years, but the six years were not picked at random. Unless there was something special about the six years that were chosen, the sample is probably a representative sample. This assumption is probably met.
 - There are 14,495 prisoners in this case. The prisoners are all indigenous Australians, so you are not mixing indigenous Australian with nonindigenous Australian prisoners. There are only two outcomes, the prisoner either dies or does not. The chance that one prisoner dies over another may not be constant, but if you consider all prisoners the same, then it may be close to the same probability. Thus, the assumptions for the binomial distribution are satisfied.
 - In this case, $x = 51$ and $n - x = 14,495 - 51 = 14,444$. Both are greater than or equal to 10. The sampling distribution for \hat{p} is a normal distribution.

- Compute the sample statistic and the confidence interval.

Sample Proportion: $\hat{p} = \frac{x}{n} = \frac{51}{14495} = .003518$,

Critical Value: $z_{\alpha/2} = 1.96$, since 95% confidence level

Margin of Error $E = z_{\alpha/2} \sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)} = 1.96 \sqrt{\left(\frac{0.003518(1-0.003518)}{14495}\right)} = 0.000964$

Confidence Interval: $\hat{p} - E < p < \hat{p} + E$

$0.003518 - 0.000964 < p < 0.003518 + 0.000964$

$0.002554 < p < 0.004482$ or $(0.002554, 0.004482)$

- Statistical Interpretation: We can be 95% confident that $0.002554 < p < 0.004482$ contains the proportion of all indigenous Australian prisoners who died.

Real World Interpretation: We can be 95% confident that the percentage of all indigenous Australian prisoners who died is between 0.26% and 0.45%.

Using Technology

Excel has no built-in shortcut key for finding a confidence interval for a proportion, but if you type in the following formulas shown below you can make your own Excel calculator where you just change the highlighted cells and all the numbers below will update with the relevant information.

Type in the following, be cognizant of cell reference numbers.

	A	B
1	Confidence Interval for a Proportion	
2	x	51
3	n	14495
4	C-Level	0.95
5	z	=-NORM.S.INV((1-B4)/2)
6	p-bar	=B2/B3
7	q-bar	=1-B6
8	E	=B5*SQRT(B6*B7/B3)
9	Lower boundary	=B6-B8
10	Upper boundary	=B6+B8

You get the following answers where the last two numbers are your confidence interval limits.

Confidence Interval for a Proportion	
x	51
n	14495
C-Level	0.95
z	1.9600
p-bar	0.003518
q-bar	0.996482
E	0.000964
Lower boundary	0.002555
Upper boundary	0.004482

Make sure to put your answer in interval notation (0.002555, 0.004482) or $0.26% < p < 0.45%$.

You can also do the calculations for the confidence interval with the TI Calculator.

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the [A:1-PropZInterval] option and press the [ENTER] key. Then type in the values for x , sample size and confidence level, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the answer in interval notation. Note: Sometimes you are not given the x value but a percentage instead. To find the x to use in the calculator, multiply \hat{p} by the sample size and round off to the nearest integer. The calculator will give you an error message if you put in a decimal for x or n . For example, if $\hat{p} = 0.22$ and $n = 124$ then $0.22 \cdot 124 = 27.28$, so use $x = 27$.



TI-89: Go to the [Apps] Stat/List Editor, then press [2nd] then F7 [Ints], then select **5: 1-PropZInt**. Type in the values for x , sample size and confidence level, and press the [ENTER] key. The calculator returns the answer in

interval notation. Note: sometimes you are not given the x value but a percentage instead. To find the x value to use in the calculator, multiply \hat{p} by the sample size and round off to the nearest integer. The calculator will give you an error message if you put in a decimal for x or n . For example, if $\hat{p} = 0.22$ and $n = 124$ then $0.22 \cdot 124 = 27.28$, so use $x = 27$.

Example 6-3: A researcher studying the effects of income levels on new mothers breastfeeding their infants hypothesizes that those countries where the income level is lower has a higher rate of infants breastfeeding than higher income countries. It is known that in Germany, considered a high-income country by the World Bank, 22% of all babies are breastfed. In Tajikistan, considered a low-income country by the World Bank, researchers found that in a random sample of 500 new mothers that 125 were breastfeeding their infants. Compute and interpret the 90% confidence interval of the proportion of mothers in low-income countries who breastfeed their infants.

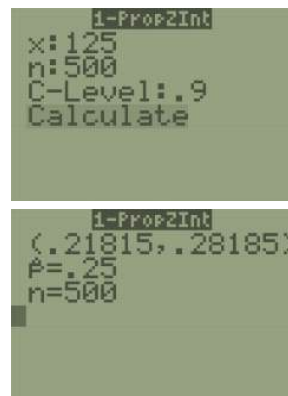
Solution:

1. State the random variable and the parameter in words.
 - x = The number of new mothers who breastfeed in a low-income country.
 - p = The proportion of new mothers who breastfeed in a low-income country.
2. State and check the assumptions for a confidence interval.
 - a. A simple random sample of 500 breastfeeding habits of new mothers in a low-income country was taken as was stated in the problem.
 - b. There were 500 women in the study. The women are considered identical, though they probably have some differences. There are only two outcomes - either the woman breastfeeds her baby or she does not. The probability of a woman breastfeeding her baby is probably not the same for each woman, but it is probably not that different for each woman. The assumptions for the binomial distribution are satisfied.
 - c. $x = 125$ and $n - x = 500 - 125 = 375$ and both are greater than or equal to 10, so the sampling distribution of \hat{p} is well approximated by a normal curve.

3. Compute the sample statistic and the confidence interval.
On the TI-83/84: Go into the STAT menu. Move the cursor over to TESTS and choose 1-PropZInt, then press Calculate.

4. Statistical Interpretation: We are 90% confident that the interval $0.219 < p < 0.282$ contains the population proportion of all women in low-income countries who breastfeed their infants.

Real World Interpretation: The proportion of women in low-income countries who breastfeed their infants is between 0.219 and 0.282 with 90% confidence.



Example 6-4: A local county has a very active adult education venue. A random sample of the population showed that 189 out of 400 persons 16 years old or older participated in some type of formal adult education activities, such as basic skills training, apprenticeships, personal interest courses, and part-time college or university degree programs. Estimate the true proportion of adults participating in some kind of formal education program with 98% confidence.

Solution: First, we are looking at the proportion of adults that participate in some kind of formal education program. Second, check the assumptions. Then find the z -score for a 98% confidence interval. One tail area would be $1 - 0.98 = 0.02$, $0.02/2 = 0.01$. The area to the left of the z -score would then be $1 - 0.01 = 0.99$. Use Excel =NORM.INV(0.99,0,1) or your calculator invNorm(0.99,0,1) which gives a z -score of 2.3263.

Compute $\hat{p} = \frac{x}{n} = \frac{189}{400} = 0.4725$, $\hat{q} = 1 - \hat{p} = 1 - 0.4725 = 0.5275$.

Substitute into the formula: $\hat{p} \pm z_{\alpha/2} \sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)} \Rightarrow 0.4725 \pm 2.3263 \sqrt{\left(\frac{0.4725 \cdot 0.5275}{400}\right)} \Rightarrow 0.4725 \pm 0.0582$

The answer can be given as an inequality $0.4143 < p < 0.5307$ or in interval notation (0.4143, 0.5307).

We can be 98% confident that the proportion of all adults participating in some kind of formal education program is between 41% and 53%.

6.3 Confidence Interval for a Mean

Suppose you want to estimate the mean weight of newborn infants, or you want to estimate the mean salary of college graduates. A confidence interval for the mean would be a way to estimate these means.

Choose a simple random sample of size n from a population having *unknown* mean μ . The $100(1 - \alpha)\%$ confidence interval estimate for μ is given by, $\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}}\right)$. The $df = \text{degrees of freedom}^*$ are $n - 1$. If the sample size is small ($n < 30$), the population we are sampling from must be normally distributed.

The confidence interval can be expressed as an inequality or an interval of values.

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \text{or} \quad \left(\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

*The degrees of freedom (df) are the number of values that are free to vary after a sample statistic has been computed. For example, if you know the mean was 50 for a sample size of 4, you could pick any 3 numbers you like, but the 4th value would have to be fixed to have the mean come out to be 50. For this class we just need to know that degrees of freedom will be based on the sample size.

The sample mean \bar{x} is the point estimate for μ , and the margin of error is $t_{\alpha/2} \left(\frac{s}{\sqrt{n}}\right)$. Where $t_{\alpha/2}$ is the positive critical value on the t-distribution curve with $df = n - 1$ and area $1 - \alpha$ between the critical values $-t_{\alpha/2}$ and $+t_{\alpha/2}$, as shown in Figure 6-3. Note: In the notation, $t_{\alpha/2}$ the $\alpha/2$ represents the area in each of the tails, see Figure 6-3.

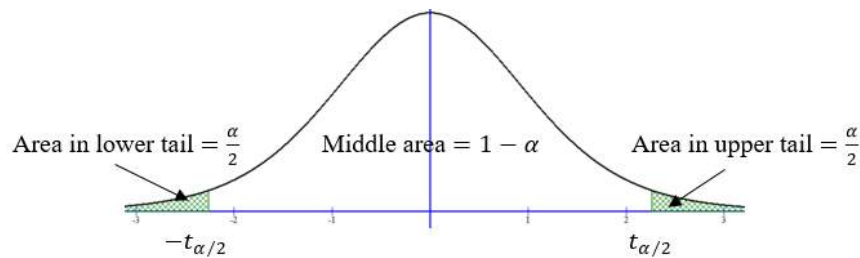


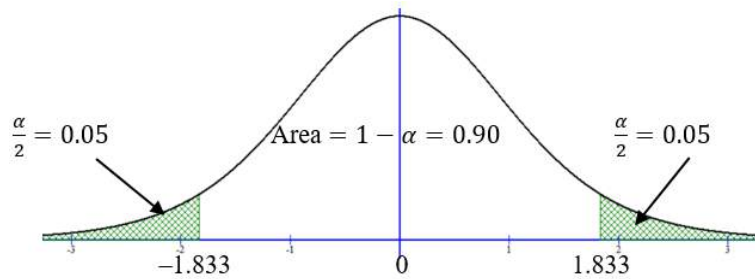
Figure 6-3

Note that we rarely have a calculation for the population standard deviation so in most cases we use the sample standard deviation as an estimate for the population standard deviation. If we have a normally distributed population with an unknown population standard deviation then the sampling distribution of the sample mean will follow a t-distribution. Assumption: If the sample size is small ($n < 30$), the population we are sampling from must be normal.

Before we compute a t-interval we will practice getting t critical values using Excel and the TI calculator's built in t-distribution. Refer back to the previous chapter for more examples.

Example 6-5: Compute the critical values $-t_{\alpha/2}$ and $+t_{\alpha/2}$ for a 90% confidence interval with a sample size of 10.

Solution: Draw a t-distribution with $df = n - 1 = 9$, see Figure 6-4. In Excel use $=T.INV(\text{lower tail area}, df)$ $=T.INV(0.95,9)$ or in the TI calculator use $\text{invT}(\text{lower tail area}, df) = \text{invT}(0.95,9)$. The critical values are $t = \pm 1.833$.



```
invT(.95,9)
1.833112923
```

Figure 6-4

We can use Excel to find the margin of error when raw data is given in a problem. The following example is first done longhand, then the T-Interval shortcut key on the TI calculator, and then using Excel's Data Analysis Tool.

Example 6-6: The yearly salary for mathematics assistant professors is normally distributed. A random sample of 8 math assistant professor's salaries are listed below in thousands of dollars. Estimate the population mean salary with a 99% confidence interval.

66.0 75.8 70.9 73.9 63.4 68.5 73.3 65.9

Solution: First find the t critical value using $df = n - 1 = 7$ and 99% confidence. Calculate the area below the critical value $(1 - 0.99)/2 = 0.995$. Calculate the critical value $t_{\alpha/2} = \text{invT}(.995,7) = 3.4995$. Then use technology to find the sample mean and sample standard deviation and substitute the numbers into the formula.

```
invT(.995,7)
3.499483292

t-Var Stats
x̄=69.7125
Σx=557.7
Σx²=39017.17
Sx=4.448254072
σx=4.160960676
n=8
```

$$\begin{aligned} \bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) &\Rightarrow 69.7125 \pm 3.4995 \left(\frac{4.4483}{\sqrt{8}} \right) \Rightarrow 69.7125 \pm 5.5037 \\ &\Rightarrow (64.2088, 75.2162) \end{aligned}$$

The answer can be given as an inequality $64.2088 < \mu < 75.2162$ or in interval notation $(64.2088, 75.2162)$.

We are 99% confident that the interval 64.2 and 75.2 contains the true population mean salary for all mathematics assistant professors.

We are 99% confident that the mean salary for mathematics assistant professors is between \$64,208.80 and \$75,216.20.

Assumption: The population we are sampling from must be normal* or approximately normal, and the population standard deviation σ is unknown. *This assumption **must be** addressed before using statistical inference for sample sizes of under 30.

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the [8:TInterval] option and press the [ENTER] key. Arrow over to the [Stats] menu and press the [ENTER] key. Then type in the mean, sample standard deviation, sample size and confidence level, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the answer in interval notation. Be careful, if you accidentally use the [7:ZInterval] option you would get the wrong answer.

Alternatively (If you have raw data in list one) Arrow over to the [Data] menu and press the [ENTER] key. Then type in the list name, L1, leave Freq:1 alone, enter the confidence level, arrow down to [Calculate] and press the [ENTER] key.

```
TInterval
Inpt: Data Stats
List: L1
Freq: 1
C-Level: .99
Calculate

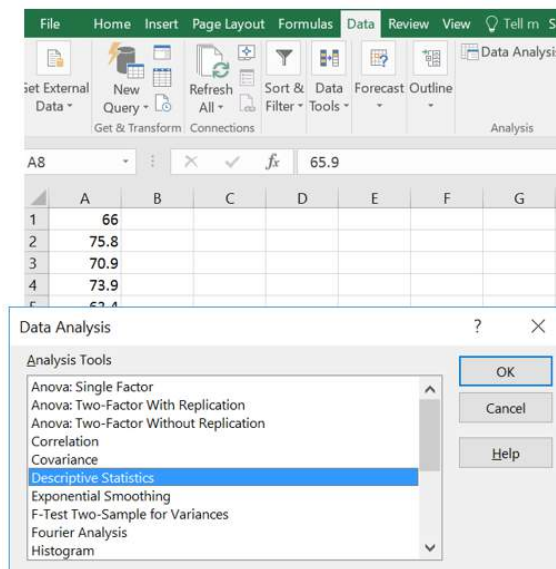
TInterval
(64.209, 75.216)
x̄=69.7125
Sx=4.448254072
n=8
```

TI-89: Go to the [Apps] **Stat/List Editor**, then press [2nd] then F7 [Ints], then select **2:TIInterval**. Choose the input method, data is when you have entered **data** into a list previously or **stats** when you are given the mean and standard deviation already. Type in the mean, standard deviation, sample size (or list name (list1), and Freq: 1) and confidence level, and press the [ENTER] key. The calculator returns the answer in interval notation. Be careful: If you accidentally use the [1:ZInterval] option you would get the wrong answer.

Excel Directions

Type the data into Excel. Select the Data Analysis Tool under the Data tab.

	A
1	66
2	75.8
3	70.9
4	73.9
5	63.4
6	68.5
7	73.3
8	65.9



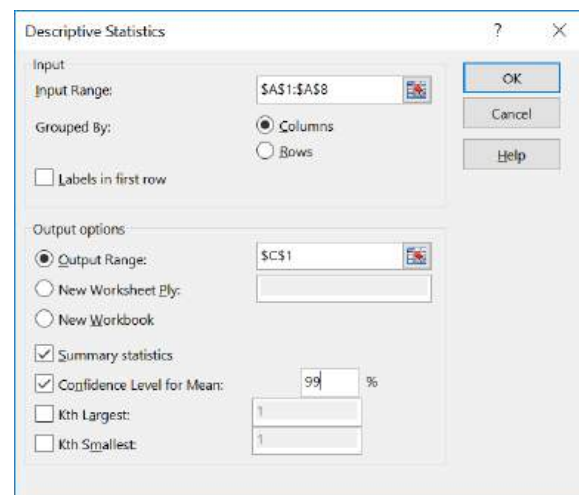
Select Descriptive Statistics. Select OK.

Use your mouse and click into the Input Range box, then select the cells containing the data. If you highlighted the label then check the box next to Labels in first row. In this case no label was typed in so the box is left blank. (Be very careful with this step. If you check the box and do not have a label then the first data point will become the label and all your descriptive statistics will be incorrect.)

Check the boxes next to Summary statistics and Confidence Level for Mean. Then change the confidence level to fit the question. Select OK.

The table output does not find the confidence interval. However, the output does give you the sample mean and margin of error.

The margin of error is the last entry labeled Confidence Level. To find the confidence interval subtract and add the margin of error to the sample mean to get the lower and upper limit of the interval in two separate cells.



The following screenshot shows the cell references to find the lower limit as =D3-D16 and the upper limit as =D3+D16. Make sure to put your answer in interval notation.

	A	B	C	D	E	F
1	66		Column1			
2	75.8					
3	70.9	Mean		69.7125		=D3-D16
4	73.9	Standard Error		1.572695309		=D3+D16
5	63.4	Median		69.7		
6	68.5	Mode		#N/A		
7	73.3	Standard Deviation		4.448254071		
8	65.9	Sample Variance		19.78696428		
9		Kurtosis		-1.53765685		
10		Skewness		-0.02822188		
11		Range		12.4		
12		Minimum		63.4		
13		Maximum		75.8		
14		Sum		557.7		
15		Count		8		
16		Confidence Level(99.0%)		5.503620966		

The answer is given as an inequality $64.2088 < \mu < 75.2162$ or in interval notation $(64.2088, 75.2162)$.

We are 99% confident that the interval 64.2 and 75.2 contains the true population mean salary for all mathematics assistant professors.

Example 6-7: The total of individual weights of garbage discarded by 19 households in one week is normally distributed with a mean of 32.1 lbs. with a sample standard deviation of 9.7 lbs. Find the 95% confidence interval of the mean.

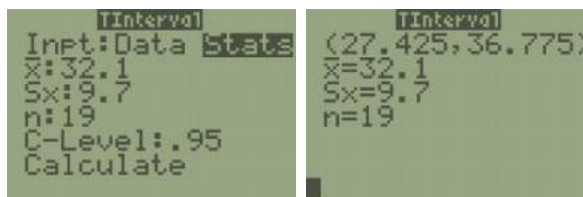
Solution: First calculate the t critical values for a 95% confidence level and $df = n - 1 = 19 - 1 = 18$. Use tail area $(1 - 0.95)/2 = 0.025$ in the invT function. In Excel use formula $=T.INV(0.025,18) = -2.100922$. Substitute in the values of $\bar{x} = 32.1$, $s = 9.7$, $n = 19$, and $t_{\alpha/2} = 2.100922$.

$$\bar{x} \pm t_{\alpha/2, n-1} \left(\frac{s}{\sqrt{n}} \right) \Rightarrow 32.1 \pm 2.100922 \left(\frac{9.7}{\sqrt{19}} \right) \Rightarrow 32.1 \pm 4.67525 \Rightarrow (27.4247, 36.7753)$$

The answer can be given as an inequality $27.4247 < \mu < 36.7753$ or in interval notation $(27.4247, 36.7753)$.

Excel does not have an option for summarized data, but you can use the TInterval shortcut in the TI calculators.

We are 95% confident that the true population mean weight of weekly household garbage is between 27.4247 and 36.7753 lbs.



Summary

A t-confidence interval is used to estimate an unknown value of the population mean for a single sample. We need to make sure that the population is normally distributed or the sample size is 30 or larger. Once this is verified, we use the interval $\bar{x} - t_{\alpha/2, n-1} \left(\frac{s}{\sqrt{n}} \right) < \mu < \bar{x} + t_{\alpha/2, n-1} \left(\frac{s}{\sqrt{n}} \right)$ to estimate the true population mean. It is important to interpret the confidence interval correctly. A general interpretation where you would change what is in the parentheses to fit the context of the problem is: "One can be $100(1 - \alpha)\%$ confident that between (lower boundary) and (upper boundary) contains the population mean of (random variable in words using context and units from problem)."

6.4 Interpreting a Confidence Interval

There is always a chance that the confidence interval would not contain the true parameter that we are looking for. Inferential statistics does not “prove” that the population parameter is within the boundaries of the confidence interval. If the sample we took had all outliers and the sample statistic is far away from the true population parameter, then when we subtract and add the margin of error to the point estimate, the population parameter may not be within the limits.

Both the sample size and confidence level affect how wide the interval is. The following discussion demonstrates what happens to the width of the interval as you get more confident.

Think about shooting an arrow into the target. Suppose you are really good at that and that you have a 90% chance of hitting the bull’s-eye. Now the bull’s-eye is very small. Since you hit the bull’s eye approximately 90% of the time, then you probably hit inside the next ring out 95% of the time. You have a better chance of doing this, but the circle is bigger. You probably have a 99% chance of hitting the target, but that is a much bigger circle to hit. As your confidence in hitting the target increases, the circle you hit gets bigger. The same is true for confidence intervals.

The higher level of confidence makes a wider interval. There is a tradeoff between width and confidence level. You can be really confident about your answer, but your answer will not be very precise. On the other hand, you can have a precise answer (small margin of error) but not be very confident about your answer.

When we increase the confidence level, the confidence interval becomes wider to be more confident that the population parameter is within the lower and upper boundaries. A wider margin of error means less accuracy. When one is more confident, one would have a harder time predicting the true parameter with the larger range of values. See Figure 6-5.

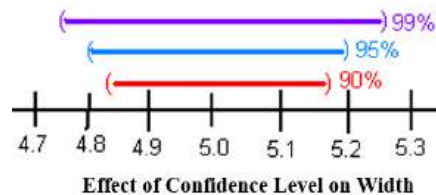


Figure 6-5

For instance, if we wanted to find the true mean grade for a statistics course using a 99% confidence critical value, we may get a very large margin of error, $75\% \pm 25\%$. This would say that we would be 99% confident that the average grade for all students is between 50% to 100%. This is of little help since that is anywhere between the grade range of F to an A. There are two ways to narrow this margin of error. The best way to reduce the margin of error is to increase the sample size, which decreases the standard deviation of the sampling distribution. When you take a larger sample, you will get a narrower interval. The other way to decrease the margin of error is to decrease your confidence level. When you decrease the confidence level, the critical value will be smaller. If we have a smaller margin of error then one can more accurately predict the population parameter.

Now look at how the sample size affects the size of the interval. Suppose the following Figure 6-6 represents confidence intervals calculated on a 95% interval.

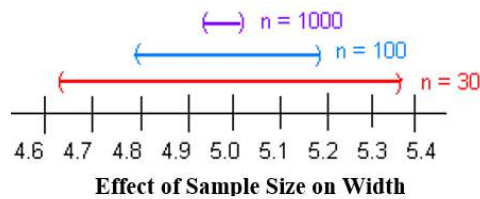


Figure 6-6

A larger sample size from a representative sample makes the standard error smaller and hence the width of the interval narrower. Large samples are closer to the true population so the point estimate is pretty close to the true value.

The probability that one confidence interval contains the mean is either zero or one. However, if we were to repeat the same sampling process, the proportion of times that the confidence intervals would capture the populations parameter is $(1 - \alpha)$, where α is the complement of the confidence level.

As an example, if you have a 95% confidence interval of $0.65 < p < 0.73$, then you would say, “If we were to repeat this process, then 95% of the time the interval 0.65 to 0.73 would contain the true population proportion.” This means that if you have 100 intervals, 95 of them will contain the true proportion, and 5% will not.

The incorrect interpretation is that there is a 95% probability that the true value of p will fall between 0.65 and 0.73. The reason that this interpretation is incorrect is that the true value is fixed out there somewhere. You are trying to capture it with this interval. This is the chance that your interval captures the true mean, and not that the true value falls in the interval.

In addition, a real-world interpretation depends on the situation. It is where you are telling people what numbers you found the parameter to lie between. Therefore, your real-world interpretation is where you tell what values your parameter is between. There is no probability attached to this statement. That probability is in the statistical interpretation.

The following website is an applet where you can simulate confidence intervals with different parameters, sample sizes and confidence levels, take a moment and play around with the applet:

<http://www.rossmanchance.com/applets/ConfSim.html>.

As an example, an instructor wants to estimate the mean score on a final exam for all calculus students at their school. The instructor takes a random sample of 20 final exam scores to estimate the mean score for all the students. The instructor calculates the confidence interval to be $65.032 < \mu < 74.573$.

Later the instructor is talking with the course coordinator and they decide to collect all the student exam scores from every class and found the population mean $\mu = 75$. The instructor’s confidence interval did not actually capture the true population mean. Why not? A confidence interval does not guarantee that the population parameter is contained within the interval. If we were to repeat the same sampling process so that we have a total of 100 intervals, we would expect that approximately $95/100 = 95\%$ of those intervals would contain the true population mean. This 95% is based off of the confidence level, but doesn’t guarantee that exactly 95% of those intervals contain the mean.

Using the applet linked above, the black vertical bar in Figure 6-7 represents the true population mean test score of 75 (which is usually unknown in real life). There are 95 confidence intervals that contain the population mean shown in green. There are 5 confidence intervals that did not capture the population mean within the interval endpoints that are shown in red. The third interval down from the top is red, represents the instructor’s interval $65.032 < \mu < 74.573$. Their sample of 20 exam scores had a sample mean of 69.803 with a standard deviation of 10.194. The sample mean was very low compared to the population mean, and hence when the margin of error was subtracted and added to the point estimate, the interval did not capture the true population mean.

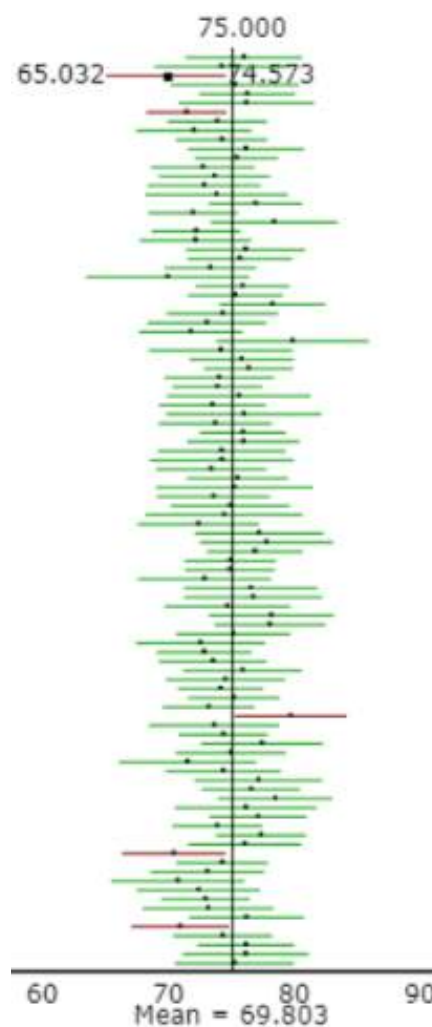


Figure 6-7

If you increase the sample size, you decrease the width of the confidence interval and hence can have a more accurate estimate, but you still have the same 5% chance that your interval does not contain the true population mean. You can change the confidence interval, which in turn will also change the width of the interval. If you want to be more confident that your interval contains the population parameter, then you need to have a larger net to capture that value.

Let's explore what happens to the width of the intervals when you change the sample sizes and the confidence level.

In the following Figure 6-8, confidence intervals were simulated using a 90% confidence level and then again using the 99% confidence level. Each confidence level was run 100 times with sample sizes of $n = 30$, then again using a sample size of $n = 100$, holding all other variables constant.

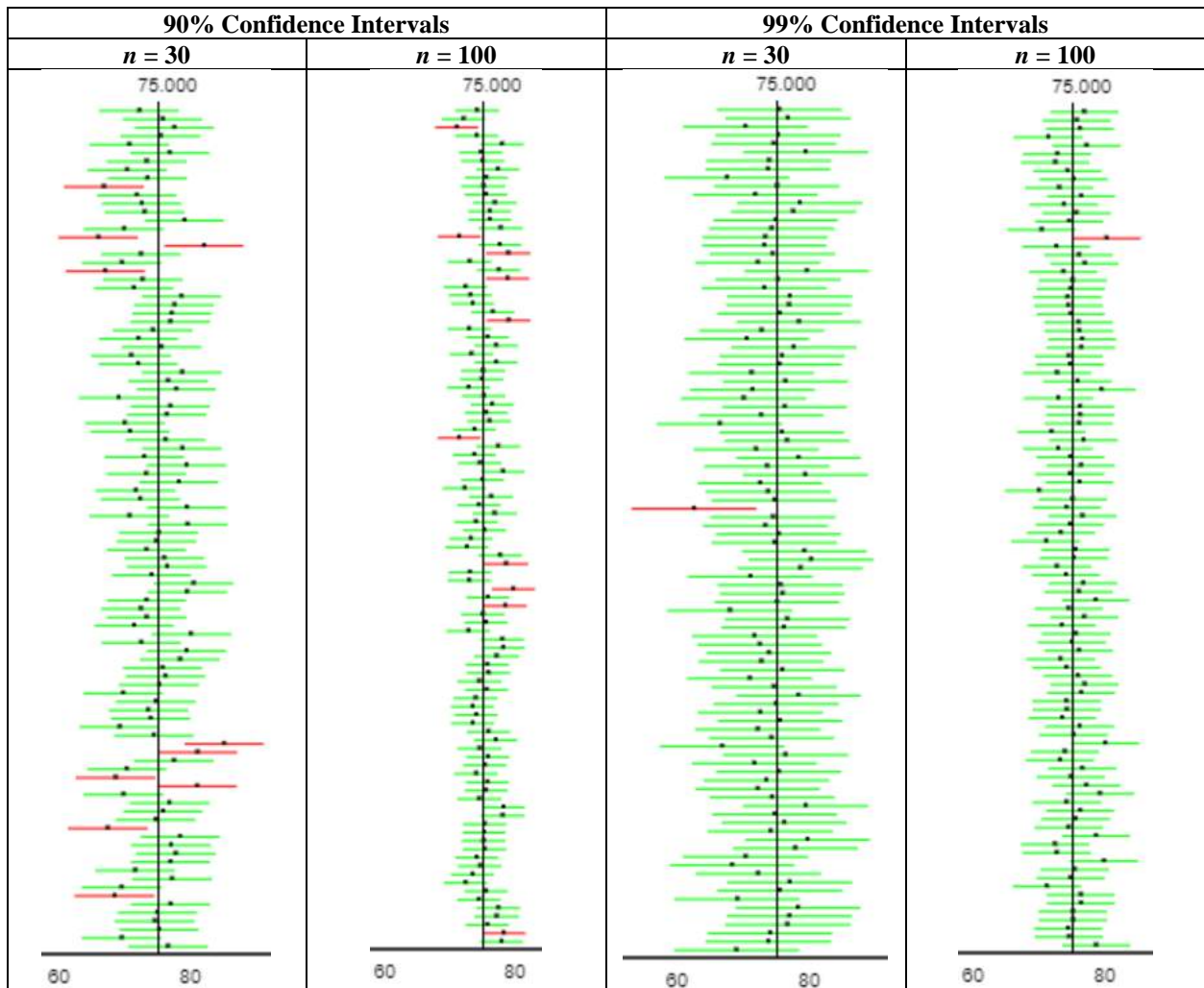


Figure 6-8

Compare columns 1 & 2 with columns 3 & 4 in Figure 6-8. For columns 1 & 2, $90/100 = 90\%$ of the confidence intervals contain the mean. For columns 3 & 4, $99/100 = 99\%$ of the confidence intervals contain the mean. Note the higher confidence level is wider for the same sample size.

Compare columns 1 & 3 in Figure 6-8 and you can see that the width of the confidence interval is wider for the 99% confidence level compared to the 90% confidence level. Holding all other variables constant the confidence interval captured the population mean 99% of the time. Then compare columns 2 & 4 to see similar results.

The wider confidence intervals will more likely capture the true population mean, however you will have less accuracy in predicting what the true mean is.

Example 6-8: State the statistical and real-world interpretations of the following confidence intervals.

- a) Suppose you have a 95% confidence interval for the mean age a woman gets married in 2023 is $26 < \mu < 28$.

Solution: Statistical Interpretation: We are 95% confident that the interval $26 < \mu < 28$ contains the population mean age of all women that got married in 2023.

Real World Interpretation: We are 95% confident that the mean age of women that got married in 2023 is between 26 and 28 years of age.

- b) Suppose a 99% confidence interval for the proportion of Americans who have tried cannabis as of 2019 is $0.55 < p < 0.61$.

Solution: Statistical Interpretation: We are 99% confident that the interval $0.55 < p < 0.61$ contains the population proportion of all Americans who have tried cannabis as of 2019.

Real World Interpretation: We are 99% confident that the proportion of all Americans who have tried cannabis as of 2019 is between 55% and 61%.

“I’m not trying to prove anything, by the way. I’m a scientist and I know what constitutes proof. But the reason I call myself by my childhood name is to remind myself that a scientist must also be absolutely like a child. If he sees a thing, he must say that he sees it, whether it was what he thought he was going to see or not. See first, think later, then test. But always see first. Otherwise you will only see what you were expecting. Most scientists forget that.”
(Adams, 2002)

6.5 Determining Sample Size

Calculating sample size is an essential part of statistics. In order to make accurate inferences about a population based on a sample, we need to ensure that our sample size is appropriate. One key factor to consider when determining sample size is the margin of error. The margin of error is the maximum amount by which our estimate may differ from the true population value, given a certain level of confidence. By understanding how to determine the appropriate sample size for a given study, we can improve the accuracy and reliability of our statistical analyses. Often, we need a specific confidence level, but we also need our margin of error to be within a set range. We are able to accomplish this by increasing the sample size. However, taking large samples is often difficult or costly to accomplish. Thus, it is useful to be able to determine the minimum sample size necessary to achieve our confidence interval.

Calculating a Sample Size to Estimate a Proportion

The following formula calculates the minimum sample size needed to construct a confidence interval for a proportion.

A confidence interval for a population proportion p and $q = 1 - p$, with specific margin of error E is given by:

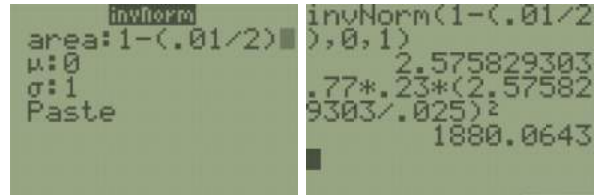
$$n = p^* \cdot q^* \left(\frac{z_{\alpha/2}}{E} \right)^2 \quad \text{Always round up the answer to the next whole number.}$$

Note: If the sample size is determined before the sample is selected, the p^* and q^* in the above equation are our best guesses. Often times statisticians will use $p^* = q^* = 0.5$; this takes the guesswork out of determining p^* and provides the “worst case scenario” for n . In other words, if $p^* = 0.5$ is used, then you are guaranteed that the margin of error

will not exceed E but you also will have to sample the largest possible sample size. Some texts will use p or π instead of p^* .

Example 6-9: You want to obtain a sample to estimate the proportion of a population that possess a particular genetic marker. Based on previous evidence, you believe approximately 77% of the population have the genetic marker. You would like to be 99% confident that your estimate is within 2.5% of the true population proportion. How large of a sample size is required?

Solution: Calculate the z critical value using 99% confidence. The area below the right hand z critical value is $1 - \alpha/2 = 1 - (0.01/2) = 0.995$. Use the invNorm function for a standard normal distribution $z_{\alpha/2} = 2.575829303$. The $p^* = 0.77$, $q^* = 1 - p^* = 1 - 0.77 = 0.23$. The margin of error is 2.5% so $E = 0.025$. Substitute into the formula:



$$n = p^* \cdot q^* \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.77 \cdot 0.23 \cdot \left(\frac{2.575829303}{0.025}\right)^2 = 1880.0643.$$

Since we cannot have 0.0643 people, we need to round up to the next whole person and use $n = 1881$. Do not round down since we may not get within our margin of error for a smaller sample size.

Note that if you round the z -score to only 2 decimal places you would get $n = 0.77 \cdot 0.23 \cdot \left(\frac{2.58}{0.025}\right)^2 = 1886.1575$ or $n = 1887$. That is 6 more people that you would need to sample by just rounding. Ideally you would round z to the same number of decimal places that would use to calculate the confidence interval. Most calculations assume that you do not round between steps and only round the final answer.

Example 6-10: A study found that 73% of prekindergarten children ages 3 to 5 whose mothers had a bachelor's degree or higher were enrolled in early childhood care and education programs.

a) How large a sample is needed to estimate the true proportion within 3% with 95% confidence?

Solution: Calculate the z critical value using 95% confidence. The area below the right hand z critical value is $1 - \alpha/2 = 1 - (0.05/2) = 0.975$. Use the invNorm function for a standard normal distribution $z_{\alpha/2} = 1.959963986$.

$$\text{Use } n = p^* \cdot q^* \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.73 \cdot 0.27 \cdot \left(\frac{1.959963986}{0.03}\right)^2 = 841.2795.$$

Round up to the next whole person and use $n = 842$. Do not round down since we may not get within our margin of error for a smaller sample size.

b) How large a sample is needed to estimate the true proportion within 3% with 99% confidence?

Solution: Calculate the z critical value using 99% confidence. The area below the right hand z critical value is $1 - \alpha/2 = 1 - (0.01/2) = 0.995$. Use the invNorm function for a standard normal distribution $z_{\alpha/2} = 2.575829303$.

$$\text{Use } n = p^* \cdot q^* \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.73 \cdot 0.27 \cdot \left(\frac{2.575829303}{0.03}\right)^2 = 1453.0424.$$

Round up to the next whole person and use $n = 1454$. Do not round down since we may not get within our margin of error for a smaller sample size. If we want to be more confident of containing the population parameter p , we would need to increase the sample size.

c) How large a sample is needed if you had no prior knowledge of the sample proportion using a 95% confidence?

Solution: Calculate the z critical value using 95% confidence. The area below the right hand z critical value is $1 - \alpha/2 = 1 - (0.05/2) = 0.975$. Use the invNorm function for a standard normal distribution $z_{\alpha/2} = 1.959963986$. Since no proportion is given, use the planning value of $p^* = 0.5$.

$$n = 0.5 \cdot 0.5 \cdot \left(\frac{1.959963986}{0.03}\right)^2 = 1067.0719 \quad \text{Round up and use } n = 1,068.$$

Note the sample sizes of 842 and 1,068. If you have a prior knowledge about the sample proportion then you may not have to sample as many people to get the same margin of error. The larger the sample size, the smaller the confidence interval.

Calculating a Sample Size to Estimate a Mean

The following formula calculates the minimum sample size needed to construct a confidence interval for a mean.

A confidence interval for a population mean μ with specific margin of error E and known population standard deviation σ is given by, $n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$.
Always round up the answer to the next whole number.

Keep in mind that we rarely know the value of the population standard deviation. We can estimate σ by using a previous year's standard deviation, a standard deviation from a similar study, a pilot sample, or by dividing the range by 4.

Example 6-11: A researcher is interested in estimating the average salary of teachers. She wants to be 95% confident that her estimate is correct. In a previous study, she found the population standard deviation was \$1,175. How large a sample is needed to be accurate within \$100?

Solution: First calculate the $z_{\alpha/2}$ for 95% confidence using Excel or your calculator, so $z_{\alpha/2} = 1.959963986$. Most of the time the margin of error = E follows the word “within” in the question, $E = 100$. The standard deviation $\sigma = 1175$. Replace each number into the formula: $n = \left(\frac{1.959963986 \cdot 1175}{100}\right)^2 = 530.36$.

If we round down, we would not get “within” the \$100 margin of error. Always round sample sizes up to the next whole number so that your margin of error will be within the specified amount. The larger the sample size, the smaller the confidence interval.

The answer is $n = 531$.

Example 6-12: The age when smokers first start from previous studies is normally distributed with a mean of 13 years old with a population standard deviation of 2.1 years old. A researcher wants to conduct a survey of smokers of this generation to estimate if the mean age has changed. The researcher wants to be 99% confident that their estimate is correct. What sample size should the researcher take to be within a margin of error of 0.5 years?

Solution: First calculate the $z_{\alpha/2}$ for 99% confidence using Excel or your calculator, so $z_{\alpha/2} = 2.575829303$. Most of the time the margin of error = E follows the word “within” in the question, $E = 0.5$. The standard deviation $\sigma = 2.1$. Replace each number into the formula: $n = \left(\frac{2.575829303 \cdot 2.1}{0.5}\right)^2 = 117.04$.

If we round down, we would not get “within” the 0.5 margin of error. Always round sample sizes up to the next whole number so that your margin of error will be within the specified amount. The larger the sample size, the smaller the confidence interval.

The answer is $n = 118$.

Chapter 6 Formulas

<p>Confidence Interval for One Proportion</p> $\hat{p} \pm z_{\alpha/2} \sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$ $\hat{p} = \frac{x}{n}$ $\hat{q} = 1 - \hat{p}$ <p>TI-Calculator: 1-PropZInt</p>	<p>t-Confidence Interval</p> $\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}}\right)$ $df = n - 1$ <p>TI-Calculator: TInterval</p>
<p>Z-Critical Values</p> <p>$\alpha = 1 - \text{confidence level}$</p> <p>Excel: $z_{\alpha/2} = \text{NORM.INV}(1 - (\alpha/2), 0, 1)$</p> <p>TI-Calculator: $z_{\alpha/2} = \text{invNorm}(1 - (\alpha/2), 0, 1)$</p>	<p>t-Critical Values</p> <p>$\alpha = 1 - \text{confidence level}$</p> <p>Excel: $t_{\alpha/2} = \text{T.INV}(1 - (\alpha/2), df)$</p> <p>TI-Calculator: $t_{\alpha/2} = \text{invT}(1 - (\alpha/2), df)$</p>
<p>Sample Size for Proportion</p> $n = p^* \cdot q^* \left(\frac{z_{\alpha/2}}{E}\right)^2$ <p>Always round up to whole number.</p> <p>If p is not given use $p^* = 0.5$.</p> <p>E = Margin of Error</p>	<p>Sample Size for Mean</p> $n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$ <p>Always round up to whole number.</p> <p>E = Margin of Error</p>

Chapter 6 Exercises

- Which confidence level would give the narrowest margin of error?
 - 80%
 - 90%
 - 95%
 - 99%
- Suppose you compute a confidence interval with a sample size of 25. What will happen to the width of the confidence interval if the sample size increases to 50, assuming everything else stays the same? Choose the correct answer below.
 - Gets smaller
 - Stays the same
 - Gets larger
- For a confidence level of 90% with a sample size of 35, find the critical z values.
- For a confidence level of 99% with a sample size of 18, find the critical z values.
- Which of the following would result in the widest confidence interval?
 - A sample size of 100 with 99% confidence.
 - A sample size of 100 with 95% confidence.
 - A sample size of 30 with 95% confidence.
 - A sample size of 30 with 99% confidence.
- Out of a sample of 200 adults ages 18 to 30, 54 still lived with their parents. Based on this, construct a 95% confidence interval for the true population proportion of adults ages 18 to 30 that still live with their parents.
- In a random sample of 200 people, 135 said that they watched educational TV. Compute and interpret the 95% confidence interval of the true proportion of people who watched educational TV.
- In a certain state, a survey of 600 workers showed that 35% belonged to a union. Compute and interpret the 95% confidence interval of true proportion of workers who belong to a union.
- A teacher wanted to estimate the proportion of students who take notes in her class. She used data from a random sample size of 82 and found that 50 of them took notes. The 99% confidence interval for the proportion of student that take notes is _____ $< p <$ _____.
- A random sample of 150 people was selected and 12% of them were left-handed. Compute and interpret the 90% confidence interval for the proportion of left-handed people.
- A 2019 survey of 2,000 adults, commissioned by the sleep-industry experts from Sleepopolis, reveals 34% sleep with a stuffed animal, blanket, or other anxiety-reducing item of sentimental value. Calculate the 90% confidence interval for the true proportion of adults that sleep with an anxiety-reducing item.
- A 2020 YouGov survey of 24,747 U.S. adults showed that 14,353 always washed their hands after using the toilet. Calculate the 99% confidence interval for the true proportion of U.S. adults that wash their hands after using the toilet.
- According to the 2023 survey of 30,300 U.S. households, 17% reported using a food pantry in the last year. Calculate the 95% confidence interval for the true proportion of U.S. households that used a food pantry in the last year.

14. A survey asked people if they were aware that maintaining a healthy weight could reduce the risk of stroke. A 95% confidence interval was found using the survey results to be (0.54, 0.62). Which of the following is the correct interpretation of this interval?
- We are 95% confident that the interval $0.54 < p < 0.62$ contains the population proportion of people who are aware that maintaining a healthy weight could reduce the risk of stroke.
 - There is a 95% chance that the sample proportion of people who are aware that maintaining a healthy weight could reduce the risk of stroke is between $0.54 < p < 0.62$.
 - There is a 95% chance of having a stroke if you do not maintain a healthy weight.
 - There is a 95% chance that the proportion of people who will have a stroke is between 54% and 62%.
15. Gallup tracks daily the percentage of Americans who approve or disapprove of the job Donald Trump is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults. Margin of error is ± 3 percentage points. On December 15, 2017, the gallop poll using a 95% confidence level showed that 34% approved of the job Donald Trump was doing. Which of the following is the correct statistical interpretation of the confidence interval?
- As of December 15, 2017, 34% of American adults approve of the job Donald Trump is doing as president.
 - We are 95% confident that the interval $0.31 < p < 0.37$ contains the proportion of American adults who approve of the job Donald Trump is doing as president as of December 15, 2017.
 - As of December 15, 2017, 95% of American adults approve of the job Donald Trump is doing as president.
 - We are 95% confident that the proportion of adult Americans who approve of the job Donald Trump is doing as president is 0.34 as of December 15, 2017.
16. Out of 500 people sampled in early October 2020, 315 preferred Biden. Based on this, compute the 95% confidence interval for the proportion of the voting population that preferred Biden.
17. For a confidence level of 90% with a sample size of 30, find the critical t values.
18. For a confidence level of 99% with a sample size of 24, find the critical t values.
19. For a confidence level of 95% with a sample size of 40, find the critical t values.
20. The amount of money in the money market accounts of 26 customers is found to be approximately normally distributed with a mean of \$18,240 and a sample standard deviation of \$1,100. Find and interpret the 95% confidence interval for the mean amount of money in the money market accounts at this bank.
21. A professor wants to estimate how long students stay connected during two-hour online lectures. From a random sample of 25 students, the mean stay time was 93 minutes with a standard deviation of 10 minutes. Assuming the population has a normal distribution, compute a 95% confidence interval estimate for the population mean.
22. A laboratory in Oregon is interested in finding the mean chloride level for a healthy resident in the state. A random sample of 25 healthy residents has a mean chloride level of 98 mEq/L and standard deviation of 17 mEq/L. If it is known that the chloride levels in healthy individuals residing in Oregon is normally distributed. Calculate and interpret the 95% confidence interval for the true mean chloride level of all healthy Oregon residents.
23. The age when smokers first start from previous studies is normally distributed with a mean of 13 years old. A survey of smokers of this generation was done to estimate if the mean age has changed. The sample of 33 smokers found that their mean starting age was 13.7 years old with a standard deviation of 2.1 years. Compute the 99% confidence interval of the mean.
24. The scores on an examination in biology are approximately normally distributed. The following is a random sample of scores from this year's examination: 403, 418, 460, 482, 511, 543, 576, 421. Compute and interpret the 99% confidence interval for the population mean scores.

25. A college advisor wants to estimate the undergraduate grade point average (GPA) for students admitted to the top graduate business schools. The advisor randomly samples 8 students admitted to the top schools and found their GPA was 3.53 with a standard deviation of 0.18. Assume that the population is normally distributed. Calculate and interpret the 99% confidence interval estimate of the mean undergraduate GPA for all students admitted to the top graduate business schools.
26. A random sample of stock prices per share (in dollars) is shown. Find and interpret the 90% confidence interval for the mean stock price. Assume the population of stock prices is normally distributed.

26.60 75.37 3.81 28.37 40.25 13.88 53.80 28.25 10.87 12.25

27. In a certain city, a random sample of executives have the following monthly personal incomes (in thousands) 35, 43, 29, 55, 63, 72, 28, 33, 36, 41, 42, 57, 38, 30. Assume the population of incomes is normally distributed. Find and interpret the 95% confidence interval for the mean income.
28. A tire manufacturer wants to estimate the average number of miles that may be driven in a tire of a certain type before the tire wears out. Assume the population is normally distributed. A random sample of tires is chosen and are driven until they wear out and the number of thousands of miles is recorded, find and interpret the 99% confidence interval for the mean using the sample data 32, 33, 28, 37, 29, 30, 22, 35, 23, 28, 30, 36.
29. Recorded here are the germination times (in days) for ten randomly chosen seeds of a new type of bean: 18, 12, 20, 17, 14, 15, 13, 11, 21, 17. Assume that the population germination time is normally distributed. Find and interpret the 99% confidence interval for the mean germination time.
30. A sample of the length in inches for newborns is given below. Assume that lengths are normally distributed. Find the 95% confidence interval of the mean length.

Length	20.8	16.9	21.9	18	15	20.8	15.2	22.4	19.4	20.5
--------	------	------	------	----	----	------	------	------	------	------

31. Suppose you are a researcher in a hospital. You are experimenting with a new tranquilizer. You collect data from a random sample of 10 patients. The period of effectiveness of the tranquilizer for each patient (in hours) is as follows:

Hours	2	2.9	2.6	2.9	3	3	2	2.1	2.9	2.1
-------	---	-----	-----	-----	---	---	---	-----	-----	-----

- What is a point estimate for the population mean length of time?
 - What must be true in order to construct a confidence interval for the population mean length of time in this situation? Choose the correct answer below.
 - The sample size must be greater than 30.
 - The population must be normally distributed.
 - The population standard deviation must be known.
 - The population mean must be known.
 - Construct a 99% confidence interval for the population mean length of time.
 - What does it mean to be "99% confident" in this problem? Choose the correct answer below.
 - 99% of all confidence intervals found using this same sampling technique will contain the population mean time.
 - There is a 99% chance that the confidence interval contains the sample mean time.
 - The confidence interval contains 99% of all sample times.
 - 99% of all times will fall within this interval.
 - Suppose that the company releases a statement that the mean time for all patients is 2 hours. Is this possible? Is it likely?
32. The world's smallest mammal is the bumblebee bat (also known as Kitti's hog-nosed bat or *Craseonycteris thonglongyai*). Such bats are roughly the size of a large bumblebee. A sample of bats, weighed in grams, is

given in the below. Assume that bat weights are normally distributed. Find the 99% confidence interval of the mean.

Weight	
2.11	1.53
2.27	1.98
2.27	2.11
1.75	2.06
1.92	2.01

33. The total of individual weights of garbage discarded by 20 households in one week is normally distributed with a mean of 30.2 lbs. with a sample standard deviation of 8.9 lbs. Find the 90% confidence interval of the mean.
34. A student was asked to find a 90% confidence interval for widget width using data from a random sample of size $n = 29$. Which of the following is a correct interpretation of the interval $14.3 < \mu < 26.8$? Assume the population is normally distributed.
- There is a 90% chance that the sample mean widget width will be between 14.3 and 26.8.
 - There is a 90% chance that the widget width is between 14.3 and 26.8.
 - With 90% confidence, the width of a widget will be between 14.3 and 26.8.
 - With 90% confidence, the mean width of all widgets is between 14.3 and 26.8.
 - The sample mean width of all widgets is between 14.3 and 26.8, 90% of the time.
35. A researcher finds a 95% confidence interval for the average commute time in minutes using public transit is (15.75, 28.25). Which of the following is the correct interpretation of this interval?
- We are 95% confident that all commute time in minutes for the population using public transit is between 15.75 and 28.25 minutes.
 - There is a 95% chance commute time in minutes using public transit is between 15.75 and 28.25 minutes.
 - We are 95% confident that the interval $15.75 < \mu < 28.25$ contains the sample mean commute time in minutes using public transportation.
 - We are 95% confident that the interval $15.75 < \mu < 28.25$ contains the population mean commute time in minutes using public transportation.
36. A researcher would like to estimate the proportion of all children that have been diagnosed with autism spectrum disorder (ASD) in their county. They are using 95% confidence level and the Centers for Disease Control and Prevention (CDC) 2018 national estimate that 1 in 68 ≈ 0.0147 children are diagnosed with ASD. What sample size should the researcher use to get a margin of error to be within 2%?
37. A political candidate has asked you to conduct a poll to determine what percentage of people support her. If the candidate only wants a 9% margin of error at a 99% confidence level, what size of sample is needed?
38. A pilot study found that 72% of adult Americans would like an Internet connection in their car.
- Use the given preliminary estimate to determine the sample size required to estimate the proportion of adult Americans who would like an Internet connection in their car to within 0.02 with 95% confidence.
 - Use the given preliminary estimate to determine the sample size required to estimate the proportion of adult Americans who would like an Internet connection in their car to within 0.02 with 99% confidence.
 - If the information in the pilot study was not given, determine the sample size required to estimate the proportion of adult Americans who would like an Internet connection in their car to within 0.02 with 99% confidence.

39. The Food & Drug Administration (FDA) regulates that fresh albacore tuna fish that is consumed is allowed to contain 0.82 ppm of mercury or less. A laboratory is estimating the amount of mercury in tuna fish for a new company and needs to have a margin of error within 0.03 ppm of mercury with 95% confidence. Assume the population standard deviation is 0.138 ppm of mercury. What sample size is needed?
40. You want to obtain a sample to estimate a population mean age of the incoming fall term transfer students. Based on previous evidence, you believe the population standard deviation is approximately 5.3. You would like to be 90% confident that your estimate is within 1.9 of the true population mean. How large of a sample size is required?
41. SAT scores are distributed with a mean of 1,500 and a standard deviation of 300. You are interested in estimating the average SAT score of first year students at your college. If you would like to limit the margin of error of your 95% confidence interval to 25 points, how many students should you sample?
42. An engineer wishes to determine the width of a particular electronic component. If she knows that the standard deviation is 1.2 mm, how many of these components should she consider to be 99% sure of knowing the mean will be within 0.5 mm?

Rough drawings from memory were futile. He didn't even know how long it had been, beyond Ford Prefect's rough guess at the time that it was "a couple of million years" and he simply didn't have the maths. Still, in the end he worked out a method which would at least produce a result. He decided not to mind the fact that with the extraordinary jumble of rules of thumb, wild approximations and arcane guesswork he was using he would be lucky to hit the right galaxy, he just went ahead and got a result. He would call it the right result. Who would know? As it happened, through the myriad and unfathomable chances of fate, he got it exactly right, though he of course would never know that. He just went up to London and knocked on the appropriate door. "Oh. I thought you were going to phone me first."
(Adams, 2002)

Chapter 7

Hypothesis Tests for One Population



- 7.1 Introduction to Hypothesis Testing
- 7.2 Type I and II Errors
- 7.3 Hypothesis Test for One Proportion
- 7.4 Hypothesis Test for One Mean

7.1 Introduction to Hypothesis Testing

A **statistic** is a characteristic or measure from a sample. A **parameter** is a characteristic or measure from a population. We use statistics to generalize about parameters, known as estimations. Every time we take a sample statistic, we would expect that estimate to be close to the parameter, but not necessarily exactly equal to the unknown population parameter. How close would depend on how large a sample we took, who was sampled, how they were sampled and other factors. **Hypothesis testing** is a scientific method used to evaluate claims about population parameters.

A **statistical hypothesis** is an educated conjecture about a population parameter. This conjecture may or may not be true. We will take sample data and infer from the sample if there is evidence to support our claim about the unknown population parameter.

The **null hypothesis** (H_0 , pronounced “H-naught” or “H-zero”), is a statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters. The null hypothesis is assumed true until there is sufficient evidence otherwise.

The **alternative hypothesis** (H_1 or H_a , pronounced “H-one” or “H-ā”), is a statistical hypothesis that states that there is a difference between a parameter and a specific value, or that there is a difference between two parameters. H_1 is always the complement of H_0 .

The researcher decides the probability that the test is true by setting the **level of significance**, also called the significance level. We use the Greek letter alpha, α , pronounced “æ1-fə,” to represent the significance level. The level of significance is the probability that the null hypothesis is rejected when it is actually true. Note: like in the previous chapter, $1 - \alpha$ is the confidence level.

When doing your own research, you should set up their hypotheses and choose the significance level before analyzing the sample data.

When reading a word problem, your first step is to identify the parameter(s), for example μ , you are testing and which direction (left, right, or two-tail) test you are being asked to perform. For this course, the homework problems will state the researcher’s claim; usually this is the alternative hypothesis.

The null-hypothesis is always set up as a parameter equal to some value (called the test value) or equal to another parameter. The null hypothesis is assumed true unless there is strong evidence from the sample to suggest otherwise. Similar to our judicial system that a person is innocent until the prosecutor shows enough evidence that they are not innocent.

For example, an investment company wants to build a new food cart. They know from experience that food carts are successful if they have on average more than 100 people a day walk by the location. They have a potential site to build on, but before they begin, they want to see if they have enough foot traffic. They observe how many people walk by the site every day over a month. The investors want to be very careful about setting up in a bad location where the food cart will fail, rather than the missed opportunity build in a prime location. We have two hypotheses. For an average of more than 100 people, we would write this in symbols as $\mu > 100$. This claim needs to go into the alternative hypothesis since there is no equality, just strictly greater than 100. The complement of greater than is $\mu \leq 100$. This has a form of equality (\leq) so needs to go in the null hypothesis.

We then would set up the hypotheses as:

$H_0: \mu \leq 100$ (Do not build)

$H_1: \mu > 100$ (Build).

When performing the hypothesis test, the test statistic assumes that the parameter is the null hypothesis equal to some value.

This still implies that the parameter could be any value less than or equal to 100 but our hypothesis test should be written as:

$$H_0: \mu = 100$$

$$H_1: \mu > 100$$

Either notation is fine, but most textbooks will always have the = sign in the null hypothesis. The null hypothesis is based off historical value, a claim or product specification.

Signs are Important

When there is a greater than sign (>) in the alternative hypothesis, we call this a right-tailed test. If we had a less than sign (<) in the alternative hypothesis, then we would have a left-tailed test. If there were a not equal sign (\neq) in the alternative hypothesis, we would have a two-tailed test. The tails will determine which side the critical region will fall on the sampling distribution. Note that you should never have an =, \leq or \geq sign appear in the alternative hypothesis.

There are three ways to set up the hypotheses for a population proportion p :

<u>Two-tailed test</u> $H_0: p = p_0$ $H_1: p \neq p_0$	<u>Right-tailed test</u> $H_0: p = p_0$ $H_1: p > p_0$	<u>Left-tailed test</u> $H_0: p = p_0$ $H_1: p < p_0$
or		
<u>Two-tailed test</u> $H_0: p = p_0$ $H_1: p \neq p_0$	<u>Right-tailed test</u> $H_0: p \leq p_0$ $H_1: p > p_0$	<u>Left-tailed test</u> $H_0: p \geq p_0$ $H_1: p < p_0$

where p_0 is a placeholder for the numeric test value.

There are three ways to set up the hypotheses for a population mean μ :

<u>Two-tailed test</u> $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	<u>Right-tailed test</u> $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	<u>Left-tailed test</u> $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$
or		
<u>Two-tailed test</u> $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	<u>Right-tailed test</u> $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	<u>Left-tailed test</u> $H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$

where μ_0 is a placeholder for the numeric test value.

- The null-hypothesis of a **two-tailed test** states that the population parameter is equal to some hypothesized test value.
- The null-hypothesis of a **right-tailed test** implies that the population parameter is less than or equal to some hypothesized test value.
- The null-hypothesis of a **left-tailed test** implies that the population parameter is greater than or equal to some hypothesized test value.

When you read a question, it is essential that you identify the population parameter of interest. The parameter determines which distribution to use. Make sure that you can recognize and distinguish which parameter you are making a conjecture about a proportion = p or a mean = μ . There will be more parameters in later chapters.

Do not use the sample statistics, like \bar{x} or \hat{p} , in the hypotheses. We are not making any inference about the sample statistics. We know the value of the sample statistic. We use the sample statistics to infer if a change has occurred in the population.

For example, if we were making a conjecture about the percent or proportion in a population, we would have the hypotheses:

$$H_0: p = p_0$$

$$H_1: p \neq p_0.$$

Use Figure 7-1 as a guide in setting up your hypotheses. The first column shows the hypotheses and how to shade in the distribution for a two-tailed test, with common phrases in the claim. The two-tailed test will always have a not equal \neq sign in the alternative hypothesis and both tails shaded. The second column is for a right-tailed test. Note that the greater than $>$ sign always will be in the alternative hypothesis and the right tail is shaded. The third column is for a left-tailed test. The left-tailed test will always have a less than $<$ sign in the alternative hypothesis and the left tail shaded in.

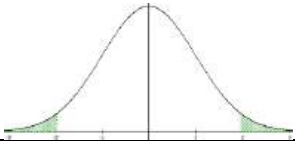
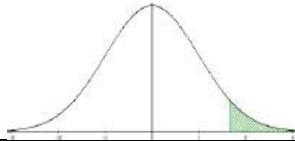
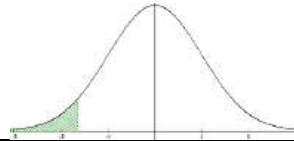
Hypothesis Testing Common Phrases		
Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ Or $H_0: p = p_0$ $H_1: p \neq p_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$ Or $H_0: p = p_0$ $H_1: p > p_0$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$ Or $H_0: p = p_0$ $H_1: p < p_0$
		
Claim is in the Null Hypothesis		
=	≤	≥
Is equal to	Is less than or equal to	Is greater than or equal to
Is exactly the same as	Is at most	Is at least
Has not changed from	Is not more than	Is not less than
Is the same as	Within	Is more than or equal to
Claim is in the Alternative Hypothesis		
≠	>	<
Is not	More than	Less than
Is not equal to	Greater than	Below
Is different from	Above	Lower than
Has changed from	Higher than	Shorter than
Is not the same as	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

Figure 7-1

Setting up the hypotheses correctly is the most important step in hypothesis testing. Here are some example research questions and how to set up the null and alternative hypotheses correctly; in a later section, we will perform the entire hypothesis test.

Example 7-1: State the hypotheses in both words and symbols for the following claims.

- a) The national mean salary for high school teachers is \$61,420. A random sample of 30 teacher’s salaries had a mean of \$49,850. A new director for a graduate teacher education program (GTEP) believes that the average salary of a teacher in Oregon is significantly less than national average.

Solution: The key phrase in the claim is “less than.” The less than sign $<$ is only allowed in the alternative hypothesis and we are testing against the national average.

H_0 : The national mean salary is \$61,420.

H_1 : The GTEP director believes the mean salary in Oregon is less than \$61,420. (claim)

$H_0: \mu = 61420$

$H_1: \mu < 61420$

- b) A high school principal is looking into assigning parking spaces at their school if the proportion of students who own their own car is more than 30%. The principal does not have the time to ask all 1,200 students at their school so instead takes a random sample of 70 students and found that 33% owned their own car.

Solution: The key phrase in the claim is “more than.” The greater than sign $>$ is only allowed in the alternative hypothesis. This is about a proportion, not a mean, so use the parameter p .

H_0 : The principal will not assign parking spaces if 30% or less of students own a car.

H_1 : The principal will assign parking spaces if more than 30% of students own a car. (claim)

$H_0: p = 0.3$

$H_1: p > 0.3$

- c) A teacher would like to know if the average age of students taking evening classes is different from the university’s average age of 26. They sample 40 students from a random sample of evening classes and found the average age to be 27.

Solution: The key word in the claim is “different.” The not equal sign \neq is only allowed in the alternative hypothesis.

H_0 : The population mean age is 26 years old.

H_1 : The evening students’ mean age is believed to be different from 26 years old. (claim)

$H_0: \mu = 26$

$H_1: \mu \neq 26$

Once we collect sample data, we need to find out how far away the sample statistic can be from the hypothesized parameter to say that a statistically significant change has occurred.

Example 7-2: Suppose a manufacturer of a new laptop battery claims the mean life of the battery is 900 days with a standard deviation of 40 days. You are the buyer of this battery and you think this claim is inflated. You would like to test your belief because without a good reason you cannot get out of your contract. You take a random sample of 35 batteries and find that the mean battery life is 890 days. What are the hypotheses for this question?

Solution: You have a guess that the mean life of a battery is less than 900 days. This is opposed to what the manufacturer claims. There really are two hypotheses, which are just guesses here – the one that the manufacturer claims and the one that you believe. For this problem:

$H_0: \mu = 900$, since the manufacturer says the mean life of a battery is 900 days.

$H_1: \mu < 900$, since you believe the mean life of the battery is less than 900 days.

Note that we do not put the sample mean of 890 in our hypotheses.

Is the sample mean of 890 days small enough to believe that you are right and the manufacturer is wrong? We would expect variation in our sample data and every time we take a new sample, the sample mean will most likely be different. How far away does the sample mean have to be from the product specification to verify our claim was correct? The sample data and the answer to these questions will be answered once we run the hypothesis test.

Three Methods for Hypothesis Testing

1. The Traditional Method (Critical Value Method)

The critical value is found from α and is the start of the shaded area called the critical region (also called rejection region or area). The test statistic is computed using sample data and may or may not be in the critical region. The critical value(s) is set before you begin (a priori) by the level of significance you are using for your test. This critical value(s) defines the shaded area known as the rejection area. The test statistic for this example is the z -score we find using the sample data that is then compared to the shaded tail(s). When the test statistic is in the shaded rejection area, you reject the null hypothesis. When your test statistic is not in the shaded rejection area, then you fail to reject the null hypothesis. Depending on if your claim is in the null or the alternative, the sample data may or may not support your claim.

There are five steps in hypothesis testing when using the traditional method:

- i. Identify the claim and formulate the hypotheses.
- ii. Compute the test statistic.
- iii. Compute the critical value(s) and state the rejection rule (the rule by which you will reject the null hypothesis (H_0)).
- iv. Make the decision to reject or not reject the null hypothesis by comparing the test statistic to the critical value(s). Reject H_0 when the test statistic is in the critical tail(s).
- v. Summarize the results and address the claim using context and units from the research question.

Steps ii and iii do not have to be in that order so make sure you know the difference between the critical value, which comes from the stated significance level α , and the test statistic, which is calculated from the sample data.

You can find the critical value(s) or test statistic in any order, but make sure you know the difference when you compare the two.

Note: The test statistic and the critical value(s) come from the same distribution and will usually have the same letter such as z , t , or F . The critical value(s) will have a subscript with the lower tail area (z_α , $z_{1-\alpha}$, $z_{\alpha/2}$) or an asterisk next to it (z^*) to distinguish it from the test statistic.

A researcher sets $\alpha = 0.05$. What z -score would represent that 5% area? It would depend on if the hypotheses were a left-tailed, two-tailed or right-tailed test. This z -score is called a critical value. Figure 7-2 shows examples of critical values for the three possible sets of hypotheses.

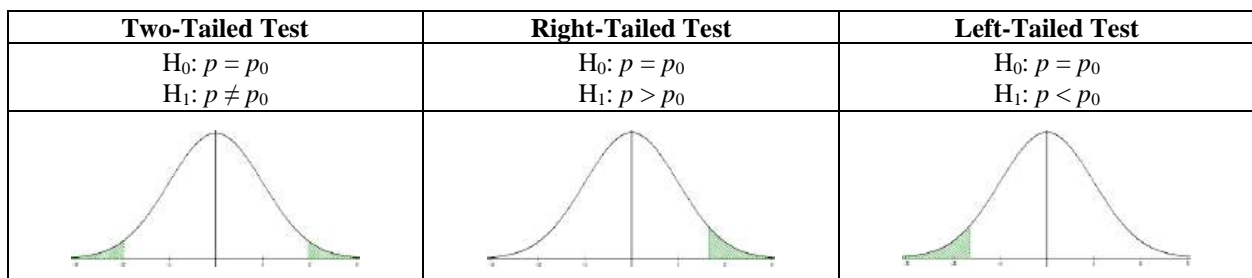


Figure 7-2

Critical Value for a Two-tailed Test

If we are doing a two-tailed test then the $\alpha = 5\%$ area gets divided into both tails. We denote these critical values $z_{\alpha/2}$ and $z_{1-\alpha/2}$. When the sample data finds a z -score (*test statistic*) that is either less than or equal to $z_{\alpha/2}$ or greater than or equal to $z_{1-\alpha/2}$ then we would reject H_0 . The area to the left of the critical value $z_{\alpha/2}$ and to the right of the critical value $z_{1-\alpha/2}$ is called the critical or rejection region. See Figure 7-3.

When $\alpha = 0.05$ then the critical values $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are found using the following technology.

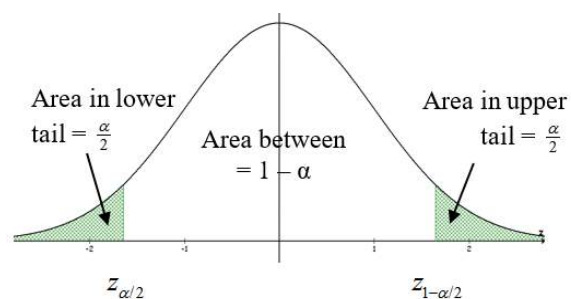


Figure 7-3

Excel: $z_{\alpha/2} = \text{NORM.S.INV}(0.025) = -1.96$ and $z_{1-\alpha/2} = \text{NORM.S.INV}(0.975) = 1.96$

TI-Calculator: $z_{\alpha/2} = \text{invNorm}(0.025,0,1) = -1.96$ and $z_{1-\alpha/2} = \text{invNorm}(0.975,0,1) = 1.96$

Since the normal distribution is symmetric, you only need to find one side's z -score and we usually represent the critical values as $\pm z_{\alpha/2}$.

Most of the time we will be finding a probability (p-value) instead of the critical values. The p-value and critical values are related and tell the same information so it is important to know what a critical value represents.

Critical Value for a Right-tailed Test

If we are doing a right-tailed test then the $\alpha = 5\%$ area goes into the right tail. We denote this critical value $z_{1-\alpha}$. When the sample data finds a z -score more than $z_{1-\alpha}$ then we would reject H_0 , reject H_0 if the *test statistic* is $\geq z_{1-\alpha}$. The area to the right of the critical value $z_{1-\alpha}$ is called the critical region. See Figure 7-4.

When $\alpha = 0.05$ then the critical value $z_{1-\alpha}$ is found using the following technology.

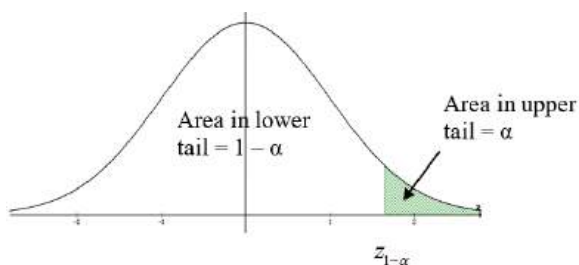


Figure 7-4

Excel: $z_{1-\alpha} = \text{NORM.S.INV}(0.95) = 1.645$

TI-Calculator: $z_{1-\alpha} = \text{invNorm}(0.95,0,1) = 1.645$

Critical Value for a Left-tailed Test

If we are doing a left-tailed test then the $\alpha = 5\%$ area goes into the left tail. If the sampling distribution is a normal distribution, then we can use the inverse normal function in Excel or calculator to find the corresponding z -score. We denote this critical value z_{α} .

When the sample data finds a z -score less than z_{α} then we would reject H_0 , reject H_0 if the *test statistic* is $\leq z_{\alpha}$. The area to the left of the critical value z_{α} is called the critical region. See Figure 7-5.

When $\alpha = 0.05$ then the critical value z_{α} is found using the following technology.

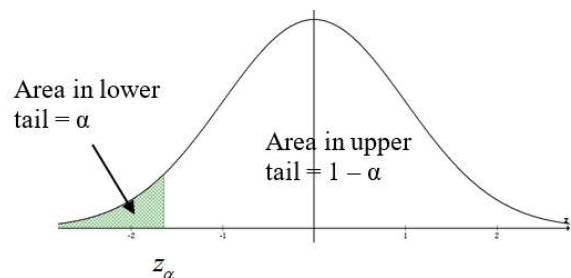


Figure 7-5

Excel: $z_{\alpha} = \text{NORM.S.INV}(0.05) = -1.645$

TI-Calculator: $z_{\alpha} = \text{invNorm}(0.05,0,1) = -1.645$

The rejection rule for critical value method: Reject H_0 when the test statistic is in the critical shaded region.

2. The P-value Method

In statistical hypothesis testing, a **p-value** is the probability of obtaining an observed result, or one more extreme, assuming that the null hypothesis is true. In the case of a single proportion test, the null hypothesis typically states that the proportion of interest is equal to a specific value, while the alternative hypothesis states that the proportion is different from that value. Most modern statistics and research methods utilize this method with the advent of computers and graphing calculators.

After collecting a sample and calculating the sample statistic, we can use statistical software to calculate the p-value. The p-value represents the probability of obtaining the observed sample statistic, or a more extreme, if the null hypothesis is true.

If the p-value is less than or equal to the significance level α , we reject the null hypothesis in favor of the alternative hypothesis, indicating that there is evidence to suggest that the population parameter of interest is significantly different from the value in the null hypothesis. On the other hand, if the p-value is greater than the significance level, we fail to reject the null hypothesis, indicating that there is insufficient evidence to suggest that the population parameter is significantly different from value in the null hypothesis.

Overall, the p-value is an important tool for determining the significance of results in statistical hypothesis testing. It allows us to assess the strength of evidence against the null hypothesis and make informed decisions about the population proportion of interest.

There are five steps in hypothesis testing when using the p-value method:

- i. Identify the claim and formulate the hypotheses.
- ii. Compute the test statistic.
- iii. Compute the p-value.
- iv. Make the decision to reject or not reject the null hypothesis by comparing the p-value with α . Reject H_0 when the p-value $\leq \alpha$.
- v. Summarize the results and address the claim.

The ideas below review the process of evaluating hypothesis tests with p-values:

- The null hypothesis represents a skeptic's position or a position of no difference. We reject this position only if the evidence strongly favors the alternative hypothesis.
- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favor of the alternative hypothesis.
- The p-value is constructed in such a way that we can directly compare it to the significance level (α) to determine whether to reject H_0 . We reject the null hypothesis if the p-value is smaller than the significance level, α , which is usually 0.05. Otherwise, we fail to reject H_0 .
- We should always state the conclusion of the hypothesis test in plain language use context and units so non-statisticians can also understand the results.

We will explore the p-value method in the next section.

The rejection rules for the previous two methods are:

- P-value method: reject H_0 when the p-value $\leq \alpha$.
- Critical value method: reject H_0 when the test statistic is in the critical region.

3. The Confidence Interval Method

A confidence interval can also be used for hypothesis testing by checking whether a hypothesized value for the population parameter falls within the interval. If the hypothesized value falls within the interval, then we fail to reject the null hypothesis that the population parameter is equal to the hypothesized value. If the hypothesized value

falls outside the interval, then we reject the null hypothesis in favor of the alternative hypothesis that the true value of the population parameter is different from the hypothesized value.

In this way, a confidence interval can be used as an alternative approach to hypothesis testing that provides a range of plausible values for the population parameter using the same units as the original data.

There are four steps in hypothesis testing when using the confidence interval method:

- i. Identify the claim and formulate the hypotheses.
- ii. Compute confidence interval.
- iii. Make the decision to reject or not reject the null hypothesis by comparing the p-value with α . Reject H_0 when the hypothesized value found in H_0 is outside the bounds of the confidence interval. We only will be doing a two-tailed version of this.
- iv. Summarize the results and address the claim.

For all 3 methods, Step i is the most important step. If you do not correctly set up your hypotheses then the next steps will be incorrect.

The Claim, Decision and Summary

The decision step is always about the null hypothesis, not about the claim. The hypothesis test's decision only has two outcomes, we either reject the null hypothesis, or fail to (do not) reject the null hypothesis. The summary should use context from the data and question to summarize the claim made in the question. The claim can be in either the null or the alternative hypothesis, but is usually in the alternative hypothesis. Don't combine the decision step with the summary statement.

Figure 7-6 is a flow chart that may help with starting your summaries, but make sure you finish the sentence with context and units from the question.

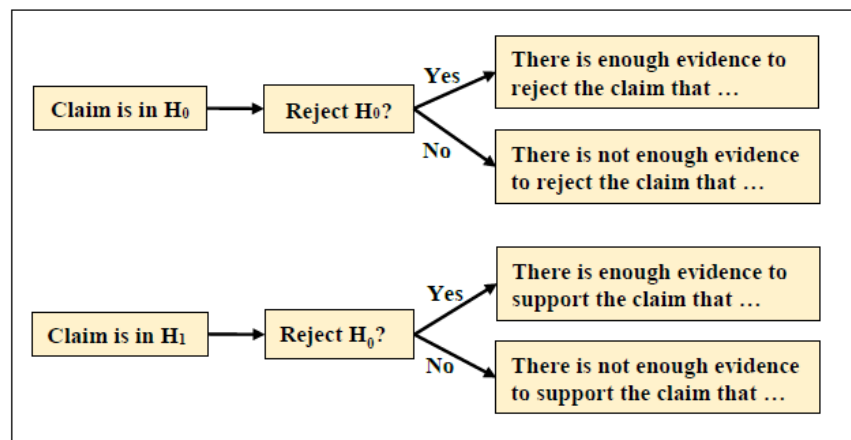


Figure 7-6

The hypothesis-testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative.

To understand the process of a hypothesis test, you need to first understand what a hypothesis is, which is an educated guess about a parameter. Once you have the alternative hypothesis, you collect data and use the data to decide to see if there is enough evidence to show that the alternative hypothesis is true. However, in hypothesis testing you actually assume something else is true, the null hypothesis, and then you look at your data to see how likely it is to get an event that your data demonstrates with that assumption. If the event is very unusual, then you might think that your assumption is actually false. If you are able to say this assumption is false, then your alternative hypothesis could be

true. You assume the opposite of your alternative hypothesis is true and show that it cannot be true. If this happens, then your alternative hypothesis is probably true. All hypothesis tests go through the same process. Once you have the process down, then the concept is much easier.

When setting up your hypotheses make sure the parameter, not the statistic, is used in the hypotheses. The equality always goes in the null hypothesis H_0 and the alternative hypothesis H_1 will be a left-tailed test with a less than sign $<$, a two-tailed test with a not equal sign \neq , or a right-tailed test with a greater than sign $>$.

“‘But alright,’ went on the rumblings, ‘so what’s the alternative?’
‘Well,’ said Ford, brightly but slowly, ‘stop doing it of course!’”
(Adams, 2002)

The wording on the summary statement changes depending on which hypothesis the researcher claims to be true. We really should always be setting up the claim in the alternative hypothesis since most of the time we are collecting evidence to show that a change has occurred, but occasionally a textbook will have the claim in the null hypothesis.

Do not use the phrase “accept H_0 ” since this implies that H_0 is true. The lack of evidence is not evidence of nothing.

There are only two possible correct answers for the decision step.

- i. Reject H_0
- ii. Fail to reject H_0

Caution! If we fail to reject the null this does not mean that there was no change, we just do not have any evidence that change has occurred. The absence of evidence is not evidence of absence. On the other hand, we need to be careful when we reject the null hypothesis, we have not proved that there is change.

When we reject the null hypothesis, there is only evidence that a change has occurred. Our evidence could have been false and lead to an incorrect decision. If we use the phrase, “accept H_0 ” this implies that H_0 was true, but we just do not have evidence that it is false. Hence you will be marked incorrect for your decision if you use accept H_0 , use instead “fail to reject H_0 ” or “do not reject H_0 .”

“Appeal to ignorance – the claim that whatever has not been proved false must be true, and vice versa (e.g. There is no compelling evidence that UFOs are not visiting the Earth; therefore, UFOs exist – and there is intelligent life elsewhere in the universe. Or: There may be seventy kazillion other worlds, but not one is known to have the moral advancement of the Earth, so we’re still central to the Universe). This impatience with ambiguity can be criticized in the phrase: absence of evidence is not evidence of absence.”
(Sagan, 1997)

7.2 Type I and II Errors

How do you quantify really small? Is 1% or 5% or 10% really small? How do you decide? That depends on your field of study and the importance of the situation. Is this a pilot study? Is someone’s life at risk? Would you lose your job? Most industry standards use 5% as the cutoff point for how small is small enough, but 1%, 5% and 10% are frequently used depending on what the situation calls for.

Now, how small is small enough? To answer that, you really want to know the types of errors you can make in hypothesis testing.

The first error is if you say that H_0 is false, when in fact it is true. This means you reject H_0 when H_0 was true. The second error is if you say that H_0 is true, when in fact it is false. This means you fail to reject H_0 when H_0 is false.

Figure 7-7 shows that if we “Reject H_0 ” when H_0 is actually true, we are committing a **type I error**. The probability of committing a type I error is denoted by the Greek letter alpha, α , pronounced “æ1-fə.” This can be controlled by the researcher by choosing a specific level of significance = α .

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Decision
Do Not Reject H_0	Correct Decision	Type II Error

Figure 7-7

Figure 7-7 shows that if we “Do Not Reject H_0 ” when H_0 is actually false, we are committing a **type II error**. The probability of committing a type II error is denoted with the Greek letter beta, β , pronounced “bei-tə.” When we increase the sample size this will reduce β . The **power of a test** is $1 - \beta$.

Example 7-3: A jury trial is about to take place to decide if a person is guilty of committing murder. The hypotheses for this situation would be:

- H_0 : The defendant is innocent
- H_1 : The defendant is not innocent

The jury has two possible decisions to make, either acquit or convict the person on trial, based on the evidence that is presented. There are two possible ways that the jury could make a mistake. They could convict an innocent person or they could let a guilty person go free. Both are bad news, but if the death penalty was sentenced to the convicted person, the justice system could be killing an innocent person. If a murderer is let go without enough evidence to convict them then they could possibly murder again. In statistics we call these two types of mistakes a type I and II error.

Figure 7-8 is a diagram to see the four possible jury decisions and two errors.

Jury’s Decision	What really happened?	
	H_0 True (the person is innocent)	H_0 False (the person is guilty)
Reject H_0 (Convict)	Type I Error $\alpha = P(\text{Type I error})$ (Level of Significance)	Correct Decision $1 - \beta$ (Power)
Do Not Reject H_0 (Acquit)	Correct Decision $1 - \alpha$ (Confidence Level)	Type II Error $\beta = P(\text{Type II error})$

Figure 7-8

Type I Error is rejecting H_0 when H_0 is true, and
Type II Error is failing to reject H_0 when H_0 is false.

Since these are the only two possible errors, one can define the probabilities attached to each error.

$$\alpha = P(\text{Type I Error}) = P(\text{Rejecting } H_0 \mid H_0 \text{ is true})$$

$$\beta = P(\text{Type II Error}) = P(\text{Failing to reject } H_0 \mid H_0 \text{ is false})$$

Example 7-4: An investment company wants to build a new food cart. They know from experience that food carts are successful if they have on average more than 100 people a day walk by the location. They have a potential site to build on, but before they begin, they want to see if they have enough foot traffic. They observe how many people walk by the site every day over a month. They will build if there is more than an average of 100 people who walk by the site each day. In simple terms, explain what the type I & II errors would be using context from the problem.

Solution: The hypotheses are: $H_0: \mu = 100$ and $H_1: \mu > 100$.

Sometimes it is helpful to use words next to your hypotheses instead of the formal symbols

$H_0: \mu \leq 100$ (Do not build)

$H_1: \mu > 100$ (Build).

A type I error would be to reject the null when in fact it is true. Take your finger and cover up the null hypothesis (our decision is to reject the null), then what is showing? The alternative hypothesis is what action we take.

If we reject H_0 then we would build the new food cart. However, H_0 was actually true, which means that the mean was less than or equal to 100 people walking by.

In more simple terms, this would mean that our evidence showed that we have enough foot traffic to support the food cart. Once we build, though, there was not on average more than 100 people that walk by and the food cart may fail.

A type II error would be to fail to reject the null when in fact the null is false. Evidence shows that we should not build on the site, but this actually would have been a prime location to build on.

The missed opportunity of a type II error is not as bad as possibly losing thousands of dollars on a bad investment.

What is more severe of an error is dependent on what side of the desk you are sitting on. For instance, if a hypothesis is about miles per gallon for a new car the hypotheses may be set up differently depending on if you are buying the car or selling the car. For this course, the claim will be stated in the problem and always set up the hypotheses to match the stated claim. In general, the research question should be set up as some type of change in the alternative hypothesis.

Visualizing α and β

If α increases that means the chances of making a type I error will increase. It is more likely that a type I error will occur with an increased α . It makes sense that you are less likely to make type II errors, only because you will be rejecting H_0 more often. You will be failing to reject H_0 less, and therefore, the chance of making a type II error will decrease. Thus, as α increases, β will decrease, and vice versa. That makes the errors seem like complements, but they are not complements. Consider one more factor – sample size.

Consider if you have a larger sample that is representative of the population, then it makes sense that you have more accuracy than with a smaller sample. Think of it this way, which would you trust more, a sample mean of 890 if you had a sample size of 35 or sample size of 350 (assuming a representative sample)? Of course, the 350 because there are more data points and so more accuracy. If you are more accurate, then there is less chance that you will make any error.

By increasing the sample size of a representative sample, you decrease β .

- For a constant sample size, n , if α increases, β decreases.
- For a constant significance level, α , if n increases, β decreases.

When the sample size becomes large, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a very small difference would likely be detected if we took a large enough sample size. Sometimes researchers will take such a large sample size that even the slightest difference is detected. While we still say that difference is statistically significant, it might not be practically significant. Statistically significant differences are sometimes so minor that they are not practically relevant. This is especially important to research: if we conduct a study, we want to focus on finding a meaningful result. We do not want to spend lots of money finding results that hold no practical value.

The role of a statistician in conducting a study often includes planning the size of the study. The statistician might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. The statistician should also obtain some reasonable estimate for the standard deviation. With these important pieces of information, they would choose a sufficiently large sample size so that the power for the meaningful difference is perhaps 80% or 90%. While larger sample sizes may still be used, the statistician might advise against using them in some cases, especially in sensitive areas of research.

If we look at the following two sampling distributions in Figure 7-9, the one on the left represents the sampling distribution for the true unknown mean. The curve on the right represents the sampling distribution based on the hypotheses the researcher is making. Do you remember the difference between a sampling distribution, the distribution of a sample, and the distribution of the population? Revisit the Central Limit Theorem in Chapter 5 if needed.

If we start with $\alpha = 0.05$, the critical value is represented by the vertical green line at $z_{\alpha} = 1.96$. Then the blue shaded area to the right of this line represents α . The area under the curve to the left of $z_{\alpha/2} = 1.96$ based on the researcher's claim would represent β .

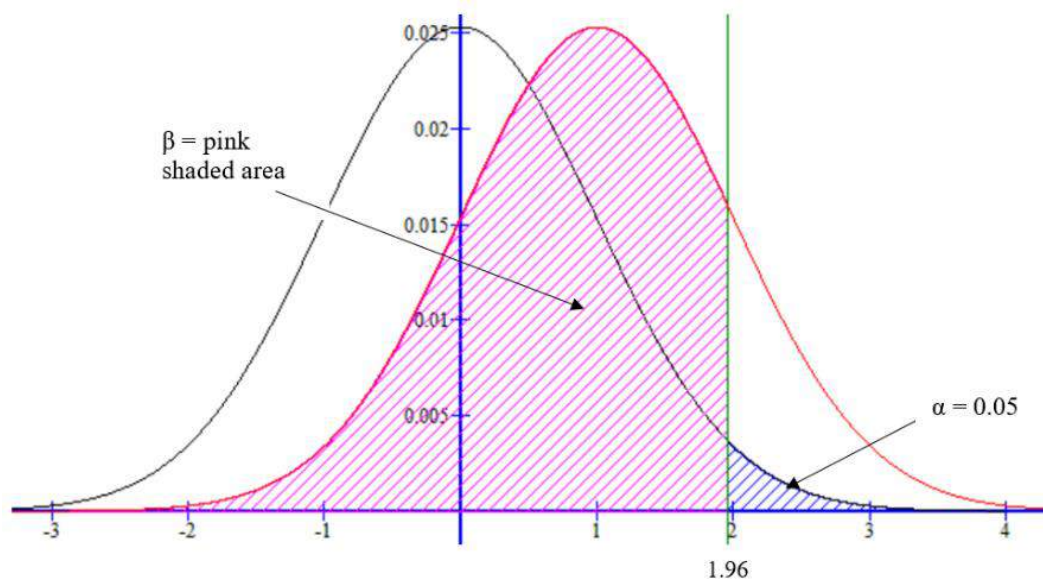


Figure 7-9

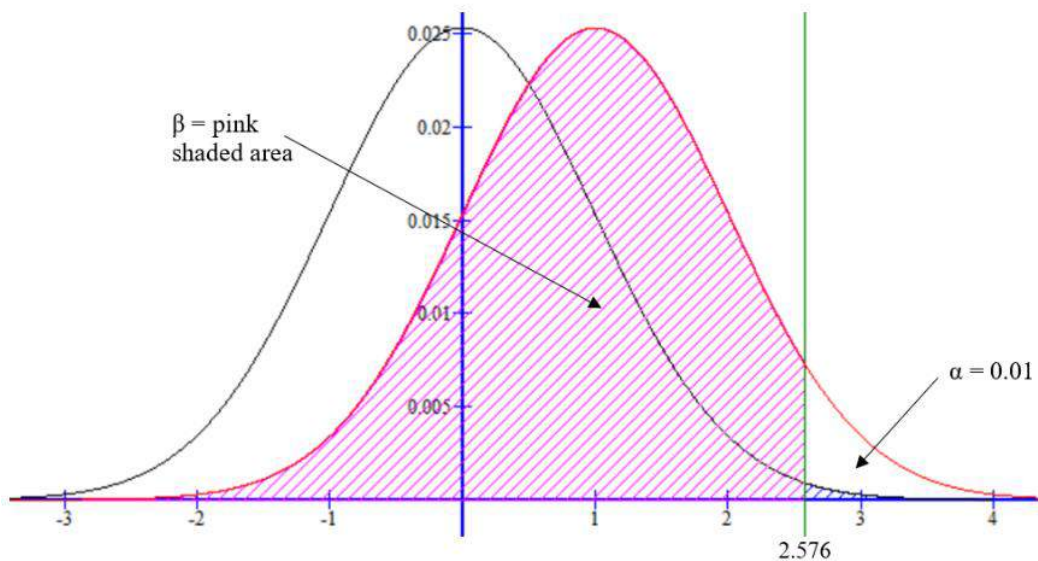


Figure 7-10

If we were to change α from 0.05 to 0.01 then we get a critical value of $z_{\alpha/2} = 2.576$. Note that when α decreases, then β increases which means your power $1 - \beta$ decreases. See Figure 7-10.

This text does not go over how to calculate β . You will need to be able to write out a sentence interpreting either the type I or II errors given a set of hypotheses. You also need to know the relationship between α , β , confidence level, and power.

Hypothesis tests are not flawless, since we can make a wrong decision in statistical hypothesis tests based on the data. For example, in the court system, innocent people are sometimes wrongly convicted and the guilty sometimes walk free, or diagnostic tests that have false negatives or false positives. However, the difference is that in statistical hypothesis tests, we have the tools necessary to quantify how often we make such errors. A type I Error is rejecting the null hypothesis when H_0 is actually true. A type II Error is failing to reject the null hypothesis when the alternative is actually true (H_0 is false).

We use the symbols $\alpha = P(\text{Type I Error})$ and $\beta = P(\text{Type II Error})$. The critical value is a cutoff point on the horizontal axis of the sampling distribution that you can compare your test statistic to see if you should reject the null hypothesis. For a left-tailed test the critical value will always be on the left side of the sampling distribution, the right-tailed test will always be on the right side, and a two-tailed test will be on both tails. Use technology to find the critical values. Most of the time in this course the shortcut menus that we use will give you the critical values as part of the output.

Controlling for Type I Error

The significance level used by the researcher should be picked prior to collection and analyzing data. This is called “a priori,” versus picking α after you have done your analysis which is called “post hoc.” When deciding on what significance level to pick, one needs to look at the severity of the consequences of the type I and type II errors. For example, if the type I error may cause the loss of life or large amounts of money the researcher would want to set α low.

Controlling for Type II Error

The power of a test is the complement of a type II error or correctly rejecting a false null hypothesis. You can increase the power of the test and hence decrease the type II error by increasing the sample size. Similar to confidence intervals, where we can reduce our margin of error when we increase the sample size. In general, we would like to have a high confidence level and a high power for our hypothesis tests. When you increase your confidence level, then in turn the power of the test will decrease. Calculating the probability of a type II error is a little more difficult and it is a conditional probability based on the researcher’s hypotheses and is not discussed in this course.

“‘That’s right!’ shouted Vroomfondel, ‘we demand rigidly defined areas of doubt and uncertainty!’”
(Adams, 2002)

7.3 Hypothesis Test for One Proportion

When you read a question, it is essential that you correctly identify the parameter of interest. The parameter determines which model to use. Make sure that you can recognize and distinguish between a question regarding a population mean and a question regarding a population proportion.

The **1 proportion z-test** is a statistical test for a population proportion. It can be used when $np \geq 10$ and $nq \geq 10$.

The formula for the *test statistic* is: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$

Where n = sample size
 $\hat{p} = \frac{x}{n}$ = sample proportion (sometimes already given as a %)
 p_0 = hypothesized population proportion, $q_0 = 1 - p_0$.

Use the phrases in Figure 7-11 to help with setting up the hypotheses.

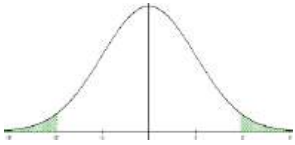
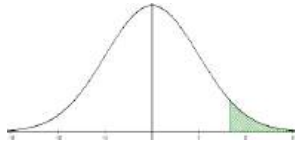
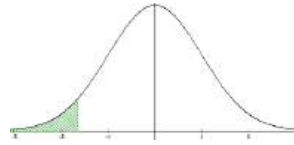
Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: p = p_0$ $H_1: p \neq p_0$	$H_0: p = p_0$ $H_1: p > p_0$	$H_0: p = p_0$ $H_1: p < p_0$
		
Claim is in the Null Hypothesis		
=	≤	≥
Is equal to	Is less than or equal to	Is greater than or equal to
Is exactly the same as	Is at most	Is at least
Has not changed from	Is not more than	Is not less than
Is the same as	Within	Is more than or equal to
Claim is in the Alternative Hypothesis		
≠	>	<
Is not	More than	Less than
Is not equal to	Greater than	Below
Is different from	Above	Lower than
Has changed from	Higher than	Shorter than
Is not the same as	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

Figure 7-11

We will use a standard normal z distribution for testing a proportion since this test uses the normal approximation to the binomial distribution. If the conditions $np \geq 10$ and $nq \geq 10$ are not met, then we could use the binomial distribution without the normal approximation. The exact binomial method is not discussed in this text.

If you are doing a left-tailed z -test the critical value will be negative. If you are performing a right-tailed z -test the critical value will be positive. If you were performing a two-tailed z -test then your critical values would be \pm critical value. The p -value, which is a probability, will always be a positive number between 0 and 1. The most important step in any method you use is setting up your null and alternative hypotheses. The critical value(s) and p -value can be found using a standard normal distribution.

Example 7-5: It has been found that 85.6% of all enrolled college students in the United States are undergraduates. A random sample of 500 enrolled college students in Oregon revealed that 420 of them were undergraduates. Is there sufficient evidence to conclude that the proportion differs from the national percentage? Use $\alpha = 0.05$. Show that all three methods of hypothesis testing yield the same results.

Solution: As you become more comfortable with the steps of a hypothesis test you do not have to number each step, but do know what each step means. Convert the national percentage of 85.6% to a proportion 0.856.

Critical Value Method

Step 1: State the hypotheses: The key words in this example, “proportion” and “differs,” give the hypotheses:

$$H_0: p = 0.856$$

$$H_1: p \neq 0.856 \text{ (claim)}$$

Step 2: Compute the test statistic. Before finding the test statistic, find the sample proportion:

$$\hat{p} = \frac{x}{n} = \frac{420}{500} = 0.84 \text{ and } q_0 = 1 - p_0 = 1 - 0.856 = 0.144.$$

Next, compute the test statistic: $z = \frac{\hat{p}-p_0}{\sqrt{\left(\frac{p_0q_0}{n}\right)}} = \frac{0.84-0.856}{\sqrt{\left(\frac{0.856 \cdot 0.144}{500}\right)}} = -1.019$.

Step 3: Draw and label the curve with the critical values. See Figure 7-12.

Use $\alpha = 0.05$ and technology to compute the critical values $z_{\alpha/2}$ and $z_{1-\alpha/2}$.

Excel: $z_{\alpha/2} = \text{NORM.S.INV}(0.025) = -1.96$ and $z_{1-\alpha/2} = \text{NORM.S.INV}(0.975) = 1.96$.

TI-Calculator: $z_{\alpha/2} = \text{invNorm}(0.025,0,1) = -1.96$ and $z_{1-\alpha/2} = \text{invNorm}(0.975,0,1) = 1.96$.

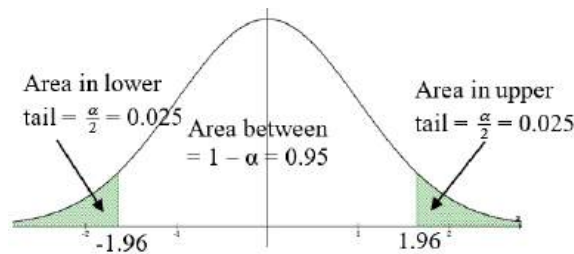


Figure 7-12

Step 4: State the decision.

Since the test statistic -1.019 is not in the shaded rejection area, the decision is to fail to reject H_0 .

Step 5: State the summary.

At the 5% level of significance, there is not enough evidence to conclude that the proportion of undergraduates in Oregon colleges differ from the national average of 85.6%.

Example 7-5 Continued using the P-value Method

Step 1: The hypotheses stay the same.

$$H_0: p = 0.856$$

$$H_1: p \neq 0.856 \text{ (claim)}$$

Step 2: The test statistic stays the same. $z = \frac{\hat{p}-p_0}{\sqrt{\left(\frac{p_0q_0}{n}\right)}} = \frac{0.84-0.856}{\sqrt{\left(\frac{0.856 \cdot 0.144}{500}\right)}} = -1.019$

Step 3: Calculate the p-value. The p-value represents the probability of seeing the sample statistic $z = \pm 1.019$ or more extreme.

To find the p-value we need to find the $P(Z > |1.019|)$ the area to the left of $z = -1.019$ and to the right of $z = 1.019$.

First, find the area below (since the test statistic is negative) $z = -1.019$ using the normalcdf we get $P(Z < -1.019) = 0.1541$. Then, double this area to get the p-value = 0.3082.

```
normalcdf(-1E99,
-1.019, 0, 1)
.1541014962
Ans*2
.3082029924
```

Figure 7-13 shows a visual representation of the p-value area shaded in blue, and the critical value area shaded in green. When the test statistic is within the critical rejection region, the p-value $\leq \alpha$. When the test statistic is not in the critical rejection region, then the p-value $> \alpha$.

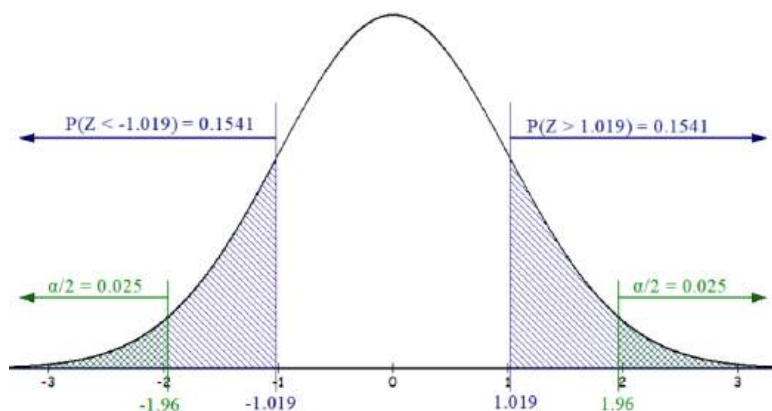


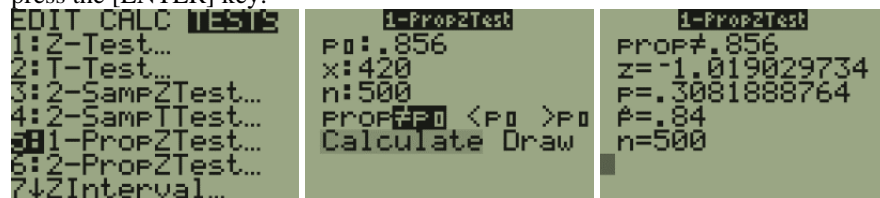
Figure 7-13

Step 4: Decision: Since the p-value $> \alpha$ the decision is to fail to reject H_0 .

Step 5: Summary: There is not enough evidence to conclude that the proportion of undergraduates in college for this state differs from the national average of 85.6%.

There is a shortcut for this test on the TI Calculators, which will quickly find the test statistic and p-value. There is no shortcut for this type of test in Excel.

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [5:1-PropZTest] and press the [ENTER] key. Type in the hypothesized proportion (p_0), x , sample size, arrow over to the \neq , $<$, $>$ sign that is the same in the problem's alternative hypothesis statement then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key.



The calculator returns the z-test statistic and the p-value. Note: sometimes you are not given the x value but a percentage instead. To find the x to use in the calculator, multiply \hat{p} by the sample size and round off to the nearest integer. The calculator will give you an error message if you put in a decimal for x or n . For example, if $\hat{p} = 0.22$ and $n = 124$ then $0.22 * 124 = 27.28$, so use $x = 27$.

TI-89: Go to the [Apps] Stat/List Editor, then press [2nd] then F6 [Tests], then select **5: 1-PropZ-Test**. Type in the hypothesized proportion (p_0), x , sample size, arrow over to the \neq , $<$, $>$ sign that is the same in the problem's alternative hypothesis statement then press the [ENTER] key to calculate. The calculator returns the z-test statistic and the p-value. Note: sometimes you are not given the x value but a percentage instead. To find the x value to use in the calculator, multiply \hat{p} by the sample size and round off to the nearest integer. The calculator will give you an error message if you put in a decimal for x or n . For example, if $\hat{p} = 0.22$ and $n = 124$ then $0.22 * 124 = 27.28$, so use $x = 27$.

Example 7-6: Nationally the percentage of adults that have their teeth cleaned by a dentist yearly is 64%. A dentist in Portland, OR believes that regionally the percent is higher. A sample of 2000 Portlanders found that 65.9% had their teeth cleaned by a dentist in the last year. Test the dentist's claim using a 5% level of significance using the p-value method.

Solution: Set up the hypotheses, the key word in the claim is "higher" which indicates a right-tailed test.

$$H_0: p = 0.64$$

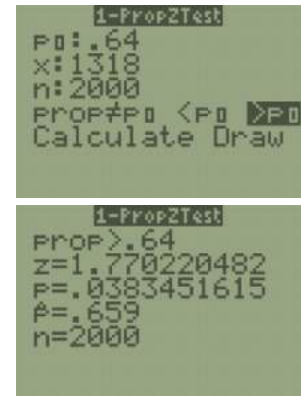
$$H_1: p > 0.64 \text{ (claim)}$$

Next, calculate the test statistic $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.659 - 0.64}{\sqrt{\frac{0.64 \cdot 0.36}{2000}}} = 1.7702$.

Use technology to calculate the p-value. Since this is a right-tailed test we want to find the $P(Z > 1.7702) = 0.0383$. Note that if you are using the TI calculator, you need x as the number of successes, not the proportion of successes. Since $\hat{p} = x/n$, we can use a little algebra and solve for x to get $x = n \cdot \hat{p} = 2000 \cdot 0.659 = 1318$.

Make the decision. Since the p-value = $0.0383 \leq \alpha = 0.05$, we reject H_0 .

Summary: There is enough evidence to support the claim that population proportion of adult Portlanders that have their teeth cleaned by a dentist yearly is higher than the national proportion of 64%.



7.4 Hypothesis Test for One Mean

Sometimes we will encounter situations where we want to test whether a certain value represents the true population mean instead of a proportion. To perform the hypothesis test, we collect representative sample data from the population of interest and calculate the sample mean (\bar{x}) and the sample standard deviation (s). We then compare the sample mean to the hypothesized value, taking into account the variability in the sample.

Example 7-7: Suppose a manufacturer of a new laptop battery claims the mean life of the battery is 900 days with a standard deviation of 40 days. You are the buyer of this battery and you think this claim is inflated. You would like to test your belief because without a good reason you cannot get out of your contract. You take a random sample of 35 batteries and find that the mean battery life is 890 days. What are the hypotheses for this question?

Solution: You have a guess that the mean life of a battery is less than 900 days. This is opposed to what the manufacturer claims. There really are two hypotheses, which are just guesses here – the one that the manufacturer claims and the one that you believe. For this problem:

- $H_0: \mu = 900$, since the manufacturer says the mean life of a battery is 900 days.
- $H_1: \mu < 900$, since you believe the mean life of the battery is less than 900 days.

Note that we do not put the sample mean of 890 in our hypotheses.

Is the sample mean of 890 days small enough to believe that you are right and the manufacturer is wrong? We would expect variation in our sample data and every time we take a new sample, the sample mean will most likely be different. How far away does the sample mean have to be from the product specification to verify our claim was correct? The sample data and the answer to these questions will be answered once we run the hypothesis test.

If you calculated a sample mean of 435, you would definitely believe the population mean is less than 900. However, even if you had a sample mean of 835 you would probably believe that the true mean was less than 900. What about 875? Or 893? There is some point where you would stop being so sure that the population mean is less than 900. That point separates the values of where you are sure or pretty sure that the mean is less than 900 from the area where you are not so sure.

How do you find that point where the sample mean is close enough to the hypothesized population mean? How close depends on how much error you want to make. Of course, you do not want to make any errors, but unfortunately, that is unavoidable in statistics since we are not measuring the entire population. You need to figure out how much error you made with your sample. Take the sample mean, and find the probability of getting another sample mean less than it, assuming for the moment that the manufacturer is right. The idea behind this is that you want to know what is the chance that you could have come up with your sample mean even if the population mean really is 900 days.

You want to find $P(\bar{X} < 890 \mid H_0 \text{ is true}) = P(\bar{X} < 890 \mid \mu = 900)$. For short, we will call this probability the p-value.

To compute this p-value, you need to know how the sample mean is distributed. Since the sample size is at least 30 you know the sample mean is approximately normally distributed, by the Central Limit Theorem (CLT). Remember $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Before calculating the probability, it is useful to see how many standard deviations away from the mean the sample mean is. Using the formula for the z-score for CLT, $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$ we can compare this z-score to a z-score based on how sure we want to be of not making a mistake.

Using our sample mean we compute the z-score: $z = \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{890 - 900}{\left(\frac{40}{\sqrt{35}}\right)} = -1.479$. This sample mean is more than one standard deviation away from the mean. Is that far enough?

Look at the probability $P(\bar{X} < 890 \mid H_0 \text{ is true}) = P(\bar{X} < 890 \mid \mu = 900) = P(Z < -1.479)$.

Using the TI Calculator `normalcdf(-1E99,890,900,40/√35) ≈ 0.0696`.

Alternatively, in Excel use `=NORM.DIST(890,900,40/SQRT(35),TRUE) ≈ 0.0696`.

Hence the p-value = 0.0696.

A picture is always useful. Figure 7-14 shows the population distribution. Figure 7-15 shows the sampling distribution of the mean.

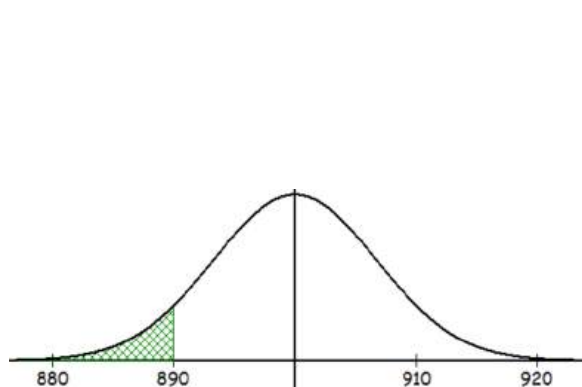


Figure 7-14

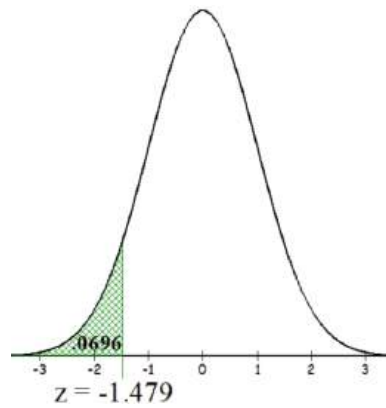


Figure 7-15

There is approximately a 6.96% chance that you could find a sample mean less than 890 when the population mean is 900 days. This is small but not really small. But how do you quantify really small? Is 1% or 5% or 10% really small? How do you decide? That depends on your field of study and the importance of the situation. We set the significance level on how small we want to quantify a rare event.

When the population standard deviation is known and stated in the problem, we will use the **z-test**.

The **z-test** is a statistical test for the mean of a population. It can be used when σ is known. The population should be approximately normally distributed when $n < 30$.

When using this model, the *test statistic* is $z = \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$ where μ_0 is the test value from the H_0 .

However, in practice, we rarely know the actual value of the population standard deviation (σ). Instead, we use the sample standard deviation (s) as an estimate for the population standard deviation. In this case, the sampling

distribution is no longer normally distributed, but instead follows a t-distribution. We will mostly use the t-distribution when testing means, but in case you come across the situation that has a known σ , then use the z instead of the t.

When the population standard deviation is unknown, use the t-test.

The **t-test** is a statistical test for the mean of a population. It will be used when σ is unknown. The population should be approximately normally distributed when $n < 30$.

When using this model, the *test statistic* is $t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$ where μ_0 is the test value from the H_0 .

The degrees of freedom are $df = n - 1$.

Use Figure 7-16 to decide when to use z versus t. In practice, σ is rarely known and usually comes from a similar study or previous year's data.

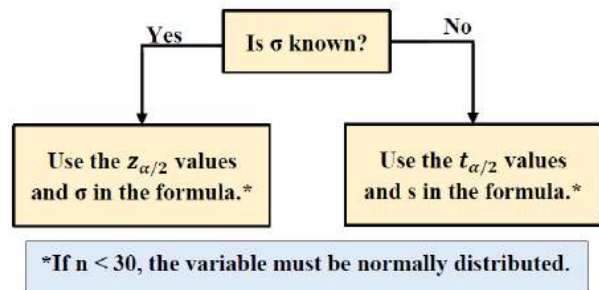


Figure 7-16

Use Figure 7-17 as a guide in setting up your hypotheses for testing a mean. The two-tailed test will always have a not equal \neq sign in H_1 and both tails shaded. The right-tailed test will always have the greater than $>$ sign in H_1 and the right tail shaded. The left-tailed test will always have a less than $<$ sign in H_1 and the left tail shaded.

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$
Claim is in the Null Hypothesis		
=	\leq	\geq
Is equal to	Is less than or equal to	Is greater than or equal to
Is exactly the same as	Is at most	Is at least
Has not changed from	Is not more than	Is not less than
Is the same as	Within	Is more than or equal to
Claim is in the Alternative Hypothesis		
\neq	$>$	$<$
Is not	More than	Less than
Is not equal to	Greater than	Below
Is different from	Above	Lower than
Has changed from	Higher than	Shorter than
Is not the same as	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

Figure 7-17

Example 7-8: M&Ms candies advertise a mean weight of 0.8535 grams. A sample of 50 M&M candies are randomly selected from a bag of M&Ms and the mean weight is found to be $\bar{x} = 0.8472$ grams with a standard deviation of $s = 0.06$ grams. A skeptic M&M consumer claims that the mean weight is less than what is advertised. Test this claim using a 5% level of significance.

Solution: By letting $\alpha = 0.05$, we are allowing a 5% chance that the null hypothesis (average weight that is at least 0.8535 grams) is rejected when in actuality it is true.

Step 1: Identify the claim and set up the hypotheses. The claim is “M&Ms candies have a mean weight that is less than 0.8535 grams.” This translates into mathematical symbols as $\mu < 0.8535$ grams. Therefore, the null and alternative hypotheses are:

$$H_0: \mu = 0.8535$$

$$H_1: \mu < 0.8535 \text{ (claim)}$$

This is a left-tailed test since the alternative hypothesis has a “less than” sign.

We are performing a test about a population mean. We use the t-test because we were given a sample standard deviation.

Step 2: Compute the critical value and draw and label the sampling distribution. The critical value for a left-tailed test with a level of significance $\alpha = 0.05$ is found in a way similar to finding the critical values from confidence interval for a mean. Because we are using the t-test, we must find the critical value t_α from the t-distribution with $df = n - 1 = 50 - 1 = 49$.

This is a left-tailed test since the sign in the alternative hypothesis has a less than $<$ sign (most of the time a left-tailed test will have a negative t-score for the test statistic).

First draw your bell curve and shade the appropriate tail with the area $\alpha = 0.05$. See Figure 7-18. Usually, the technology you are using only asks for the area in the left tail, which in this case is $\alpha = 0.05$. For the TI calculators, under the DISTR menu use $\text{invT}(0.05,49) = -1.6766$. For Excel use the formula $=\text{T.INV}(0.05,49)$ to get the critical value.

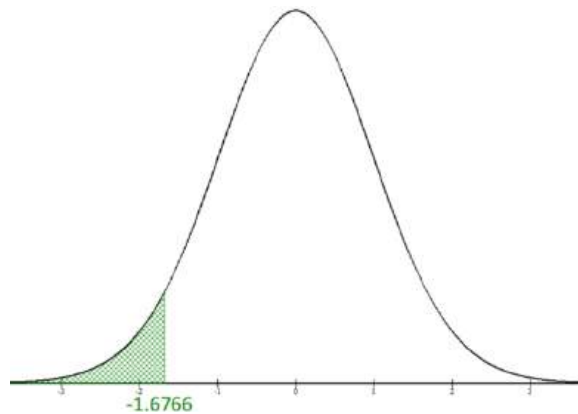
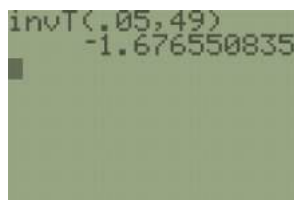


Figure 7-18

Step 3: Compute the test statistic.

$$\text{The formula for the test statistic is } t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{0.8472 - 0.8535}{\left(\frac{0.06}{\sqrt{50}}\right)} = -0.7425.$$

Step 4: Make the decision by comparing the test statistic to the critical value. Figure 7-19 shows both the critical value and the test statistic. There are only two possible correct answers for the decision step.

- i. Reject H_0
- ii. Fail to reject H_0

To make the decision whether to “Fail to reject H_0 ” or “Reject H_0 ” using the traditional method, we must compare the test statistic $t = -0.7425$ with the critical value $t_\alpha = -1.6766$.

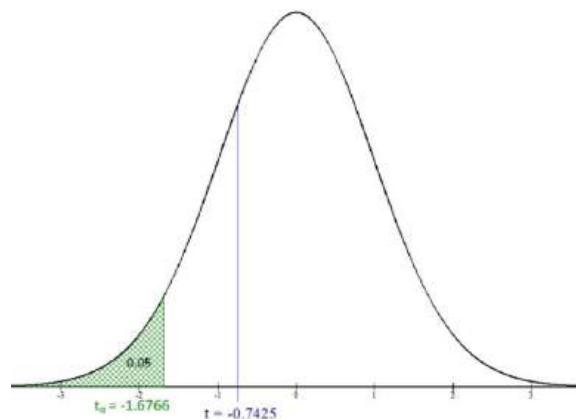


Figure 7-19

When the test statistic is in the shaded tail, called the rejection area, then we would reject H_0 , if not then we fail to reject H_0 . Since the test statistic $t = -0.7425$ is in the unshaded region, the decision is: Do not reject H_0 .

Step 5: Summarize the results. At 5% level of significance, there is not enough evidence to support the claim that the mean weight of an M&M candy is less than 0.8535 grams.

Example 7-8 used the traditional critical value method. With the onset of computers, this method is outdated and the p-value and confidence interval methods are becoming more popular.

Most statistical software packages will give a p-value and confidence interval but not the critical value.

Example 7-9: The label on a particular brand of cream of mushroom soup states that (on average) there is 870 mg of sodium per serving. A nutritionist would like to test if the average is actually more than the stated value. To test this, 13 servings of this soup were randomly selected and amount of sodium measured. The sample mean was found to be 882.4 mg and the sample standard deviation was 24.3 mg. Assume that the amount of sodium per serving is normally distributed. Test this claim using the traditional method of hypothesis testing. Use the $\alpha = 0.05$ level of significance.

Solution:

Step 1: State the hypotheses and identify the claim: The statement “the average is more ($>$) than 870” must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

$$H_0: \mu = 870$$

$$H_1: \mu > 870 \text{ (claim)}$$

This is a right-tailed test with the claim in the alternative hypothesis.

Step 2: Compute the test statistic: We are using the t-test because we are performing a test about a population mean. We must use the t-test (instead of the z-test) because the population standard deviation σ is unknown. (Note: be sure that you know why we are using the t-test instead of the z-test in general.)

$$\text{The formula for the test statistic is } t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{882.4 - 870}{\left(\frac{24.3}{\sqrt{13}}\right)} = 1.8399.$$

Note: If you were given raw data use 1-var Stats on your calculator to find the sample mean, sample size and sample standard deviation.

Step 3: Compute the critical value(s): The critical value for a right-tailed test with a level of significance $\alpha = 0.05$ is found in a way similar to finding the critical values from confidence intervals.

Since we are using the t-test, we must find the critical value $t_{1-\alpha}$ from a t-distribution with the degrees of freedom, $df = n - 1 = 13 - 1 = 12$. Use the DISTR menu **invT** option. Note that if you have an older TI-84 or a TI-83 calculator you need to have the invT program installed or use Excel.

Draw and label the t-distribution curve with the critical value as shown in Figure 7-20.

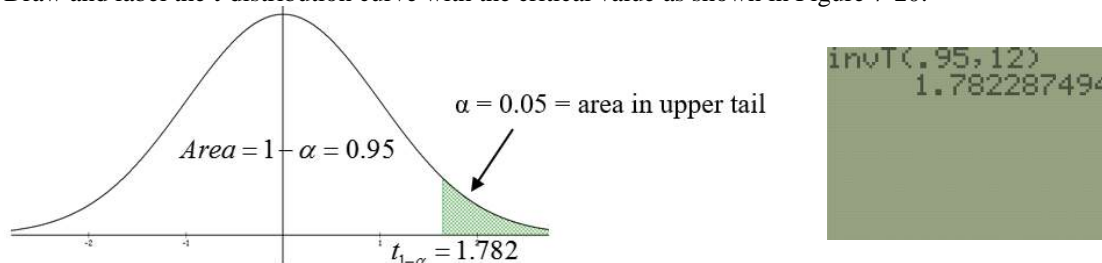


Figure 7-20

The critical value is $t_{1-\alpha} = 1.782$ and the rejection rule becomes: Reject H_0 if the test statistic $t \geq t_{1-\alpha} = 1.782$.

Step 4: State the decision. Decision: Since the test statistic $t = 1.8399$ is in the critical region, we should reject H_0 .

Step 5: State the summary. Summary: At the 5% significance level, we have sufficient evidence to say that the average amount of sodium per serving of cream of mushroom soup exceeds the stated 870 mg amount.

Example 7-9 Continued:

Use the prior example, but this time use the **p-value method**. Again, let the significance level be $\alpha = 0.05$.

Solution:

Step 1: The hypotheses remain the same. $H_0: \mu = 870$
 $H_1: \mu > 870$ (claim)

Step 2: The test statistic remains the same, $t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{882.4 - 870}{\left(\frac{24.3}{\sqrt{13}}\right)} = 1.8399$.

Step 3: Compute the p-value.

The p-value is the probability of observing an effect as least as extreme as in your sample data, assuming that the null hypothesis is true. The p-value is calculated based on the assumptions that the null hypothesis is true for the population and that the difference in the sample is caused entirely by random chance.

For a right-tailed test, the p-value is found by finding the area to the right of the test statistic $t = 1.8399$ under a t-distribution with 12 degrees of freedom. See Figure 7-21.

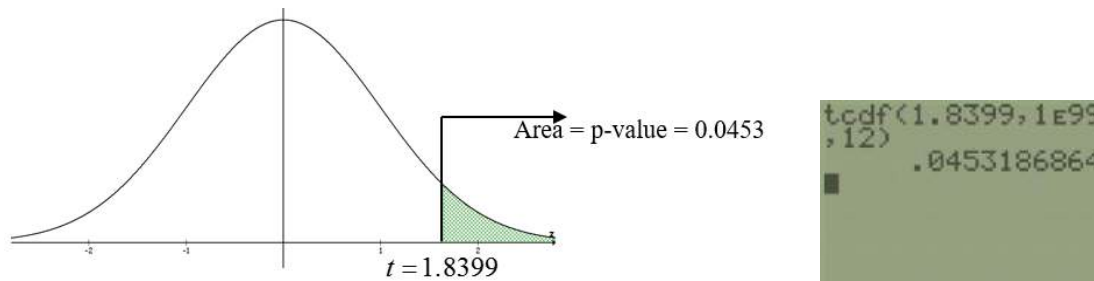


Figure 7-21

Note that exact p-values for a t-test is found using a computer or calculator. For the TI calculators this is in the DISTR menu. Use $\text{tcdf}(\text{lower}, \text{upper}, df)$. In Excel use the T.DIST function.

TI calculator use $\text{tcdf}(1.8399, \infty, 12)$. Excel use $=1 - \text{T.DIST}(1.8399, 12, \text{TRUE})$. The p-value = 0.0453.

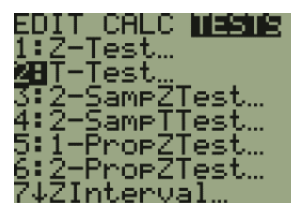
The p-value is the probability of observing an effect as least as extreme as in your sample data, assuming that the null hypothesis is true. The p-value is calculated based on the assumptions that the null hypothesis is true for the population and that the difference in the sample is caused entirely by random chance.

Step 4: State the decision. The rejection rule: reject the null hypothesis if the p-value $\leq \alpha$. Decision: Since the p-value = 0.0453 is less than $\alpha = 0.05$, we Reject H_0 . This agrees with the decision from the traditional method. (These two methods should always agree!)

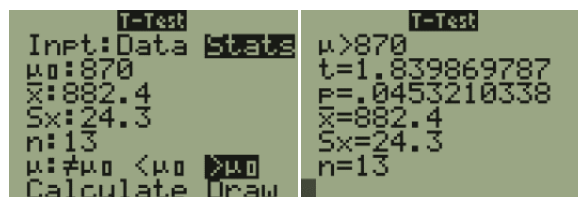
Step 5: State the summary. The summary remains the same as in the previous method. At the 5% significance level, we have sufficient evidence to say that the average amount of sodium per serving of cream of mushroom soup exceeds the stated 870 mg amount.

We can use technology to get the test statistic and p-value.

TI-84: If you have raw data, enter the data into a list before you go to the test menu. Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the [2:T-Test] option and press the [ENTER] key. Arrow over to the [Stats] menu and press the [ENTER] key. Then type in the hypothesized mean (μ_0), sample or population standard deviation, sample mean, sample size, arrow over to the $\neq, <, >$ sign that is the same as the problem's alternative hypothesis statement then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the t-test statistic and p-value.



Alternatively (If you have raw data in list one) Arrow over to the [Data] menu and press the [ENTER] key. Then type in the hypothesized mean (μ_0), L_1 , leave Freq:1 alone, arrow over to the $\neq, <, >$ sign that is the same in the problem's alternative hypothesis statement then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the t-test statistic and the p-value.



TI-89: Go to the [Apps] **Stat/List Editor**, then press [2nd] then F6 [Tests], then select **2: T-Test**. Choose the input method, data is when you have entered data into a list previously or stats when you are given the mean and standard deviation already. Then type in the hypothesized mean (μ_0), sample standard deviation, sample mean, sample size (or list name (list1), and Freq: 1), arrow over to the $\neq, <, >$ and select the sign that is the same as the problem's alternative hypothesis statement then press the [ENTER] key to calculate. The calculator returns the t-test statistic and p-value.



Example 7-10: The weight of the world's smallest mammal is the bumblebee bat (also known as Kitti's hog-nosed bat or *Craseonycteris thonglongyai*) is approximately normally distributed with a mean 1.9 grams. Such bats are roughly the size of a large bumblebee. A chiropterologist believes that the Kitti's hog-nosed bats in a new geographical region under study has a different average weight than 1.9 grams. A sample of 10 bats weighed in grams in the new region are shown below. Use the confidence interval method to test the claim that mean weight for all bumblebee bats is not 1.9 g using a 10% level of significance.

Weight	1.9	2.24	2.13	2	1.54	1.96	1.79	2.18	1.81	2.3
--------	-----	------	------	---	------	------	------	------	------	-----

Solution:

Step 1: State the hypotheses and identify the claim. The key phrase is “mean weight not equal to 1.9 g.” In mathematical notation, this is $\mu \neq 1.9$. The not equal \neq symbol is only allowed in the alternative hypothesis so the hypotheses would be:

$$\begin{aligned} H_0: \mu &= 1.9 \\ H_1: \mu &\neq 1.9 \end{aligned}$$

Step 2: Compute the confidence interval. First, find the t critical value using $df = n - 1 = 9$ and 90% confidence. In Excel $t_{\alpha/2} = T.INV(.1/2,9) = -1.833113$.

Then use technology to find the sample mean and sample standard deviation and substitute in your numbers to the formula.

$$\begin{aligned} \bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) \\ \Rightarrow 1.985 \pm 1.833113 \left(\frac{0.235242}{\sqrt{10}} \right) \\ \Rightarrow 1.985 \pm 1.833113(0.07439) \\ \Rightarrow 1.985 \pm 0.136365 \\ \Rightarrow (1.8486, 2.1214) \end{aligned}$$

The answer can be given as an inequality $1.8486 < \mu < 2.1214$

or in interval notation $(1.8486, 2.1214)$.

Step 3: Make the decision: The rejection rule is to reject H_0 when the hypothesized value found in H_0 is outside the bounds of the confidence interval. The null hypothesis was $\mu = 1.9$ g. Since 1.9 is between the lower and upper boundary of the confidence interval $1.8486 < \mu < 2.1214$ then we would fail to reject H_0 .

The sampling distribution, assuming the null hypothesis is true, will have a mean of $\mu = 1.9$ and a standard error of $\frac{0.2352}{\sqrt{10}} = 0.07439$. When we calculated the confidence interval using the sample mean of 1.985 the confidence interval captured the hypothesized mean of 1.9. See Figure 7-22.

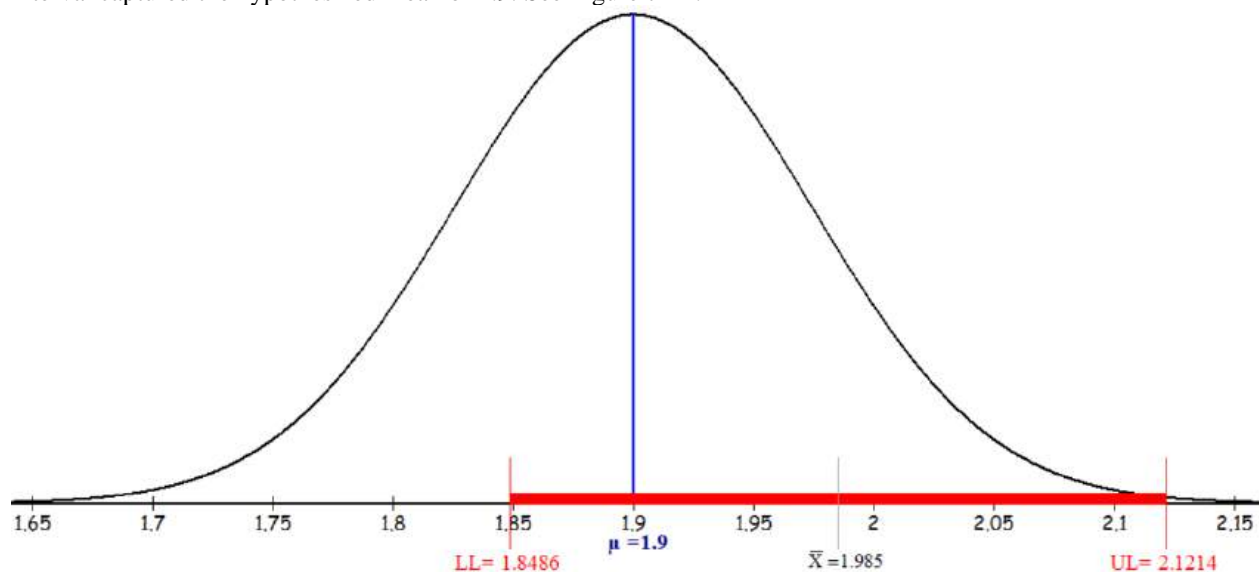


Figure 7-22

Step 4: State the summary: At the 10% significance level, there is not enough evidence to support the claim that the population mean weight for bumblebee bats in the new geographical region is different from 1.9 g.

This confidence interval can also be computed using a TI calculator or Excel (although Excel only works when you have raw data).

TI-84: Enter the data in a list, choose Tests > TInterval. Select and highlight Data, change the list and confidence level to match the question. Choose Calculate.



Excel: Select Data Analysis > Descriptive Statistics: Note, you will need to change the cell reference numbers to where you copy and paste your data, only check the label box if you selected the label in the input range, and change the confidence level to $1 - \alpha$.

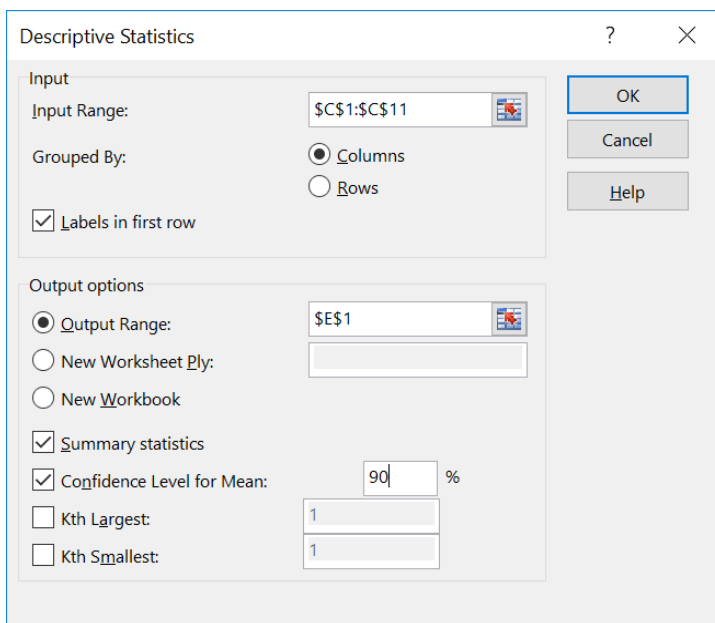


Figure 7-23 shows the Excel output. Excel only calculates the descriptive statistics with the margin of error.

<i>Weight</i>	
Mean	1.985
Standard Error	0.07439
Median	1.98
Mode	#N/A
Standard Deviation	0.235242
Sample Variance	0.055339
Kurtosis	-0.23332
Skewness	-0.46112
Range	0.76
Minimum	1.54
Maximum	2.3
Sum	19.85
Count	10
Confidence Level(90.0%)	0.136365

Figure 7-23

Use Excel to find each piece of the confidence interval $\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$.

Excel $t_{\alpha/2} = T.INV(0.1/2,9) = -1.83311$.

$$\begin{aligned} & \bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) \\ \Rightarrow & 1.985 \pm 1.83311 \left(\frac{0.235242}{\sqrt{10}} \right) \\ \Rightarrow & 1.985 \pm 1.83311(0.07439) \end{aligned}$$

Can you find the mean and standard error $\frac{s}{\sqrt{n}} = 0.07439$ in the Excel output in Figure 7-23?

$$\Rightarrow 1.985 \pm 0.136365$$

Can you find the margin of error $t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) = 0.136365$ in the Excel output in Figure 7-23?

Subtract and add the margin of error from the sample mean to get each confidence interval boundary. The answer can be in interval notation (1.8486, 2.1214) or standard notation $1.8486 < \mu < 2.1214$.

If we have raw data, Excel has a shortcut for both the critical value and p-value method.

Example 7-10 Continued:

Use the prior example, but this time use the **p-value method**. Again, let the significance level be $\alpha = 0.10$.

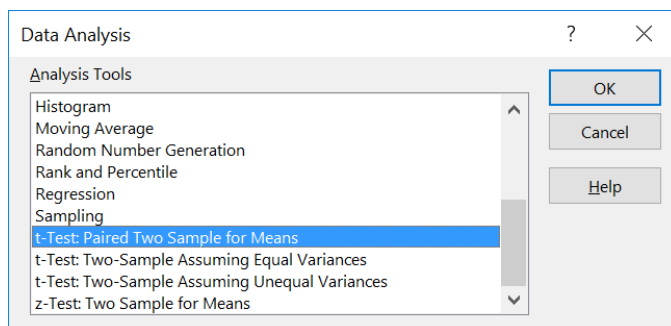
Solution:

Step 1: State the hypotheses. The hypotheses are: $H_0: \mu = 1.9$
 $H_1: \mu \neq 1.9$

Step 2: Compute the test statistic, $t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}} \right)} = \frac{1.985 - 1.9}{\left(\frac{0.235242}{\sqrt{10}} \right)} = 1.142625$.

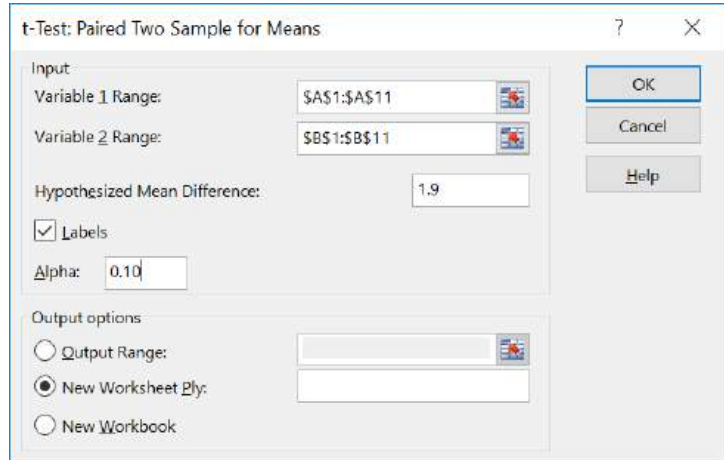
Verify using Excel. Excel does not have a one-sample t-test, but it does have a two-sample t-test that can be used with a dummy column of zeros as the second sample to get the results for just one sample. Copy over the data into cell A1. In column B, next to the data, type in a dummy column of zeros, and label it Dummy. (We frequently use placeholders in statistics called dummy variables.)

Select the Data Analysis tool and then select t-Test: Paired Two Sample for Means, then select OK.



	A	B	C
1	Weight	Dummy	
2	1.9	0	
3	2.24	0	
4	2.13	0	
5	2	0	
6	1.54	0	
7	1.96	0	
8	1.79	0	
9	2.18	0	
10	1.81	0	
11	2.3	0	
12			

For the Variable 1 Range select the data in cells A1:A11, including the label. For the Variable 2 Range select the dummy column of zeros in cells B1:B11, including the label. Change the hypothesized mean to 1.9. Check the Labels box and change the alpha value to 0.10, then select OK.



Excel provides the following output:

t-Test: Paired Two Sample for Means

	<i>Weight</i>	<i>Dummy</i>	What it Means:
Mean	1.985	0	Sample Mean
Variance	0.055339	0	Sample Variance
Observations	10	10	Sample Size
Pearson Correlation	#DIV/0!		Ignore this
Hypothesized Mean Difference	1.9		Hypothesized Mean
df	9		Degrees of Freedom
t Stat	1.142625		Test Statistic
P(T<=t) one-tail	0.14134		p-value for a one-tail test
t Critical one-tail	1.383029		Critical Value
P(T<=t) two-tail	0.282679		p-value for a two-tail test
t Critical two-tail	1.833113		±Critical Value

Step 3: Compute the p-value. Since the alternative hypothesis has a \neq symbol, use the Excel output next two-tailed p-value = 0.2826.

Step 4: Make the decision. For the p-value method we would compare the two-tailed p-value = 0.2826 to $\alpha = 0.10$. The rule is to reject H_0 if the p-value $\leq \alpha$. In this case the p-value $> \alpha$, therefore we fail to reject H_0 . Again, the same decision as the confidence interval method.

For the critical value method, we would compare the test statistic $t = 1.142625$ with the critical values for a two-tailed test $t_{\alpha/2} = \pm 1.833113$. Since the test statistic is between -1.8331 and 1.8331 we would fail to reject H_0 , which is the same decision using the p-value method or the confidence interval method.

Step 5: State the summary. At the 10% significance level, there is not enough evidence to support the claim that the population mean weight for all bumblebee bats is not equal to 1.9 g.

One-Tailed Versus Two-Tailed Tests

Most software packages do not ask which tailed test you are performing. Make sure you look at the sign in the alternative hypothesis to and determine which p-value to use. The difference is just what part of the picture you are looking at. In Excel, the critical value shown is for a one-tail test and does not specify left or right tail. The critical value in the output will always be positive, it is up to you to know if the critical value should be a negative or positive value. For example, Figures 7-24, 7-25, and 7-26 uses $df = 9$, $\alpha = 0.10$ to show all three tests comparing either the test statistic with the critical value or the p-value with α .

Two-Tailed Test

The test statistic can be negative or positive depending on what side of the distribution it falls; however, the p-value is a probability and will always be a positive number between 0 and 1. See Figure 7-24.

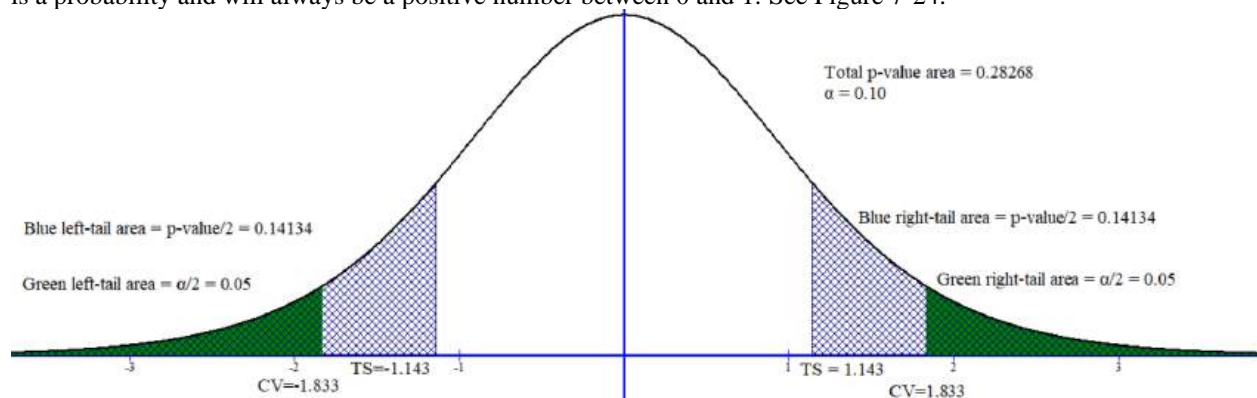


Figure 7-24

Right-Tailed Test

If we happened to do a right-tailed test with $df = 9$ and $\alpha = 0.10$, the critical value $t_{1-\alpha} = 1.383$ will be in the right tail and usually the test statistic will be a positive number. See Figure 7-25.

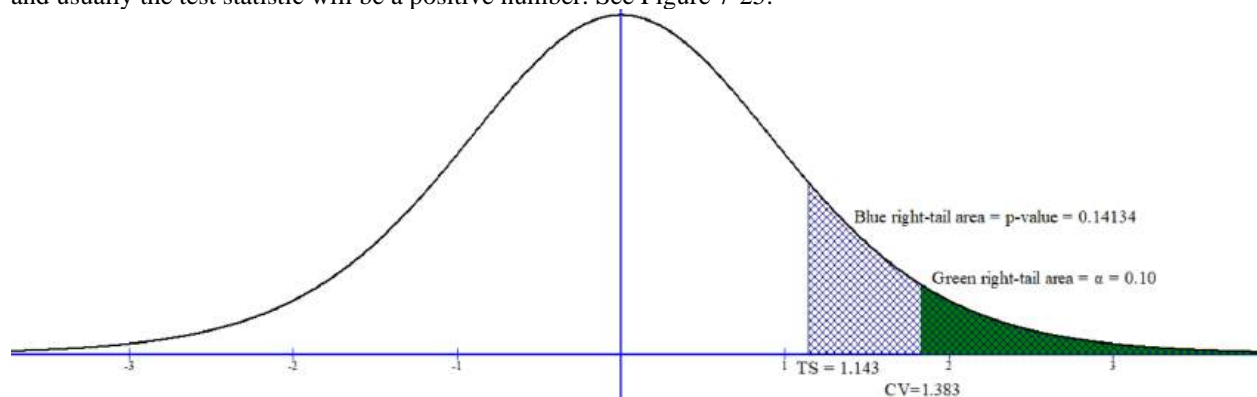


Figure 7-25

Left-Tailed Test

If we happened to do a left-tailed test with $df = 9$ and $\alpha = 0.10$, the critical value $t_{\alpha} = -1.383$ will be in the left tail and usually the test statistic will be a negative number. See Figure 7-26.

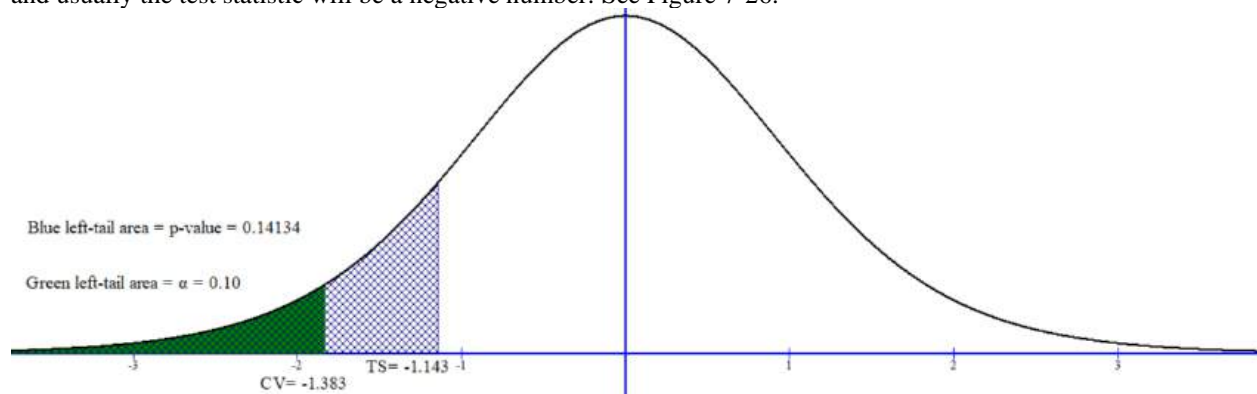


Figure 7-26

Chapter 7 Formulas

<p>Type I Error- Reject H_0 when H_0 is true.</p> <p>Type II Error- Fail to reject H_0 when H_0 is false.</p>	<p>Rejection Rules:</p> <ul style="list-style-type: none"> • P-value method: reject H_0 when the p-value $\leq \alpha$. • Critical value method: reject H_0 when the test statistic is in the critical region (shaded tails). • Confidence interval method: reject H_0 when the hypothesized value $\mu = \mu_0$ is not captured within the limits of the confidence interval.
<p>Hypothesis Test for One Proportion</p> <p>$H_0: p = p_0$ $H_1: p \neq p_0$ $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$</p> <p>TI-Calculator: 1-PropZTest</p>	<p>Hypothesis Test for One Mean</p> <p>$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ $t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$</p> <p>TI-Calculator: T-Test</p>
<p>z-Critical Values</p> <p>Excel: Two-tail: $z_{\alpha/2} = \text{NORM.INV}(1-(\alpha/2),0,1)$ Right-tail: $z_{1-\alpha} = \text{NORM.INV}(1-\alpha,0,1)$ Left-tail: $z_{\alpha} = \text{NORM.INV}(\alpha,0,1)$</p> <p>TI-Calculator: Two-tail: $z_{\alpha/2} = \text{invNorm}(1-(\alpha/2),0,1)$ Right-tail: $z_{1-\alpha} = \text{invNorm}(1-\alpha,0,1)$ Left-tail: $z_{\alpha} = \text{invNorm}(\alpha,0,1)$</p>	<p>t-Critical Values</p> <p>Excel: Two-tail: $t_{\alpha/2} = \text{T.INV}(1-(\alpha/2),df)$ Right-tail: $t_{1-\alpha} = \text{T.INV}(1-\alpha,df)$ Left-tail: $t_{\alpha} = \text{T.INV}(\alpha,df)$</p> <p>TI-Calculator: Two-tail: $t_{\alpha/2} = \text{invT}(1-(\alpha/2),df)$ Right-tail: $t_{1-\alpha} = \text{invT}(1-\alpha,df)$ Left-tail: $t_{\alpha} = \text{invT}(\alpha,df)$</p>

Chapter 7 Exercises

1. The plant-breeding department at a major university developed a new hybrid boysenberry plant called Stumptown Berry. Based on research data, the claim is made that from the time shoots are planted 90 days on average are required to obtain the first berry. A corporation that is interested in marketing the product tests 60 shoots by planting them and recording the number of days before each plant produces its first berry. The sample mean is 92.3 days. The corporation wants to know if the mean number of days is different from the 90 days claimed. Which one is the correct set of hypotheses?

- a) $H_0: p = 90\%$ $H_1: p \neq 90\%$
 b) $H_0: \mu = 90$ $H_1: \mu \neq 90$
 c) $H_0: p = 92.3\%$ $H_1: p \neq 92.3\%$
 d) $H_0: \mu = 92.3$ $H_1: \mu \neq 92.3$
 e) $H_0: \mu \neq 90$ $H_1: \mu = 90$

2. Match the symbol with the correct phrase.

<	More than
>	At most
\leq	Reduced
\geq	Different from
\neq	At least

3. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, 23% of all complaints in 2007 were for identity theft. In that year, Alaska had 321 complaints of identity theft out of 1,432 consumer complaints. Does this data provide enough evidence to show that Alaska had a lower proportion of identity theft than 23%? Which one is the correct set of hypotheses?

Federal Trade Commission, (2008). *Consumer fraud and identity theft complaint data: January-December 2007*. Retrieved from website: <http://www.ftc.gov/opa/2008/02/fraud.pdf>.

- a) $H_0: p = 23\%$ $H_1: p < 23\%$
 b) $H_0: \mu = 23$ $H_1: \mu < 23$
 c) $H_0: p < 23\%$ $H_1: p \geq 23\%$
 d) $H_0: p = 0.224$ $H_1: p < 0.224$
 e) $H_0: \mu < 0.224$ $H_1: \mu \geq 0.224$
4. Compute the z critical value for a right-tailed test when $\alpha = 0.01$.
5. Compute the z critical value for a two-tailed test when $\alpha = 0.01$.
6. Compute the z critical value for a left-tailed test when $\alpha = 0.05$.
7. Compute the z critical value for a two-tailed test when $\alpha = 0.05$.
8. As of 2018, the Centers for Disease Control and Protection's (CDC) national estimate that 1 in 68 ≈ 0.0147 children have been diagnosed with autism spectrum disorder (ASD). A researcher believes that the proportion of children in their county is different from the CDC estimate. Which one is the correct set of hypotheses?

- a) $H_0: p = 0.0147$ $H_1: p \neq 0.0147$
- b) $H_0: \mu = 0.0147$ $H_1: \mu \neq 0.0147$
- c) $H_0: p \neq 0.0147$ $H_1: p = 0.0147$
- d) $H_0: \mu = 68$ $H_1: \mu \neq 68$
- e) $H_0: = 0.0147$ $H_1: \neq 0.0147$

9. Match the phrase with the correct symbol.

- | | |
|----------------------------------|-----------------|
| a. Sample Size | i. α |
| b. Population Mean | ii. n |
| c. Sample Variance | iii. σ^2 |
| d. Sample Mean | iv. s^2 |
| e. Population Standard Deviation | v. s |
| f. P(Type I Error) | vi. \bar{x} |
| g. Sample Standard Deviation | vii. σ |
| h. Population Variance | viii. μ |

10. The Food & Drug Administration (FDA) regulates that fresh albacore tuna fish contains at most 0.82 ppm of mercury. A scientist at the FDA believes the mean amount of mercury in tuna fish for a new company exceeds the ppm of mercury. Which one is the correct set of hypotheses?

- a) $H_0: p = 82\%$ $H_1: p > 82\%$
- b) $H_0: \mu = 0.82$ $H_1: \mu > 0.82$
- c) $H_0: p > 82\%$ $H_1: p \leq 82\%$
- d) $H_0: \mu = 0.82$ $H_1: \mu \neq 0.82$
- e) $H_0: \mu > 0.82$ $H_1: \mu \leq 0.82$

11. Match the symbol with the correct phrase.

$100(1 - \alpha)\%$	Parameter
$1 - \beta$	P(Type II Error)
β	Power
μ	Significance Level
α	Confidence Level

12. A 2019 survey by the Bureau of Labor Statistics reported that 92% of Americans working in large companies have paid leave. In January 2021, a random survey of workers showed that 89% had paid leave. The resulting p-value is 0.009; thus, the null hypothesis is rejected. It is concluded that there has been a decrease in the proportion of people, who have paid leave from 2019 to January 2021. What type of error is possible in this situation?

- a) Type I Error
- b) Type II Error
- c) Standard Error
- d) Margin of Error
- e) No error was made.

13. The SAT exam in previous years is normally distributed with an average score of 1,000 points. The test writers for this upcoming year want to make sure that the new test does not have a significantly different mean score. A sample of 20 students take the new SAT exam and their mean score was 1,050 points with a standard deviation of 150 points.
- State the hypotheses to test to see if the mean time has significantly changed using a 5% level of significance.
 - What is a type I error for this problem?
 - What is a type II error for this problem?
14. The plant-breeding department at a major university developed a new hybrid boysenberry plant called Stumptown Berry. Based on research data, the claim is made that from the time shoots are planted 90 days on average are required to obtain the first berry. A corporation that is interested in marketing the product tests 60 shoots by planting them and recording the number of days before each plant produces its first berry. The sample mean is 92.3 days. The corporation will not market the product if the mean number of days is more than the 90 days claimed. The hypotheses are $H_0: \mu = 90$ $H_1: \mu > 90$. Which answer is the correct type I error in the context of this problem?
- The corporation will not market the Stumptown Berry even though the berry does produce fruit within the 90 days.
 - The corporation will market the Stumptown Berry even though the berry does produce fruit within the 90 days.
 - The corporation will not market the Stumptown Berry even though the berry does produce fruit in more than 90 days.
 - The corporation will market the Stumptown Berry even though the berry does produce fruit in more than 90 days.
15. The Food & Drug Administration (FDA) regulates that fresh albacore tuna fish contains at most 0.82 ppm of mercury. A scientist at the FDA believes the mean amount of mercury in tuna fish for a new company exceeds the ppm of mercury. The hypotheses are $H_0: \mu = 0.82$ $H_1: \mu > 0.82$. Which answer is the correct type II error in the context of this problem?
- The fish is rejected by the FDA when in fact it had less than 0.82 ppm of mercury.
 - The fish is accepted by the FDA when in fact it had less than 0.82 ppm of mercury.
 - The fish is rejected by the FDA when in fact it had more than 0.82 ppm of mercury.
 - The fish is accepted by the FDA when in fact it had more than 0.82 ppm of mercury.
16. A two-tailed z -test found a test statistic of $z = 2.153$. At a 1% level of significance, which would the correct decision?
- Do not reject H_0
 - Reject H_0
 - Accept H_0
 - Reject H_1
 - Do not reject H_1
17. A left-tailed z -test found a test statistic of $z = -1.99$. At a 5% level of significance, what would the correct decision be?
- Do not reject H_0
 - Reject H_0
 - Accept H_0
 - Reject H_1
 - Do not reject H_1

18. A right-tailed z -test found a test statistic of $z = 0.05$. At a 5% level of significance, what would the correct decision be?
- Reject H_0
 - Accept H_0
 - Reject H_1
 - Do not reject H_0
 - Do not reject H_1
19. A two-tailed z -test found a test statistic of $z = -2.19$. At a 1% level of significance, which would the correct decision?
- Do not reject H_0
 - Reject H_0
 - Accept H_0
 - Reject H_1
 - Do not reject H_1
20. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, 23% of all complaints in 2007 were for identity theft. In that year, Alaska had 321 complaints of identity theft out of 1,432 consumer complaints. Does this data provide enough evidence to show that Alaska had a lower proportion of identity theft than 23%? The hypotheses are $H_0: p = 23\%$ $H_1: p < 23\%$. Which answer is the correct type I error in the context of this problem? Federal Trade Commission, (2008). *Consumer fraud and identity theft complaint data: January-December 2007*. Retrieved from website: <http://www.ftc.gov/opa/2008/02/fraud.pdf>.
- It is believed that less than 23% of Alaskans had identity theft and there really was 23% or less that experienced identity theft.
 - It is believed that more than 23% of Alaskans had identity theft and there really was 23% or more that experience identity theft.
 - It is believed that less than 23% of Alaskans had identity theft even though there really was 23% or more that experienced identity theft.
 - It is believed that more than 23% of Alaskans had identity theft even though there really was less than 23% that experienced identity theft
21. A hypothesis test was conducted during a clinical trial to see if a new COVID-19 vaccination reduces the risk of contracting the virus. What is the Type I and II errors in terms of approving the vaccine for use?
22. A manufacturer of rechargeable laptop batteries claims its batteries have, on average, 500 charges. A consumer group decides to test this claim by assessing the number of times 30 of their laptop batteries can be recharged and finds a p-value is 0.1111; thus, the null hypothesis is not rejected. What is the Type II error for this situation?
23. A commonly cited standard for one-way length (duration) of school bus rides for elementary school children is 30 minutes. A local government office in a rural area conducts a study to determine if elementary schoolers in their district have a longer average one-way commute time. If they determine that the average commute time of students in their district is significantly higher than the commonly cited standard, they will invest in increasing the number of school buses to help shorten commute time. What would a Type II error mean in this context?
24. The Centers for Disease Control and Prevention (CDC) 2018 national estimate that 1 in 68 ≈ 0.0147 children have been diagnosed with autism spectrum disorder (ASD). A researcher believes that the proportion of children in their county is different from the CDC estimate. The hypotheses are $H_0: p = 0.0147$ $H_1: p \neq 0.0147$. Which answer is the correct type II error in the context of this problem?
- The proportion of children diagnosed with ASD in the researcher's county is believed to be different from the national estimate, even though the proportion is the same.

- b) The proportion of children diagnosed with ASD in the researcher's county is believed to be different from the national estimate and the proportion is different.
 - c) The proportion of children diagnosed with ASD in the researcher's county is believed to be the same as the national estimate, even though the proportion is different.
 - d) The proportion of children diagnosed with ASD in the researcher's county is believed to be the same as the national estimate and the proportion is the same.
25. You are conducting a study to see if the accuracy rate for fingerprint identification is significantly different from 0.34. Thus, you are performing a two-tailed test. Your sample data produce the test statistic $z = 2.504$. Use your calculator to find the p-value and state the correct decision and summary.

For exercises 26-31, show all 5 steps for hypothesis testing:

- a) State the hypotheses.
 - b) Compute the test statistic.
 - c) Compute the critical value or p-value.
 - d) State the decision.
 - e) Write a summary.
26. Test the claim that the proportion of people who own dogs is less than 32%. A random sample of 1,000 people found that 28% owned dogs. Do the sample data provide convincing evidence to support the claim? Test the relevant hypotheses using a 10% level of significance.
27. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.3% of American adults suffer from depression or a depressive illness. Suppose that in a survey of 2,000 people in a certain city, 11.1% of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that city suffering from depression or a depressive illness is more than the 9.3% in the general adult American population. Test the relevant hypotheses using a 5% level of significance.
28. The United States Department of Energy reported that 48% of homes were heated by natural gas. A random sample of 333 homes in Oregon found that 149 were heated by natural gas. Test the claim that the proportion of homes in Oregon that were heated by natural gas is different from what was reported. Use a 1% significance level.
29. You are conducting a study to see if the proportion of men over the age of 50 who regularly have their prostate examined is significantly less than 0.31. A random sample of 735 men over the age of 50 found that 208 have their prostate regularly examined. Do the sample data provide convincing evidence to support the claim? Test the relevant hypotheses using a 5% level of significance.
30. Nationwide 40.1% of employed teachers are union members. A random sample of 250 Oregon teachers showed that 110 belonged to a union. At $\alpha = 0.10$, is there sufficient evidence to conclude that the proportion of union membership for Oregon teachers is higher than the national proportion?
31. Nationally the percentage of adults that have their teeth cleaned by a dentist yearly is 64%. A dentist in Portland, Oregon believes that regionally the percent is higher. A sample of 2,000 Portlanders found that 1,312 had their teeth cleaned by a dentist in the last year. Test the relevant hypotheses using a 10% level of significance.
32. Compute the t critical value for the following.
- a) A left-tailed test when $\alpha = 0.10$ and $df = 10$.
 - b) A right-tailed test when $\alpha = 0.05$ and $n = 20$.
 - c) A two-tailed test when $\alpha = 0.05$ with a sample size of 18.
 - d) A two-tailed test when $\alpha = 0.01$ with a sample size of 29.

33. A student is interested in becoming an actuary. They know that becoming an actuary takes a lot of schooling and they will have to take out student loans. They want to make sure the starting salary will be higher than \$55,000/year. They randomly sample 30 starting salaries for actuaries and find a p-value of 0.0392. Use $\alpha = 0.05$.
- a) Choose the correct hypotheses.
- $H_0: \mu = 55,000$ $H_1: \mu < 55,000$
 - $H_0: \mu > 55,000$ $H_1: \mu \leq 55,000$
 - $H_0: \mu = 55,000$ $H_1: \mu > 55,000$
 - $H_0: \mu < 55,000$ $H_1: \mu \geq 55,000$
 - $H_0: \mu = 55,000$ $H_1: \mu \neq 55,000$
- b) Should the student pursue an actuary career?
- Yes, since we reject the null hypothesis.
 - Yes, since we reject the claim.
 - No, since we reject the claim.
 - No, since we reject the null hypothesis.
34. Honda advertises the 2024 Honda Civic as getting 34 mpg for city driving. A skeptical consumer about to purchase this model believes the mpg is less than the advertised amount and randomly selects 35, of the 2024 Honda Civic owners, and asks them what their car's mpg is. Use a 1% significance level. They find a p-value of 0.0436.
- a) Choose the correct hypotheses.
- $H_0: \mu = 34$ $H_1: \mu < 34$
 - $H_0: \mu < 34$ $H_1: \mu \geq 34$
 - $H_0: \mu = 34$ $H_1: \mu > 34$
 - $H_0: \mu = 35$ $H_1: \mu \neq 35$
 - $H_0: \mu = 34$ $H_1: \mu \neq 34$
- b) Choose the correct decision based off the reported p-value.
- Reject H_0
 - Do not reject H_0
 - Do not reject H_1
 - Reject H_1
35. The Food & Drug Administration (FDA) regulates that fresh albacore tuna fish contains at most 0.82 ppm of mercury. A scientist at the FDA believes the mean amount of mercury in tuna fish for a new company exceeds the ppm of mercury. A test statistic was found to be 2.576 and a critical value was found to be 1.708, what is the correct decision and summary?
- Reject H_0 , there is enough evidence to support the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.
 - Accept H_0 , there is not enough evidence to reject the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.
 - Reject H_1 , there is not enough evidence to reject the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.
 - Reject H_0 , there is not enough evidence to support the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.
 - Do not reject H_0 , there is not enough evidence to support the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.

36. The plant-breeding department at a major university developed a new hybrid boysenberry plant called Stumptown Berry. Based on research data, the claim is made that from the time shoots are planted 90 days on average are required to obtain the first berry. A corporation that is interested in marketing the product tests 60 shoots by planting them and recording the number of days before each plant produces its first berry. The corporation wants to know if the mean number of days is different from the 90 days claimed. A random sample was taken and the following test statistic was $t = -2.15$ and critical values of $t = \pm 1.86$ were found. What is the correct decision and summary?
- Do not reject H_0 , there is not enough evidence to support the corporation's claim that the mean number of days until a berry is produced is different from the 90 days claimed by the university.
 - Reject H_0 , there is enough evidence to support the corporation's claim that the mean number of days until a berry is produced is different from the 90 days claimed by the university.
 - Accept H_0 , there is enough evidence to support the corporation's claim that the mean number of days until a berry is produced is different from the 90 days claimed by the university.
 - Reject H_1 , there is not enough evidence to reject the corporation's claim that the mean number of days until a berry is produced is different from the 90 days claimed by the university.
 - Reject H_0 , there is not enough evidence to support the corporation's claim that the mean number of days until a berry is produced is different from the 90 days claimed by the university.

For exercises 37-42, show all 5 steps for hypothesis testing:

- State the hypotheses.
 - Compute the test statistic.
 - Compute the critical value or p-value.
 - State the decision.
 - Write a summary.
37. The total of individual pounds of garbage discarded by 17 households in one week is shown below. The current waste removal system company has a weekly maximum weight policy of 36 pounds. Test the claim that the average weekly household garbage weight is less than the company's weekly maximum. Use a 5% level of significance. Assume the population of garbage weights are approximately normally distributed.

Weight				
34.5	32.9	42.9	32.9	31.8
40	33.8	35.8	35.4	30.5
31.4	39.2	26.8	30.6	34.5
34.7	32.8			

38. The workweek for adults in the United States work full-time is normally distributed with a mean of 47 hours. A newly hired engineer at a start-up company believes that employees at her start-up company work more on average than working adults in the U.S. The newly hired engineer asks 12 engineering co-workers for the lengths in hours of their workweek. Their responses are shown in the table below. Test the claim using a 5% level of significance. Assume the population of working hours is approximately normally distributed.

Hours	46	42	54	52	48	45	49	49	50	46	55	55
-------	----	----	----	----	----	----	----	----	----	----	----	----

39. The average number of calories from a fast food meal for adults in the United States is 842 calories. A nutritionist believes that the average is higher than reported. They sample 11 meals that adults ordered and measure the calories for each meal shown below. Test the claim using a 5% level of significance. Assume that fast food calories are normally distributed.

Calories	855	854	785	854	952	860	853	760	862	851	919
----------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

40. The world's smallest mammal is the bumblebee bat (also known as Kitti's hog-nosed bat or *Craseonycteris thonglongyai*). Such bats are roughly the size of a large bumblebee. A sample of 10 bats weighed in grams are

shown below. Test the claim that mean weight for all bumblebee bats is not equal to 2.1 g using a 1% level of significance. Assume that the bat weights are normally distributed.

Weight	2.22	1.6	1.78	1.52	1.61	1.98	1.56	2.24	1.55	2.28
--------	------	-----	------	------	------	------	------	------	------	------

41. A sample of 45 body temperatures of athletes had a mean of 98.8°F and a standard deviation of 0.62°F. Test the claim that the mean body temperature for all athletes is more than 98.6°F. Use a 1% level of significance.
42. The average age of an adult's first vacation without a parent or guardian was reported to be 23 years old. A travel agent believes that the average age is different from what was reported. They sample 28 adults and they asked their age in years when they first vacationed as an adult without a parent or guardian, data shown below. Test the claim using a 10% level of significance. Assume the population ages are approximately normally distributed.

Age						
21	22	25	26	24	27	22
23	24	27	20	18	24	22
22	28	26	25	26	22	23
24	27	25	22	21	24	22

“For instance, a race of hyperintelligent pan-dimensional beings once built themselves a gigantic supercomputer called Deep Thought to calculate once and for all the Answer to the Ultimate Question of Life, the Universe, and Everything.

For seven and a half million years, Deep Thought computed and calculated, and in the end announced that the answer was in fact Forty-two - and so another, even bigger, computer had to be built to find out what the actual question was.”
 (Adams, 2002)

Chapter 8

Hypothesis Tests & Confidence Intervals for Two Populations



- 8.1 Introduction to Inference for Two Populations
- 8.2 Two Proportions Hypothesis Test & Confidence Interval
- 8.3 Two Dependent Means Hypothesis Test & Confidence Interval
- 8.4 Two Independent Means Hypothesis Test & Confidence Interval
 - 8.4.1 Unequal Variance Method
 - 8.4.2 Equal Variance Method

8.1 Introduction to Inference For Two Populations

There are many instances where researchers wish to compare two groups. A clinical trial may want to use a control group and an experiment group to see if a new medication is effective. Identical twin studies help geneticists learn more about inherited traits. Educators may want to test to see if there is a difference between before and after test scores. A farmer may wish to see if there is a difference between two types of fertilizer. A marketing firm may want to see if there is a preference between two different bottle designs. Hypothesis testing for two groups takes on similar steps as one group. It is important to know if the two groups are dependent (related) or independent (not related) from one another.

In order to use formulas that compare the proportions or means from two populations, we use subscripts to show which sample statistic or population parameter we are referencing.

Statistics

n_1 = sample size from population 1

n_2 = sample size from population 2

\hat{p}_1 = proportion of sample from population 1

\hat{p}_2 = proportion of sample from population 2

\bar{x}_1 = mean of sample from population 1

\bar{x}_2 = mean of sample from population 2

s_1 = standard deviation of sample from population 1

s_2 = standard deviation of sample from population 2

s_1^2 = variance of sample from population 1

s_2^2 = variance of sample from population 2

Parameters

p_1 = population proportion of population 1

p_2 = population proportion of population 2

μ_1 = population mean of population 1

μ_2 = population mean of population 2

σ_1 = population standard deviation of population 1

σ_2 = population standard deviation of population 2

σ_1^2 = population variance of population 1

σ_2^2 = population variance of population 2

You do not need to use the subscripts 1 and 2. You may use a letter or symbol that helps you differentiate between the two groups. For instance, if you have two manufacturers labeled A and B, you may want to use μ_A and μ_B .

Most of the time the groups are numbered from the order in which their statistics or data first appear in the problem. To keep the correct sign of the test, make sure you do not switch the order of the groups.

For instance, if we were comparing the mean Scholastic Assessment Test (SAT) score between high school juniors and seniors and our hypothesis is that the mean for seniors is higher, we could set up the alternative hypotheses as either $\mu_j < \mu_s$ if we had the juniors be group 1 and $\mu_j > \mu_s$ if we had the seniors be group 1. This change would switch the sign of both the test statistic and the critical value.

When performing a one-tailed test, most of the time the sign of the test statistic and critical value will match. For example, if your test statistic came out to be $z = -1.567$ and your critical value was $z = 1.645$ you most likely have the incorrect order in your hypotheses.

8.2 Two Proportion Hypothesis Test & Confidence Interval

This section will look at how to analyze a difference in the proportions for two independent samples. As with all other hypothesis tests and confidence intervals, the process of testing is the same, though the formulas and assumptions are different.

There are three ways to set up the hypotheses for comparing the difference in two population proportions $p_1 - p_2$, see Figure 8-1.

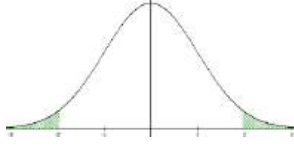
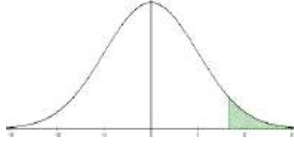
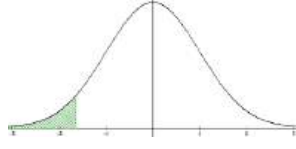
Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$	$H_0: p_1 = p_2$ $H_1: p_1 > p_2$	$H_0: p_1 = p_2$ $H_1: p_1 < p_2$
		
$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 \neq 0$	$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 > 0$	$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 < 0$

Figure 8-1

When the two proportions are assumed to be equal, $p_1 = p_2$, if we were to subtract from both sides of the equality, $p_1 - p_2 = p_2 - p_2$, we would get $p_1 - p_2 = 0$. Note that for our purposes we will always use, $p_1 - p_2 = 0$. We could also use a variant of this model to test the difference in the proportions $p_1 - p_2$ is set to a number other than zero for a known magnitude in the difference, but we will not cover that scenario.

The **2 proportion z-test** is a statistical test for comparing the proportions from two populations. It can be used when the samples are independent, $n_1\hat{p}_1 \geq 10$, $n_1\hat{q}_1 \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2\hat{q}_2 \geq 10$. The formula for the z-test statistic is:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p} \cdot \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where $\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)} = \frac{(n_1\hat{p}_1 + n_2\hat{p}_2)}{(n_1 + n_2)}$, $\hat{q} = 1 - \hat{p}$, $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$.

The pooled proportion \hat{p} is a weighted mean of the proportions and \hat{q} is the complement of \hat{p} . Some texts or software may use different notation for the pooled proportion, for example \hat{p} or \bar{p} .

Example 8-1: A vice principal wants to see if there is a difference between the number of students who are late to class for the first class of the day compared to the student's class right after lunch. To test their claim to see if there is a difference in the proportion of late students between first class of the day and after lunch classes, the vice-principal randomly selects 200 students from first class and records if they are late, then randomly selects 200 students in their class after lunch and records if they are late. At the 0.05 level of significance, can a difference be concluded?

	First Class	After Lunch Class
Sample size	200	200
Number of late students	13	16

Solution:

Assumptions: We are comparing the proportion of late students' first and after lunch classes. The number of "successes" and "failures" from each population must be greater than 10 ($n_1\hat{p}_1 = 13 \geq 10$, $n_1\hat{q}_1 = 187 \geq 10$, $n_2\hat{p}_2 = 16 \geq 10$, and $n_2\hat{q}_2 = 184 \geq 10$). We can assume that the samples were independent since they are randomly selected.

Example 8-1 Using the Traditional Method

The claim is that there is a difference between the proportion of late students. Let population 1 be the first class, and population 2 be the class after lunch. Our claim would then be $p_1 \neq p_2$.

The correct hypotheses are: $H_0: p_1 = p_2$
 $H_1: p_1 \neq p_2$.

Compute the $z_{\alpha/2}$ critical values. Draw and label the sampling distribution.

Use the inverse normal function $\text{invNorm}(0.025,0,1)$ to get the critical values $z_{\alpha/2} = \pm 1.96$. See Figure 8-2.

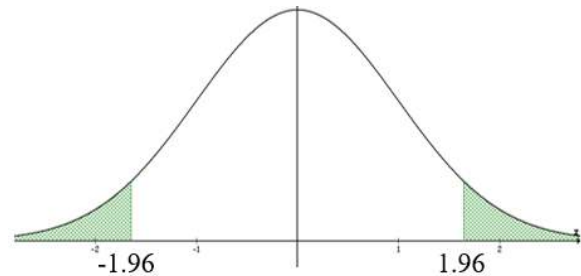


Figure 8-2

In order to compute the test statistic, we must first compute the following proportions:

$$\hat{p} = \frac{(x_1+x_2)}{(n_1+n_2)} = \frac{(13+16)}{(200+200)} = 0.0725 \qquad \hat{q} = 1 - \hat{p} = 1 - 0.0725 = 0.9275$$

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{13}{200} = 0.065 \qquad \hat{p}_2 = \frac{x_2}{n_2} = \frac{16}{200} = 0.08$$

The test statistic is, $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p} \cdot \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(0.065 - 0.08) - 0}{\sqrt{(0.0725 \cdot 0.9275) \left(\frac{1}{200} + \frac{1}{200} \right)}} = -0.5784$.

Decision: Because the test statistic is between the critical values, we do not reject H_0 .

Summary: There is not enough evidence to support any difference in the proportion of students that are late for their first class compared to the class after lunch.

Example 8-1 Using the p-Value Method

The correct hypotheses are: $H_0: p_1 = p_2$
 $H_1: p_1 \neq p_2$.

Compute the test statistic, we must first compute the following proportions:

$$\hat{p} = \frac{(x_1+x_2)}{(n_1+n_2)} = \frac{(13+16)}{(200+200)} = 0.0725 \qquad \hat{q} = 1 - \hat{p} = 1 - 0.0725 = 0.9275$$

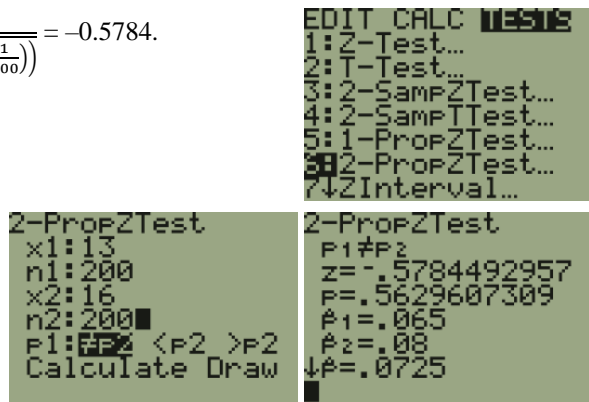
$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{13}{200} = 0.065 \qquad \hat{p}_2 = \frac{x_2}{n_2} = \frac{16}{200} = 0.08$$

The test statistic is, $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p} \cdot \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(0.065 - 0.08) - 0}{\sqrt{(0.0725 \cdot 0.9275) \left(\frac{1}{200} + \frac{1}{200} \right)}} = -0.5784$.

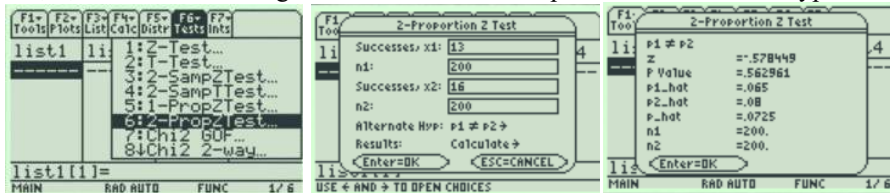
Use technology to compute the p-value = 0.563.

Using the TI Calculators:

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [6:2-PropZTest] and press the [ENTER] key. Type in the x_1 , n_1 , x_2 , and n_2 arrow over to the \neq , $<$, $>$ sign that is the same in the problem's alternative hypothesis statement, then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the z-test statistic and the p-value.



TI-89: Go to the [Apps] **Stat/List Editor**, then press [2nd] then F6 [Tests], then select **6: 2-PropZTest**. Type in the x_1 , n_1 , x_2 , and n_2 arrow over to the \neq , $<$, $>$ and select the sign that is the same in the problem's alternative hypothesis statement. Press the [ENTER] key to calculate. The calculator returns the z-test statistic, sample proportions, pooled proportion, and the p-value.



Excel: There is no shortcut key for a 2 Proportion Hypothesis Test in Excel. However, recall that the p-value is the probability of seeing the test statistic or more extreme, given the null hypothesis is true. Since this is a two tailed test, we can use the NORM.S.DIST function. The p-value would be $2 * P(Z \leq -0.578449) = 2 * \text{NORM.S.DIST}(-0.578449, \text{TRUE}) = 0.563$.

Decision: Since the p-value $> \alpha$, we do not reject H_0 .

Summary: There is not enough evidence to support any difference in the proportion of students that are late for their first class compared to the class after lunch.

Example 8-2: A superintendent of a school district in Portland wants to compare their student preference of chocolate milk over regular milk to students in Tucson. A random sample of 273 students in Portland found that 187 preferred chocolate milk. A random sample of 665 students in Tucson found that 475 preferred chocolate milk. Test the claim that the proportion of all students in Portland is less than the proportion of all students in Tucson that prefer chocolate milk. Use $\alpha = 0.10$.

Solution: The correct hypotheses are: $H_0: p_1 = p_2$
 $H_1: p_1 < p_2$.

Compute the test statistic, we must first compute the following proportions:

$$\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)} = \frac{(187 + 475)}{(273 + 665)} = 0.705757 \quad \hat{q} = 1 - \hat{p} = 1 - 0.705757 = 0.294243$$

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{187}{273} = 0.684982 \quad \hat{p}_2 = \frac{x_2}{n_2} = \frac{475}{665} = 0.714286$$

$$\text{The test statistic is, } z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p} \cdot \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(0.684982 - 0.714286) - 0}{\sqrt{(0.705757 \cdot 0.294243) \left(\frac{1}{273} + \frac{1}{665} \right)}} = -0.8946.$$

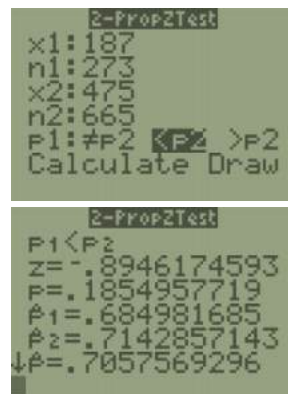
Use technology to compute the p-value = 0.1855.

Since the p-value $> \alpha$, we fail to reject H_0 . At the 10% significance level, there is not enough evidence to support the claim that the proportion of Portland students that prefer chocolate milk is less than Tucson students.

Two Proportions Z-Interval

When comparing two proportions, we are often interested in determining if there is a significant difference between them. The confidence interval helps us answer this question by providing a range of values that encompasses the likely difference between the two population proportions.

Interpreting the confidence interval involves examining whether the interval captures zero. If the interval captures zero, it suggests that there may not be a significant difference between the two proportions. On the other hand, if the interval does not capture zero, there is evidence of a statistically significant difference in the proportions.



A $100(1 - \alpha)\%$ confidence interval for the difference between two population proportions $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\left(\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}\right)} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\left(\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}\right)}$$

Or more compactly as $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\left(\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}\right)}$

The requirements are identical to the 2-proportion hypothesis test, the samples are independent, $n_1 \hat{p}_1 \geq 10$, $n_1 \hat{q}_1 \geq 10$, $n_2 \hat{p}_2 \geq 10$, and $n_2 \hat{q}_2 \geq 10$. Note that the standard error does not rely on a pooled proportion that we use for the hypothesized proportions being equal, so be cautious using the confidence interval to make decisions based on a hypothesis statement.

Example 8-3: Calculate the 95% confidence interval for the difference in the proportion of late students in their first class and the proportion who are late to their class after lunch.

	First Class	After Lunch Class
Sample size	200	200
Number of late students	13	16

Solution: First, compute the following:

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{13}{200} = 0.065 \quad \hat{q}_1 = 1 - \hat{p}_1 = 1 - 0.065 = 0.935$$

$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{16}{200} = 0.08 \quad \hat{q}_2 = 1 - \hat{p}_2 = 1 - 0.08 = 0.92$$

Calculate the $z_{\alpha/2}$ critical values. Use the inverse normal to get $z_{\alpha/2} = \pm 1.96$.

Now substitute the numbers into the interval estimate: $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\left(\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}\right)}$

$$\Rightarrow (0.065 - 0.08) \pm 1.96 \sqrt{\left(\frac{0.065 \cdot 0.935}{200} + \frac{0.08 \cdot 0.92}{200}\right)}$$

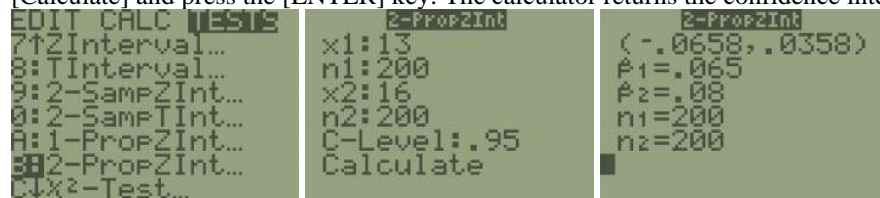
$$\Rightarrow -0.015 \pm 0.0508$$

$$\Rightarrow (-0.0658, 0.0358).$$

Use interval notation $(-0.0658, 0.0358)$ or standard notation $-0.0658 < p_1 - p_2 < 0.0358$. Note that we can have negative numbers here since we are taking the difference of two proportions.

Since $p_1 - p_2 = 0$ is in the interval, we are 95% confident that there is no difference in the proportion of late students between their first class or those who are late for their class after lunch.

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [2-PropZInterval] and press the [ENTER] key. Type in the x_1 , n_1 , x_2 , n_2 , the confidence level, then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the confidence interval.



TI-89: Go to the [Apps] Stat/List Editor, then press [2nd] then F7 [Ints], then select **6: 2-PropZInt**. Type in the x_1 , n_1 , x_2 , n_2 , the confidence level, then press the [ENTER] key to calculate. The calculator returns the confidence interval.

8.3 Two Dependent Means Hypothesis Test & Confidence Interval

Dependent samples occur when the subjects are paired up, or matched in some way. Most often, this model is characterized by selection of a random sample where each member is observed under two different conditions, before/after some experiment or treatment, or subjects that are similar (matched) to each other are studied under two different conditions. The dependent t-test for means is often referred to as repeated samples or **matched pairs**.

The hypothesis test for comparing two **dependent** population means μ_1 and μ_2 , where, μ_D is the expected difference of the matched pairs can be set up as a two-tailed, right-tailed or left-tailed test. See Figure 8-3.

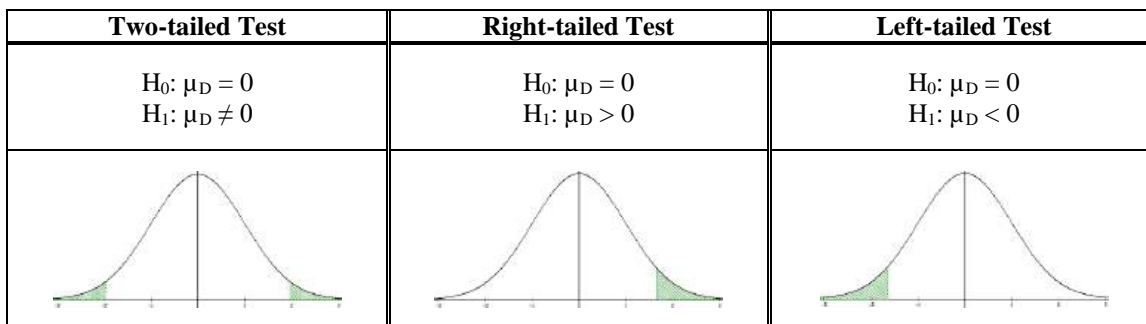


Figure 8-3

Note: If each pair were equal to one another then the mean of the differences would be zero. We could also use this model to test with a magnitude of a difference, but we rarely cover that scenario, therefore we are usually test against the difference of zero. We will usually only use the case where μ_D equals zero.

The t-test for dependent samples is a statistical test for comparing the means from two dependent populations (or the difference between the means from two populations). The t-test is used when the differences are normally distributed. The two samples also must be dependent.

The formula for the t-test statistic is: $t = \frac{\bar{D} - \mu_D}{\left(\frac{s_D}{\sqrt{n}}\right)}$.

Where the t-distribution with degrees of freedom, $df = n - 1$.

The subscript “D” denotes the difference between population one and two. It is important to compute $D = x_1 - x_2$ for each pair of observations. However, this makes setting up the hypotheses more challenging for one-tailed tests.

If we were looking for an increase in test scores from before to after, then we would expect the after score to be larger. When we take a smaller number minus a larger number then the difference would be negative. If we put the before group first and the after group second then we would need a left-tailed test $\mu_D < 0$ to test the “increase” in test scores. This is opposite of the sign we associate for “increase.” If we swap the order and use the after group first, then the before group would have a larger number minus a smaller number which would be positive and we would do a right-tailed test $\mu_D > 0$.

Always subtract in the same order the data is presented in the question. An easier way to decide on the one-tailed test is to write down the two labels and then put a less than (<) or greater than (>) symbol between them depending on the question. For example, if the research statement is a weight loss program significantly decreases the average weight, the sign of the test would change depending on which group came first. If we subtract before weight – after weight, then we would want to have before > after and use $\mu_D > 0$. If we have the after weight as the first measurement then we would subtract the after weight – before weight and want after < before and use $\mu_D < 0$. If you keep your labels in the same order as they appear in the question, compare them and carry this sign down to the alternative hypothesis.

The traditional method (critical value method), the p-value method, or the confidence interval method are performed with steps that are identical to those when performing hypothesis tests for one population.

Example 8-4: A dietician is testing to see if a new diet program reduces the average weight. They randomly sample 35 patients and measure them before they start the program and then weigh them again after 2 months on the program. What are the correct hypotheses?

Solution: Let x_1 = weight before a weight-loss program and x_2 = weight after the weight-loss program. We want to test if, on average, participants lose weight. Therefore, the difference $D = x_1 - x_2$. This gives D = before weight – after weight, thus if on average people do lose weight, then in general the **before** > **after** and the D s are positive. How we define our differences determines that this example is a right-tailed test (carry the > sign down to the alternative hypothesis) and the correct hypotheses are:

$$\begin{aligned} H_0: \mu_D &= 0 \\ H_1: \mu_D &> 0 \end{aligned}$$

If we were to do the same problem but reverse the order and take D = after weight – before weight the correct alternative hypothesis is $H_1: \mu_D < 0$ since after weight < before weight. Just be consistent throughout your problem, and never switch the order of the groups in a problem.

Example 8-5: In an effort to increase production of an automobile part, the factory manager decides to play music in the manufacturing area. Eight workers are selected, and the number of items each produced for a specific day is recorded. After one week of music, the same workers are monitored again. The data are given in the table. At $\alpha = 0.05$, can the manager conclude that the music has increased production? Assume production is normally distributed. Use the p-value method.

Worker	1	2	3	4	5	6	7	8
Before	6	8	10	9	5	12	9	7
After	10	12	9	12	8	13	8	10

Solution:

Assumptions: We are comparing production rates before and after music is played in the manufacturing area. We are given that the production rates are normally distributed. Because these are consecutive times from the same population, they are dependent samples, so we must use the t-test for matched pairs. Let population 1 be the number of items before the music, and population 2 be after. The claim is that music increases production so before production < after production. Carry this same sign to the alternative hypothesis.

The correct hypotheses are: $H_0: \mu_D = 0$; $H_1: \mu_D < 0$, this is a left-tailed test.

In order to compute the t-test statistic, we must first compute the differences between each of the matched pairs.

Before (x_1)	6	8	10	9	5	12	9	7
After (x_2)	10	12	9	12	8	13	8	10
$D = x_1 - x_2$	-4	-4	1	-3	-3	-1	1	-3

Using the 1-var stats on the differences in your calculator, we compute $\bar{D} = \bar{x} = -2$, $s_D = s_x = 2.0702$, $n = 8$.

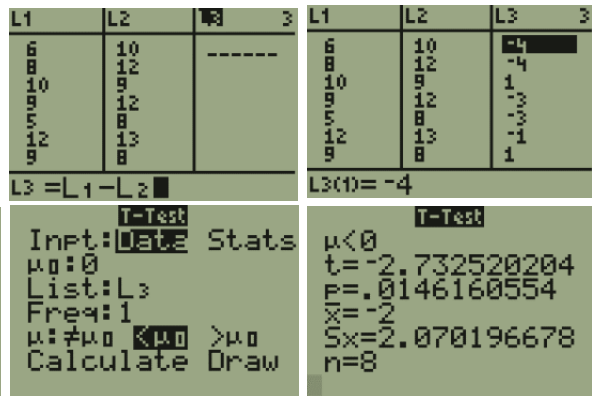
The test statistic is: $t = \frac{\bar{D} - \mu_D}{\left(\frac{s_D}{\sqrt{n}}\right)} = \frac{-2 - 0}{\left(\frac{2.0702}{\sqrt{8}}\right)} = -2.7325$.

The p-value for a two-tailed t-test with degrees of freedom $= n - 1 = 7$, is found by finding twice the area to the left of the test statistic -2.7325 using technology. See screenshots below for the calculator to find the p-value = 0.0146.

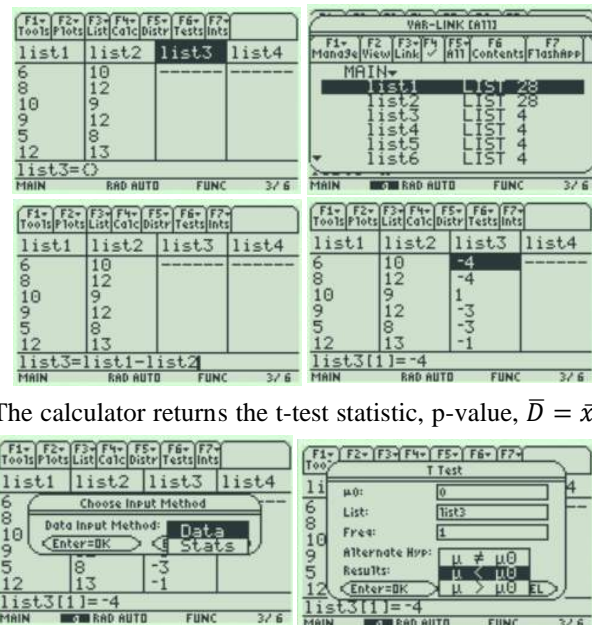
Decision: Since the p-value = 0.0146 is less than $\alpha = 0.05$, we reject H_0 .

Summary: At the 5% level of significance, there is enough evidence to support the claim that the mean production rate increases when music is played in the manufacturing area.

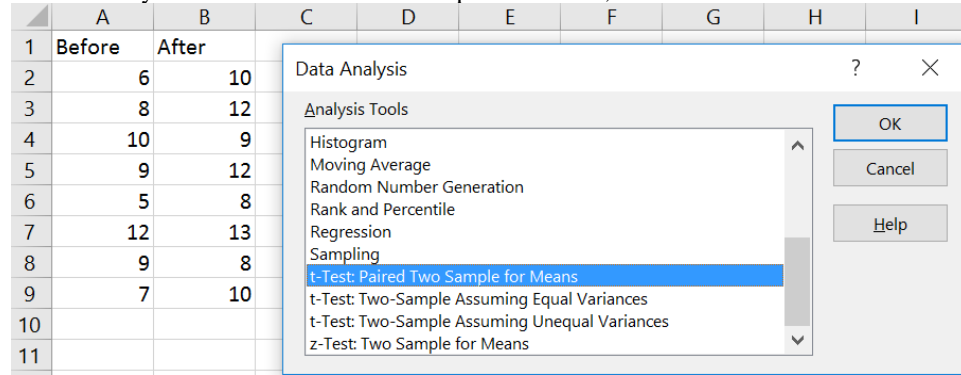
TI-84: Find the differences between the sample pairs (you can subtract two lists to do this). Press the [STAT] key and then the [EDIT] function, enter the difference column into list one. Press the [STAT] key, arrow over to the [TESTS] menu. Arrow down to the option [2:T -Test] and press the [ENTER]. Arrow over to the [Data] menu and press the [ENTER] key. Then type in the hypothesized mean as 0, List: L3, leave Freq:1 alone, arrow over to the $\neq, <, >$ sign that is the same in the problem's alternative hypothesis statement then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the t-test statistic, the p-value, mean of the differences $\bar{D} = \bar{x}$ and standard deviation of the differences $s_D = s_x$.



TI-89: Find the differences between the sample pairs (you can subtract two lists to do this). Go to the [Apps] Stat/List Editor, enter the two data sets in lists 1 and 2. Move the cursor so that it is highlighted on the header of list3. Press [2nd] Var-Link and move down to list1 and press [Enter]. This brings the name list1 back to the list3 at the bottom, select the minus [-] key, then select [2nd] Var-link and this time highlight list2 and press [Enter]. You should now see list1-list2 at the bottom of the window. Press [Enter] then the differences will be stored in list3. Press [2nd] then F6 [Tests], select 2: T-Test. Select the [Data] menu. Then type in the hypothesized mean as 0, List: list1, Freq:1, arrow over to the $\neq, <, >$ and select the sign that is the same in the problem's alternative hypothesis, press the [ENTER] key to calculate. The calculator returns the t-test statistic, p-value, $\bar{D} = \bar{x}$ and $s_D = s_x$.



Excel: Start by entering the data in two columns in the same order that they appear in the problem. Then select Data > Data Analysis > t-test: Paired Two Sample for Means, then select OK.



Select the Before data (including the label) into the Variable 1 Range, and the After data (including the label) in the Variable 2 Range. Type in zero for the Hypothesized Mean Difference box. Select the box for Labels (do not select

this if you do not have labels in the variable range selected). Change alpha to fit the problem. You can leave the default to open in a new worksheet or change output range to be one cell where you want the top left of the output table to start (make sure this cell does not overlap any existing data). Then select OK. See below for example.

t-Test: Paired Two Sample for Means

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

Labels

Alpha:

Output options

Output Range:

New Worksheet Ply:

New Workbook

OK Cancel Help

You get the following output:

t-Test: Paired Two Sample for Means			
	Before	After	
Mean	8.25	10.25	Sample means for each group.
Variance	5.0714	3.6429	Sample variance for each group.
Observations	8	8	Sample size for each group.
Pearson Correlation	0.5152		Ignore this for now.
Hypothesized Mean Difference	0		This is the zero we set for H_0 .
df	7		$df = \text{number of pairs} - 1$.
t Stat	-2.7325		Test Statistic t-score.
P(T<=t) one-tail	0.0146		p-value for a 1-tailed test
t Critical one-tail	1.8946		Absolute value of the critical value for a one-tailed test.
P(T<=t) two-tail	0.0292		p-value for a two-tailed test.
t Critical two-tail	2.3646		Positive critical value for a two-tailed test.

One nice feature in Excel is that you get the p-value and the critical value in the output. The critical value can be taken from the Excel output; however, Excel never gives negative critical values. Since we are doing a left-tailed test we will need to use the t-score = -1.8946.

If we were to draw and shade the critical region for the sampling distribution, it would look like Figure 8-4.

The decision is made by comparing the test statistic $t = -2.7325$ with the critical value $t_\alpha = -1.8946$.

Since the test statistic is in the shaded critical region, we would reject H_0 .

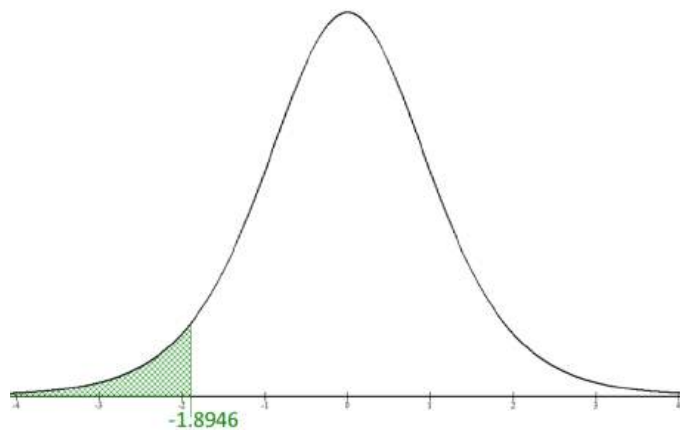


Figure 8-4

At the 5% level of significance, there is enough evidence to support the claim that the mean production rate increases when music is played in the manufacturing area.

The decision and summary should not change from using the p-value method.

Example 8-6: Hands-On Café records the number of online orders for eight randomly selected locations for two consecutive days. Assume the number of online orders is normally distributed. Test to see if there is a difference in mean number of online orders between Thursday and Friday using a 5% level of significance.

Location	1	2	3	4	5	6	7	8
Thursday	67	65	68	68	68	70	69	70
Friday	68	70	69	71	72	69	70	70

Solution: First set up the hypotheses. We are testing to see if Thursday \neq Friday orders. The hypotheses would be:

$$H_0: \mu_D = 0$$

$$H_1: \mu_D \neq 0$$

Next, to calculate the test statistic, compute the differences of Thursday – Friday for each pair.

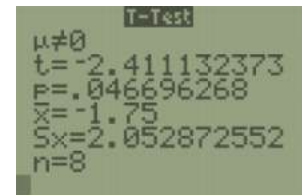
Thursday	67	65	68	68	68	70	69	70
Friday	68	70	69	71	72	69	70	70
D	-1	-5	-1	-3	-4	1	-1	0

Use technology to compute the mean, standard deviation and sample size. Note if you use a TI calculator then $\bar{D} = \bar{x} = -1.75$, $s_D = s_x = 2.05287$, $n = 8$.

Substitute these values into the test statistic equation $t = \frac{\bar{D} - \mu_D}{\left(\frac{s_D}{\sqrt{n}}\right)} = \frac{-1.75 - 0}{\left(\frac{2.05287}{\sqrt{8}}\right)} = -2.4111$.

The p-value = 0.0467.

Since the p-value $< \alpha$, reject H_0 . At the 5% level of significance, there is enough evidence to support the claim that there is a difference in the mean sales for online orders between Thursday and Friday.



Confidence Interval For Dependent Means

The t-interval for dependent samples is a confidence interval for comparing the means from two dependent populations (or the difference between the means from two populations). The t-interval is used when the differences are normally distributed. The two samples also must be dependent.

A $(1 - \alpha) * 100\%$ confidence interval for the difference between two population means with matched pairs: $\mu_D =$ mean of the differences.

$$\bar{D} - t_{\alpha/2} \left(\frac{s_D}{\sqrt{n}}\right) < \mu_D < \bar{D} + t_{\alpha/2} \left(\frac{s_D}{\sqrt{n}}\right)$$

$$\text{Or more compactly as } \bar{D} \pm t_{\alpha/2} \left(\frac{s_D}{\sqrt{n}}\right)$$

Where the t-distribution has degrees of freedom, $df = n - 1$, where n is the number of pairs.

Example 8-7: Hands-On Café records the number of online orders for eight randomly selected locations for two consecutive days. Assume the number of online orders is normally distributed. Compute the 95% confidence interval for the mean difference of online orders between Thursday and Friday.

Location	1	2	3	4	5	6	7	8
Thursday	67	65	68	68	68	70	69	70
Friday	68	70	69	71	72	69	70	70

Solution: Compute the $t_{\alpha/2}$ critical value for a 95% confidence interval and $df = 7$. Use the t-distribution with technology using confidence level 95%, lower tail area of $\alpha/2 = 0.025$ to get $t_{\alpha/2} = t_{0.025} = \pm 2.36462$. Compute the differences of Thursday – Friday for each pair.

Thursday	67	65	68	68	68	70	69	70
Friday	68	70	69	71	72	69	70	70
D	-1	-5	-1	-3	-4	1	-1	0

Use technology to compute the mean, standard deviation and sample size. Note if you use a TI calculator then $\bar{D} = \bar{x} = -1.75$ and $s_D = s_x = 2.05287$.

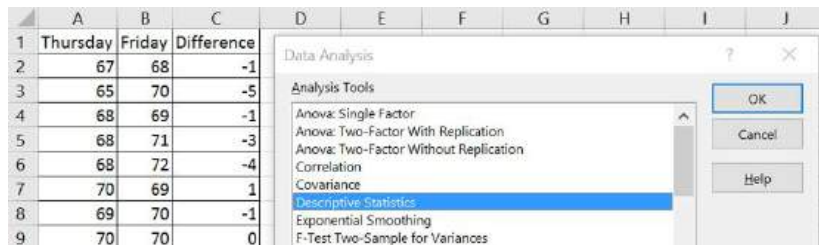
Calculate the interval estimate: $\bar{D} \pm t_{\alpha/2} \left(\frac{s_D}{\sqrt{n}} \right) \Rightarrow -1.75 \pm 2.36462 \left(\frac{2.05287}{\sqrt{8}} \right) \Rightarrow -1.75 \pm 1.7162$.

Write the answer using standard notation $-3.4662 < \mu_D < -0.0338$ or interval notation $(-3.4662, -0.0338)$.

For an interpretation of the interval, if we were to use the same sampling techniques, approximately 95 out of 100 times the confidence interval $(-3.4662, -0.0338)$ would contain the population mean difference in the number of orders between Thursday and Friday. Since both endpoints are negative, we can be 95% confident that the population mean number of orders for Thursday is between 3.4662 and 0.0338 orders lower than Friday.

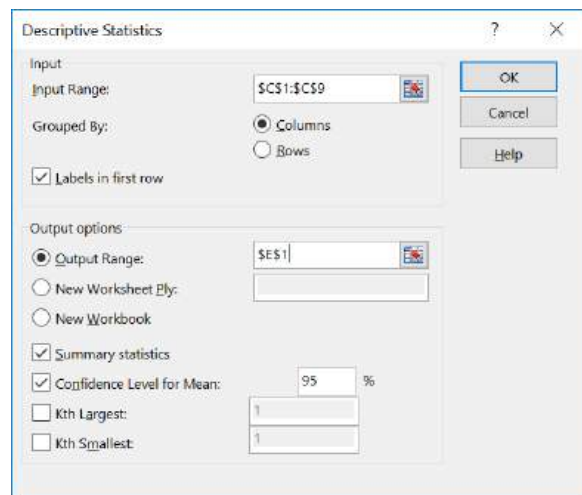
Excel: Type in both samples in two adjacent columns, and then subtract each pair in a third column and label the column Difference.

Thursday	Friday	Difference
67	68	=A2-B2
65	70	=A3-B3
68	69	=A4-B4
68	71	=A5-B5
68	72	=A6-B6
70	69	=A7-B7
69	70	=A8-B8
70	70	=A9-B9



Select Data > Data Analysis > Descriptive Statistics and click OK.

Select the Difference column for the input range including the label, then check the box next to Labels in first row (do not select this box if you did not highlight a label in the input range). Use the default new worksheet or select a single cell for the Output Range where you want your top left-hand corner of the table to start. Check the boxes Summary Statistics and Confidence Level for Mean. Change the confidence level to fit the question, and then select OK.



You get the following output:

<i>Difference</i>		
Mean	-1.75	\bar{D} = Mean of the differences
Standard Error	0.7258	
Median	-1	
Mode	-1	
Standard Deviation	2.0529	s_D = Standard Deviation of the differences
Sample Variance	4.2143	
Kurtosis	-0.9009	
Skewness	-0.4458	
Range	6	
Minimum	-5	
Maximum	1	
Sum	-14	
Count	8	Sample Size = n = Number of Pairs
Confidence Level(95.0%)	1.7162	Margin of Error

The confidence interval is the mean \pm margin of error. In two different cells subtract and then add the margin of error from the mean to get the confidence interval limits and then put your answer in interval notation ($-3.4662, -0.0338$).

TI-84: First, find the differences between the samples. Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the [8:TInterval] option and press the [ENTER] key. Arrow over to the [Data] menu and press the [ENTER] key. The defaults are List: L₁, Freq:1. If this is set with a different list, arrow down and use [2nd] [1] to get L₁. Then type in the confidence level. Arrow down to [Calculate] and press the [ENTER] key. The calculator returns the confidence interval, $\bar{D} = \bar{x}$ and $s_D = s_x$.

TI-89: First, find the differences between the samples. Go to the [Apps] **Stat/List Editor**, then enter the differences into list 1. Press [2nd] then F7 [Ints], then select **2: T-Interval**. Select the [Data] menu. Enter in List: list1, Freq:1. Then type in the confidence level. Press the [ENTER] key to calculate. The calculator returns the confidence interval, $\bar{D} = \bar{x}$ and $s_D = s_x$.

8.4 Two Independent Means Hypothesis Test & Confidence Interval

This section will look at how to analyze a difference in the mean for two independent samples. As with all other hypothesis tests and confidence intervals, the process is the same, though the formulas and assumptions are different.

When you are making a conjecture about a population mean, we have **two** different situations, depending on if we know that population standard deviation, or not, called the z -test and t -test, respectively. Use Figure 8-5 to help decide when to use the z -test and t -test. Note that you should never use the value of σ_x on your calculator since you would rarely ever have an entire population of raw data to input into a calculator. The problem may give you raw data, but σ or σ^2 would be stated in the problem and you should be using a z -test, otherwise use the t -test with the sample standard deviation s_x . Usually, σ is unknown, but could come from a previous year or similar study. We rarely have access to the population standard deviations and this text will use the σ unknown case.

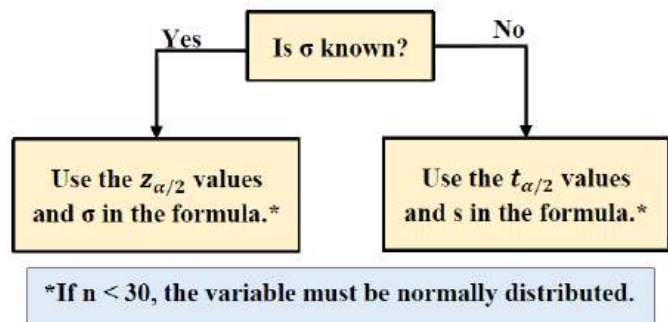


Figure 8-5

If the sample sizes are below 30, we need to check that the population is approximately normally distributed. We can do this with a normal probability plot. Most examples that we deal with assume the population is normally distributed, but in practice, you should always check this assumption. If the populations are not normally distributed, then nonparametric methods discussed in more advanced courses may be used.

When setting up the null hypothesis we are testing if the difference in the two means is equal to some known difference. $H_0: \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$. We will focus on the case where $(\mu_1 - \mu_2)_0 = 0$, which says that, tentatively, we assume that there is no difference in population means $H_0: \mu_1 - \mu_2 = 0$. If we were to subtract μ_2 from both sides of the equation $\mu_1 - \mu_2 = 0$ we would get $\mu_1 = \mu_2$. For instance, if the average age for group one was 25 and the average age for group two was also 25, then the difference between the two means would be $25 - 25 = 0$.

Figure 8-6 shows the three ways to set up hypotheses for comparing two **independent** population means μ_1 and μ_2 .

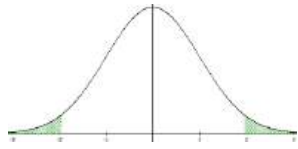
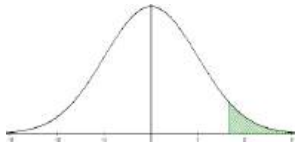
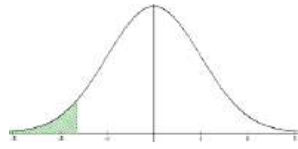
Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$
		
$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 > 0$	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 < 0$

Figure 8-6

This text mostly will use an equal (=) sign in the null hypothesis. For a one-tailed test, one could alternatively write the null hypotheses as:

<u>Right-tailed test</u>	<u>Left-tailed test</u>
$H_0: \mu_1 \leq \mu_2$	$H_0: \mu_1 \geq \mu_2$
$H_1: \mu_1 > \mu_2$	$H_1: \mu_1 < \mu_2$

The traditional method (or critical value method), the p-value method, and the confidence interval method are performed with steps that are similar to those when performing hypothesis tests for one population.

8.4.1 Unequal Variance Method

The t-test is a statistical test for comparing the means from two independent populations. The following t-test is used when σ_1 and/or σ_2 are both unknown and assumed unequal. The samples must be independent and if the sample sizes are less than 30 then the populations need to be normally distributed. The sample sizes both need to be 30 or more, or the populations need to be approximately normally distributed. The traditional method (or critical value method), the p-value method, and the confidence interval method are performed with steps similar to those when performing hypothesis tests for one population.

If we assume the variances are **unequal** ($\sigma_1^2 \neq \sigma_2^2$), the formula for the t test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

Use the t-distribution where the degrees of freedom are $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1-1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2-1}\right)}$.

Some older calculators only accept the df as an integer, in this case round the df **down** to the nearest integer if needed. For most technology, you would want to keep the decimal df . Some textbooks use an approximation for the df as the smaller of $n_1 - 1$ or $n_2 - 1$, so you may find a different answer using your calculator compared to examples found elsewhere.

Note that $\mu_1 - \mu_2$ is the hypothesized difference found in the null hypothesis and is usually zero.

Example 8-8: The general United States adult population volunteer an average of 4.2 hours per week. A random sample of 18 undergraduate college students and 20 graduate college students indicated the results below concerning the amount of time spent in volunteer service per week. At $\alpha = 0.01$ level of significance, is there sufficient evidence to conclude that a difference exists between the mean number of volunteer hours per week for undergraduate and graduate college students? Assume that number of volunteer hours per week is normally distributed.

	Undergraduate	Graduate
Sample Mean	2.5	3.8
Sample Variance	2.2	3.5
Sample Size	18	20

Solution: Assumptions: The two populations we are comparing are undergraduate and graduate college students. We are given that the number of volunteer hours per week is normally distributed. We are told that the samples were randomly selected and should therefore be independent. We do not know the two population standard deviations (we only have the sample standard deviations as the square root of the sample variances), so we must use the t -test. Using the critical value method steps, we get the following.

The question is asking if there is a difference between the mean number of volunteer hours per week for undergraduate and graduate level college students. We let population 1 be undergraduate students, and population 2 be graduate students.

The correct hypotheses for a two-tailed test are: $H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2$.

$$\text{The test statistic is } t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{(2.5 - 3.8) - 0}{\sqrt{\left(\frac{2.2}{18} + \frac{3.5}{20}\right)}} = -2.3845.$$

The critical value for a two-tailed t -test with degrees of freedom is found by using tail area $\alpha/2 = 0.005$ with

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)\right)} = \frac{\left(\frac{2.2}{18} + \frac{3.5}{20}\right)^2}{\left(\left(\frac{2.2}{18}\right)^2 \left(\frac{1}{17}\right) + \left(\frac{3.5}{20}\right)^2 \left(\frac{1}{19}\right)\right)} = 35.0753.$$

Draw the curve and label the critical values as shown in Figure 8-7.

Use the invT function on your calculator to compute the critical value $\text{invT}(0.005, 35.0753) = -2.724$ (older calculators may require you to use a whole number, round down to $df = 35$), or use Excel $=\text{T.INV}(0.005, 35.0753)$ to compute the critical value.

The test statistic of -2.3845 is between the critical values -2.724 and 2.724 , therefore do not reject H_0 .

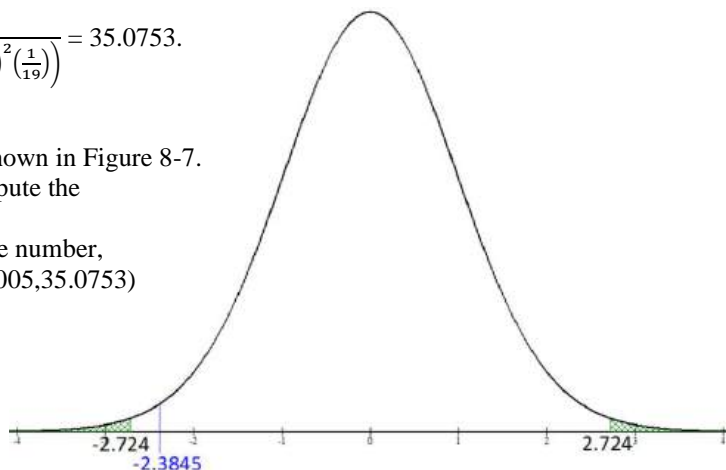
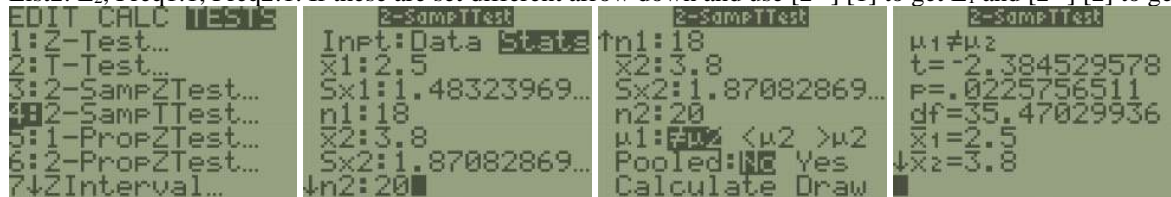


Figure 8-7

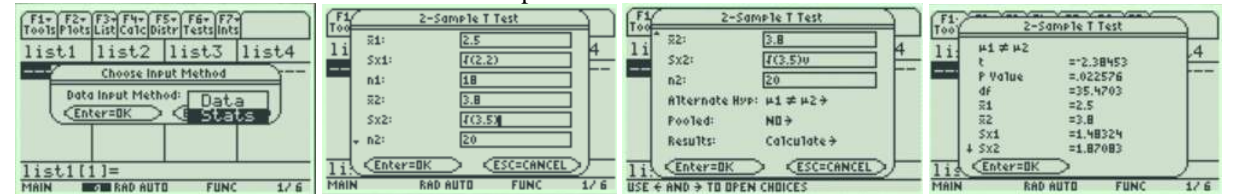
There is not enough evidence to suggest a difference between the population mean number of volunteer hours per week for undergraduate and graduate college students.

Note that if we had decided to have population 1 be graduate students, the test statistic would be positive 2.3845, but this would not change our decision for a two-tailed test. If you are doing a one-tailed test, then you need to be consistent on which sign your test statistic has. Most of the time for a left-tailed test both the critical value and the test statistic will be negative and for a right-tailed test both the critical value and test statistic will be positive.

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [4:2-SampTTest] and press the [ENTER] key. Arrow over to the [Stats] menu and press the [Enter] key. Enter the means, standard deviations, sample sizes, confidence level. Then arrow over to the \neq , $<$, $>$ sign that is the same in the problem's alternative hypothesis statement, then press the [ENTER] key. Highlight the No option under Pooled for unequal variances. Arrow down to [Calculate] and press the [ENTER] key. The calculator returns the test statistic and the p-value. If you have raw data, press the [STAT] key and then the [EDIT] function, then enter the data into list one and list two. Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [4:2-SampTTest] and press the [ENTER] key. Arrow over to the [Data] menu and press the [ENTER] key. The defaults are List1: L₁, List2: L₂, Freq1:1, Freq2:1. If these are set different arrow down and use [2nd] [1] to get L₁ and [2nd] [2] to get L₂.



TI-89: Go to the [Apps] Stat/List Editor, then press [2nd] then F6 [Tests], then select **4: 2-SampT-Test**. Enter the sample means, sample standard deviations, and sample sizes (or list names (list3 & list4), and Freq1:1 & Freq2:1). Then arrow over to the \neq , $<$, $>$ and select the sign that is the same in the problem's alternative hypothesis statement. Highlight the No option under Pooled. Press the [ENTER] key to calculate. The calculator returns the t-test statistic and the p-value.



Example 8-9: A researcher is studying how much electricity (in kilowatt hours) households from two different cities use in their homes. Random samples of 17 days in Sacramento and 16 days in Portland are given below. Test to see if there is a difference using all 3 methods (critical value, p-value and confidence interval). Assume that electricity use is normally distributed and the population variances are unequal. Use $\alpha = 0.10$.

Sacramento			
474	414	692	467
443	605	419	277
670	696	783	
577	813	694	
565	663	884	

Portland			
783	587	527	546
442	107	728	662
371	427	277	474
605	293	320	555

Solution: The populations are independent and normally distributed.

The hypotheses for all 3 methods are: $H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2$.

Use technology to find the sample means, standard deviations and sample sizes.

Enter the Sacramento data into list 1, then do 1-Var Stats L₁ and you should get $\bar{x}_1 = 596.2353$, $s_1 = 163.2362$, and $n_1 = 17$.

Enter the Portland data into list 2, then do 1-Var Stats L₂ and you should get $\bar{x}_2 = 481.5$, $s_2 = 179.3957$, and $n_2 = 16$.

$$\text{The test statistic is } t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{(596.2353 - 481.5) - 0}{\sqrt{\left(\frac{163.2362^2}{17} + \frac{179.3957^2}{16}\right)}} = 1.9179.$$

$$\text{The } df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)} = \frac{\left(\frac{163.2362^2}{17} + \frac{179.3957^2}{16}\right)^2}{\left(\frac{163.2362^2}{17}\right)^2 \left(\frac{1}{16}\right) + \left(\frac{179.3957^2}{16}\right)^2 \left(\frac{1}{15}\right)} = 30.2598.$$

The p-value would be double the area to the right of $t = 1.9179$. Using the TI calculator or Excel we get the p-value = 0.0646. Stop and see if you can find this p-value using the same process from previous sections.

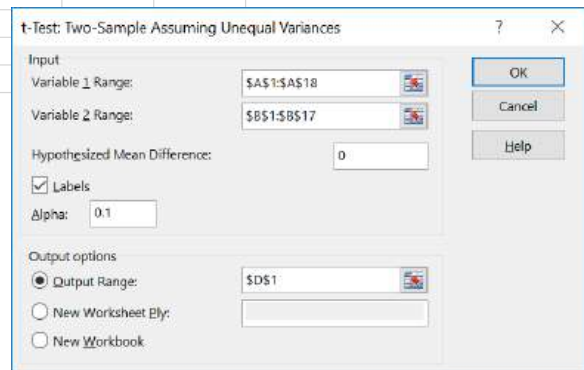
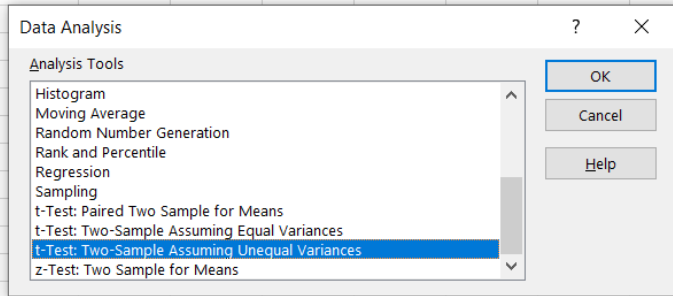
Since the p-value is less than alpha, we would reject H₀.

At the 10% level of significance, there is a statistically significant difference between the mean electricity use in Sacramento and Portland.

Excel

When you have raw data, you can use Excel to find all this information using the Data Analysis tool. Enter the data into Excel, then choose Data > Data Analysis > t-Test: Two Sample Assuming Unequal Variances.

	A	B	C	D	E	F	G	H	I	J
1	Sacramento	Portland								
2		474	783							
3		443	442							
4		670	371							
5		577	605							
6		565	587							
7		414	107							
8		605	427							
9		696	293							
10		813	527							
11		663	728							
12		692	277							
13		419	320							
14		783	546							
15		694	662							
16		884	474							
17		467	555							
18		277								
19										



Enter the necessary information as we did in previous sections (see output below) and select OK.

You can use this Excel shortcut only if you have raw data given in the question.

We get the following output, which has both p-values and critical values.

t-Test: Two-Sample Assuming Unequal Variances			
	<i>Sacramento</i>	<i>Portland</i>	
Mean	596.2352941	481.5	Sample means for each group.
Variance	26646.06618	32182.8	Sample variance for each group.
Observations	17	16	Sample size for each group.
Hypothesized Mean Difference	0		Zero that we used from the null hypothesis.
df	30		df rounded down.
t Stat	1.917899527		Test statistic.
P(T<=t) one-tail	0.032343507		P-value for either a left- or right-tailed test.
t Critical one-tail	1.310415025		Absolute value of the critical value for a one-tailed test.
P(T<=t) two-tail	0.064687015		P-value for a two-tailed test.
t Critical two-tail	1.697260887		Positive critical value for a two-tailed test.

Example 8-9 Using the Traditional Method

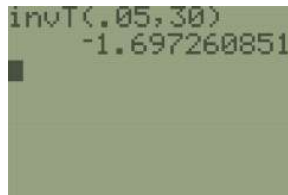
The hypotheses and test statistic steps do not change compared to the p-value method.

Hypotheses: $H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2.$

$$\text{Test Statistic: } t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{(596.2353 - 481.5) - 0}{\sqrt{\left(\frac{163.2362^2}{17} + \frac{179.3957^2}{16}\right)}} = 1.9179.$$

Compute the t critical values. First, calculate the degrees of freedom: $df = \frac{\left(\frac{163.2362^2}{17} + \frac{179.3957^2}{16}\right)^2}{\left(\frac{163.2362^2}{17}\right)^2 \left(\frac{1}{16}\right) + \left(\frac{179.3957^2}{16}\right)^2 \left(\frac{1}{15}\right)} = 30.2598.$

We can use the t Critical two-tail value given in the Excel output or use the TI-calculator $\text{invT}(0.05, 30.2598) = -1.697$. Some older calculators do not let you use a decimal for df so round down and use $\text{invT}(0.05, 30)$.



Draw and label the t-distribution with the critical values. See Figure 8-8.

Since the test statistic is in the critical region, we would reject H_0 . This agrees with the same decision that we had using the p-value method.

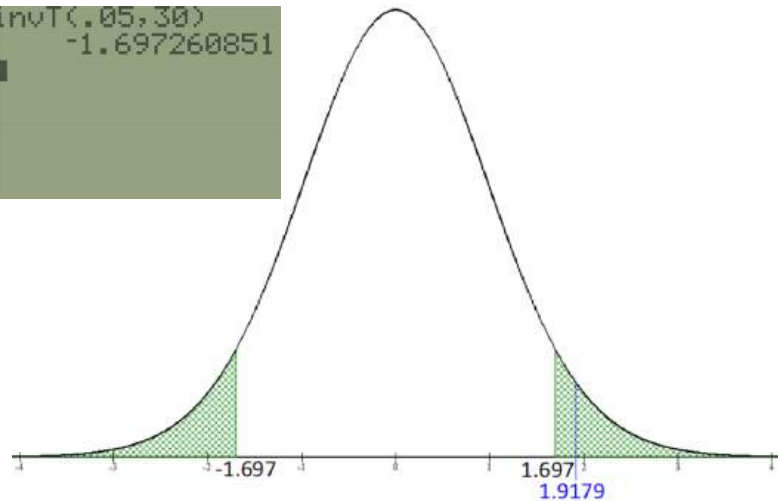


Figure 8-8

Summary: At the 10% level of significance, there is statistically significant difference between the mean electricity use between Sacramento and Portland.

Two-Sample T-Interval

For independent samples, we take the mean of each sample, then take the difference in the means. If the means are equal, then the difference of the two means would be equal to zero. We can then compare the null hypothesis, that there is no difference in the means $\mu_1 - \mu_2 = 0$, with the confidence interval limits to decide whether to reject the null hypothesis. If zero is contained within the confidence interval, then we fail to reject H_0 . If zero is not contained within the confidence interval, then we reject H_0 .

A $(1 - \alpha) * 100\%$ confidence interval for the difference between two population means $\mu_1 - \mu_2$ for independent samples with unequal variances: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$.

The requirements and degrees of freedom are identical to the above hypothesis test.

Example 8-9 Using the Confidence Interval Method

The hypotheses are the same. The main difference is that we would find a confidence interval and compare $H_0: \mu_1 - \mu_2 = 0$ with the endpoints to make the decision.

Hypotheses: $H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2$.

Calculate the confidence interval. First, compute the $t_{\alpha/2}$ critical value for a 90% confidence interval since $\alpha = 0.10$.

$$\text{Use } df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1-1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2-1}\right)} = \frac{\left(\frac{163.2362^2}{17} + \frac{179.3957^2}{16}\right)^2}{\left(\frac{163.2362^2}{17}\right)^2 \left(\frac{1}{16}\right) + \left(\frac{179.3957^2}{16}\right)^2 \left(\frac{1}{15}\right)} = 30.2598.$$



The critical value is $t_{\alpha/2} = \text{invT}(0.05, 30.2598) = -1.697$.

The older TI-83 invT program only accepts integer df , use $df = 30$. Alternatively, use the output from the Excel output under the t Critical two-tail row.

Next, substitute the values into the interval estimate formula $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$

$$\Rightarrow (596.2353 - 481.5) \pm 1.697 \sqrt{\left(\frac{163.2362^2}{17} + \frac{179.3957^2}{16}\right)}$$

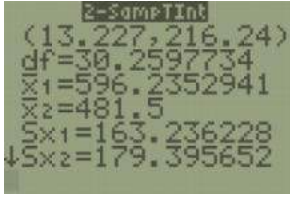
$$\Rightarrow 114.7353 \pm 101.5203.$$

Use interval notation (13.215, 216.2556) or standard notation $13.215 < \mu_1 - \mu_2 < 216.2556$.

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option **2-SampTInt** and press the [ENTER] key. Arrow over to the [Stats] menu and press the [Enter] key. Enter the means, standard deviations, sample sizes, confidence level. Highlight the **No** option under Pooled for unequal variances. Arrow down to [Calculate] and press the [ENTER] key. The calculator returns the confidence interval.

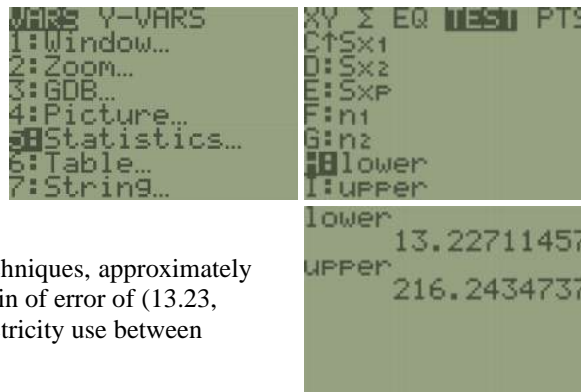


Or (if you have raw data in list one and list two) press the [STAT] key and then the [EDIT] function, type the data into list one for sample one and list two for sample two. Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [0:2-SampTInt] and press the [ENTER] key. Arrow over to the [Data] menu and press the [ENTER] key. The defaults are List1: L1, List2: L2, Freq1:1, Freq2:1. If these are set different, arrow down and use [2nd] [1] to get L1 and [2nd] [2] to get L2. Then type in the confidence level. Highlight the No option under Pooled for unequal variances. Arrow down to [Calculate] and press the [ENTER] key. The calculator returns the confidence interval.



TI-89: Go to the [Apps] **Stat/List Editor**, then press [2nd] then F5 [Ints], then select **4: 2-SampTInt**. Enter the sample means, sample standard deviations, sample sizes (or list names (list3 & list4), and Freq1:1 & Freq2:1), confidence level. Highlight the No option under Pooled. Press the [ENTER] key to calculate. The calculator returns the confidence interval. If you have the raw data, select Data and enter the list names.

Note the calculator does not round between steps and gives a more accurate answer of (13.23, 216.24). To see more decimal places on the TI-84, select the VARS key, select the Statistics option, then arrow over to TESTS and select lower then enter. Do the same for the upper. This will give you the lower and upper values of the confidence interval on your home screen.



For an interpretation, if we were to use the same sampling techniques, approximately 90 out of 100 times a confidence interval with the same margin of error of (13.23, 216.24) would contain the population mean difference in electricity use between Sacramento and Portland.

We are 90% confident that the population mean household electricity use for Sacramento is between 13.23 and 216.24 kilowatt hours more than Portland households.

Since both endpoints are positive, zero would not be captured in the confidence interval so we would reject H_0 .

Summary: At the 10% level of significance, there is statistically significant difference between the mean electricity use between Sacramento and Portland.

All 3 methods should yield the same result. This text is only using the two-sided confidence interval.

8.4.2 Equal Variance Method

The t-test for two independent samples has two different versions depending on if a particular assumption that the unknown population variances are unequal or equal. Since we do not know the true value of the population variances, we usually will use the first version and assume that the population variances are not equal $\sigma_1^2 \neq \sigma_2^2$. Both versions are presented, so make sure to check with your instructor if you are using both versions.

The equal variance method assumes that we know the population's standard deviations have approximately the same spread. Be careful with this since both populations could be normally distributed and independent, but one population may be way more spread out (larger variance) than the other so you would want to use the unequal variance version. For this text, we will state in the problem whether or not the population's variances (or standard deviations) are equal. Also, be careful when distinguishing between when to use the z-test versus t-test, just because we assume the population variances or standard deviations are equal does not mean we know their numeric values. We also need to assume the populations are normally distributed if either sample size is below 30.

If we assume the population variances are **equal** ($\sigma_1^2 = \sigma_2^2$), the formula for the t test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Use the t-distribution with pooled degrees of freedom $df = n_1 + n_2 - 2$.

The value $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$ under the square root is called the **pooled variance** and is a weighted mean of the two sample variances, weighted on the corresponding sample sizes.

In some textbooks, they may find the pooled variance first, then place into the formula as $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2 + s^2}{n_1 + n_2}}}$.

Note: The df formula matches what your calculator gives you when you select **Yes** under the Pooled option.

The traditional method (or critical value method), the p-value method, and the confidence interval method are performed with steps that are identical to those when performing hypothesis tests for one population.

Two-Sample t-Interval Assuming Equal Variances

For independent samples, we take the mean of each sample, then take the difference in the means. If the means are equal, then the difference of the two means would be equal to zero. We can then compare the null hypothesis, that there is no difference in the means $\mu_1 - \mu_2 = 0$, with the confidence interval limits to decide whether to reject the null hypothesis. If zero is contained within the confidence interval, then we fail to reject H_0 . If zero is not contained within the confidence interval, then we reject H_0 .

A $(1 - \alpha) * 100\%$ confidence interval for the difference between two population means $\mu_1 - \mu_2$ for independent samples assuming equal population variances ($\sigma_1^2 = \sigma_2^2$):

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

The requirements and degrees of freedom are identical to the above hypothesis test.

Example 8-10: A manager believes that the average sales in coffee at their Portland store is more than the average sales at their Cannon Beach store. They take a random sample of weekly sales from the two stores over the last year. Assume that the sales are normally distributed with equal variances. Test the manager's claim using $\alpha = 0.05$.

Portland	
1510	1257
4125	4677
1510	3055
5244	1764
4125	6128
6128	3319
3319	6433
3319	5244
3055	

Cannon Beach	
3585	1510
4399	5244
1764	1510
3853	4399
5244	1510
1510	5244
2533	4125
3585	2275
2533	2275
4399	3585
4125	5244

Solution:

Assumptions: The sample sizes are both less than 30, but the problem states that the populations are normally distributed. We are testing two means. We do not have population standard deviations or variances given in the problem so this will be a t-test not a z-test. The sales at each store are independent and the problem states that we are assuming, $\sigma_1^2 = \sigma_2^2$.

Set up the hypotheses, where group 1 is Portland, and group 2 is Cannon Beach.

We want to test if the Portland mean $>$ Cannon Beach mean, so carry this sign down to the alternative hypothesis to get a right-tailed test:

$$\begin{aligned} H_0: \mu_1 &= \mu_2 \\ H_1: \mu_1 &> \mu_2. \end{aligned}$$

Use technology to compute the sample means, standard deviations and sample sizes to get the following test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(3777.176471 - 3384.136364) - 0}{\sqrt{\left(\frac{(16 \cdot 1692.22728^2 + 21 \cdot 1361.86966^2)}{(17+22-2)}\right)\left(\frac{1}{17} + \frac{1}{22}\right)}} = 0.804135.$$

Use technology to find the critical value.

The $df = n_1 + n_2 - 2 = 17 + 22 - 2 = 37$.

On the TI Calculator use the invT function to compute the critical value
 $\text{invT}(0.95, 37) = 1.6871$.

In Excel use $=\text{T.INV}(0.95, 37) = 1.6871$ to compute the critical value.

Draw the curve and label the critical values as shown in Figure 8-9.

The test statistic 0.804135 is not in the critical region, therefore we do not reject H_0 .

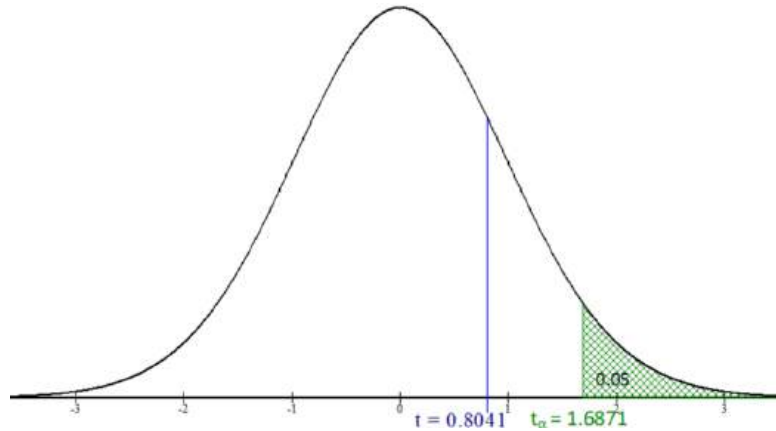


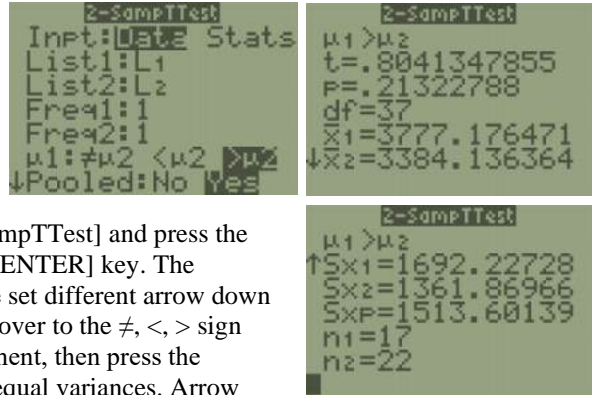
Figure 8-9

There is not enough evidence to conclude that there is a significant difference in the average sales for Portland and Cannon Beach.

For the p-value method on the TI calculator DIST menu with $\text{tcdf}(0.804135, 1E99, 37) = 0.2132$, or use shortcut shown below. In Excel use $=1 - \text{T.DIST}(0.804135, 37, \text{TRUE}) = 0.2132$ or if you have raw data see shortcut below.

The p-value = 0.2132 is larger than $\alpha = 0.05$, therefore we do not reject H_0 .

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [4:2-SampTTest] and press the [ENTER] key. Arrow over to the [Stats] menu and press the [Enter] key. Enter the means, standard deviations, sample sizes, confidence level. If you have raw data, press the [STAT] key and then the [EDIT] function, enter the data into list one and list two. Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [4:2-SampTTest] and press the [ENTER] key. Arrow over to the [Data] menu and press the [ENTER] key. The defaults are List1: L1, List2: L2, Freq1:1, Freq2:1. If these are set different arrow down and use [2nd] [1] to get L1 and [2nd] [2] to get L2. Then arrow over to the $\neq, <, >$ sign that is the same in the problem's alternative hypothesis statement, then press the [ENTER] key. Highlight the Yes option under Pooled for unequal variances. Arrow down to [Calculate] and press the [ENTER] key. The calculator returns the test statistic and the p-value.



TI-89: Go to the [Apps] Stat/List Editor, then press [2nd] then F6 [Tests], then select **4: 2-SampT-Test**. Enter the sample means, sample standard deviations, and sample sizes (or list names (list3 & list4), and Freq1:1 & Freq2:1). Then arrow over to the $\neq, <, >$ and select the sign that is the same in the problem's alternative hypothesis statement. Highlight the Yes option under Pooled. Press the [ENTER] key to calculate. The calculator returns the t-test statistic and the p-value.

Excel: Follow the same steps with the Data Analysis tool, except choose the t-Test: Two-Sample Assuming Equal Variances.

You can only use this Excel shortcut if you have raw data given in the question.

The screenshot shows an Excel spreadsheet with two columns of data: 'Portland' and 'Cannon Beach'. The 'Data Analysis' dialog box is open, and 't-Test: Two-Sample Assuming Equal Variances' is selected. The dialog box shows the following settings:

- Variable 1 Range: \$A\$1:\$A\$18
- Variable 2 Range: \$B\$1:\$B\$23
- Hypothesized Mean Difference: 0
- Labels
- Alpha: 0.05
- Output options:
 - Output Range: \$D\$1
 - New Worksheet Ply:
 - New Workbook

Enter the necessary information as we did in previous sections (see screenshot below) and select OK.

You get the following output:

t-Test: Two-Sample Assuming Equal Variances			
	<i>Portland</i>	<i>Cannon Beach</i>	
Mean	3777.176471	3384.136364	Sample means for each group.
Variance	2863633.154	1854688.981	Sample variance for each group.
Observations	17	22	Sample size for each group.
Pooled Variance	2290989.164		Pooled Variance.
Hypothesized Mean Difference	0		Zero that we used from the null hypothesis.
df	37		<i>df</i>
t Stat	0.804135		Test Statistic.
P(T<=t) one-tail	0.213228		P-value for either a left- or right-tailed test.
t Critical one-tail	1.687094		Absolute value of the critical value for a one-tailed test.
P(T<=t) two-tail	0.426456		P-value for a two-tailed test.
t Critical two-tail	2.026192		Positive critical value for a two-tailed test.

When reading the Excel output for a t-test, be careful with your signs.

- For a left-tailed t-test the critical value will be negative.
- For a right-tailed t-test the critical value will be positive.
- For a two-tailed t-test then your critical values would be \pm critical value.

Note you can only use the Excel shortcut if you have the raw data. If you have summarized data then you would need to do everything by hand or use the calculator shortcut.

Example 8-11: A Portland State University advisor wants to see whether there is a significant difference in ages of full-time students and part-time students. The advisor randomly selects a sample of 50 full-time students and finds a mean age of 22.12 years with a standard deviation of 3.68. The advisor randomly selects a sample of 50 part-time students and finds a mean age of 22.76 years with a standard deviation of 4.7. At $\alpha = 0.05$, decide if there is enough evidence to support the claim that there is a difference in the ages of the two groups. Assume the population variances are equal.

Solution:

Assumptions: The two populations we are sampling from are not necessarily normal, but the sample sizes are greater than 30. The two groups are randomly selected from a large population so we can assume the samples are independent; therefore, we use the t-test for comparing two population means μ_1 and μ_2 .

The claim is that there is a difference in the ages of the two student groups. Let full-time students be population 1 and part-time students be population 2 (always go in the same order as the data are presented in the problem unless otherwise stated). Then μ_1 would be the average age for full-time students and μ_2 would be the average age for part-time students. The key phrase is difference: $\mu_1 \neq \mu_2$.

The correct hypotheses are $H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2$.

This is a two-tailed test and the claim is in the alternative hypothesis.

Note that if we had decided to have population 1 be part-time students, the test statistic would be negated from that given below, but the p-value and result would be identical. In general, you should take population 1 as whatever group comes first in the problem.

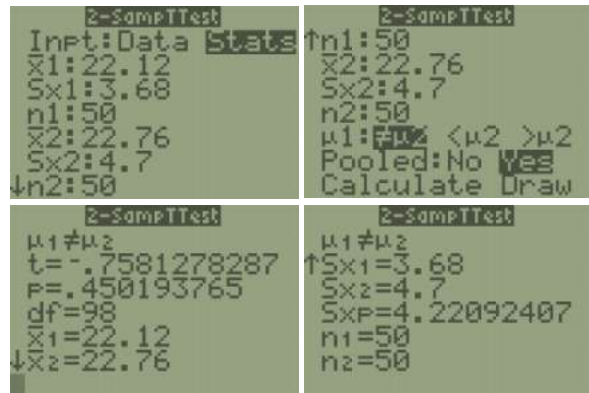
From the question we are given $n_1 = 50$, $\bar{x}_1 = 22.12$, $s_1 = 3.68$, $n_2 = 50$, $\bar{x}_2 = 22.76$, and $s_2 = 4.7$.

Since $\mu_1 = \mu_2$ then we know that $\mu_1 - \mu_2 = 0$. Substitute these values into the test statistic formula.

$$\text{The test statistic is: } t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(22.12 - 22.76) - 0}{\sqrt{\left(\frac{(49 \cdot 3.68^2 + 49 \cdot 4.7^2)}{(50 + 50 - 2)}\right)\left(\frac{1}{50} + \frac{1}{50}\right)}} = -0.7581.$$

The p-value for a two-tailed t-test is found by finding the area to the left (since t is negative) of the test statistic using a t-distribution and multiplying the area by two. We first need to find the degrees of freedom $df = n_1 + n_2 - 2 = 50 + 50 - 2 = 98$. Using the tcdf($-\infty, -0.7581278, 98$) we get an area of 0.22511. Since this is a two-tailed test we need to double the area, which gives a p-value = 0.4502. Note that if the t-score was positive, find the area to the right of t, then double.

Or use the calculator shortcut to get the p-value.



Decision: Because the p-value = 0.4502 is larger than $\alpha = 0.05$, we do not reject H_0 .

Summary: At the 5% level of significance, there is not enough evidence to support the claim that there is a difference in the ages of full-time students and part-time students.

Example 8-12: Calculate the 95% confidence interval for the previous example data.

Solution: From the question we are given $n_1 = 50$, $\bar{x}_1 = 22.12$, $s_1 = 3.68$, $n_2 = 50$, $\bar{x}_2 = 22.76$, and $s_2 = 4.7$. Calculate the critical value for $df = n_1 + n_2 - 2 = 50 + 50 - 2 = 98$. Using a TI Calculator use the $\text{invT}(0.025,98)$ or in Excel use $\text{T.INV}(0.025,98) = -1.984467$.

Substitute the numbers into the confidence interval formula: $(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$$\Rightarrow (22.12 - 22.76) \pm 1.984467 \sqrt{\left(\frac{(49 \cdot 3.68^2 + 49 \cdot 4.7^2)}{(50+50-2)}\right) \left(\frac{1}{50} + \frac{1}{50}\right)}$$

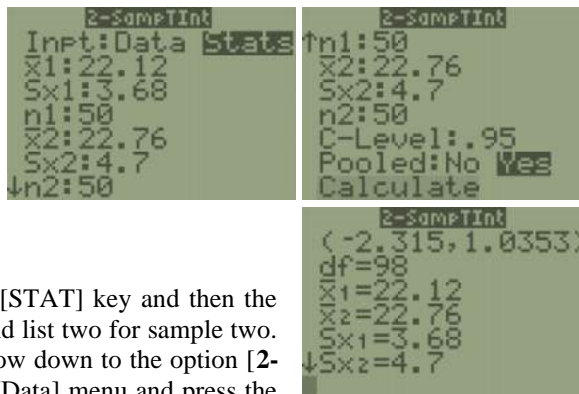
$$\Rightarrow -0.64 \pm 1.675257.$$

Use interval notation $(-2.3153, 1.0353)$ or standard notation $-2.3153 < \mu_1 - \mu_2 < 1.0353$.

For an interpretation, if we were to use the same sampling techniques, approximately 95 out of 100 times the confidence interval $(-2.3153, 1.0353)$ would contain the population mean difference in the ages of full-time students and part-time students. Since one endpoint is negative, and the other endpoint is positive, we fail to reject H_0 . We can be 95% confident that there is no difference in the mean ages of full-time students and part-time students.

There is no shortcut option for a two-sample t confidence interval in Excel.

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [2-SampTInt] and press the [ENTER] key. Arrow over to the [Stats] menu and press the [Enter] key. Enter the means, sample standard deviations, sample sizes, confidence level. Highlight the Yes option under Pooled for unequal variances. Arrow down to [Calculate] and press the [ENTER] key. The calculator returns the confidence interval.



Or (if you have raw data in list one and list two) press the [STAT] key and then the [EDIT] function, type the data into list one for sample one and list two for sample two. Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [2-SampTInt] and press the [ENTER] key. Arrow over to the [Data] menu and press the [ENTER] key. The defaults are List1: L_1 , List2: L_2 , Freq1:1, Freq2:1. If these are set different, arrow down and use [2nd] [1] to get L_1 and [2nd] [2] to get L_2 . Then type in the confidence level. Highlight the **Yes** option under Pooled for unequal variances. Arrow down to [Calculate] and press the [ENTER] key. The calculator returns the confidence interval.

TI-89: Go to the [Apps] Stat/List Editor, then press [2nd] then F5 [Ints], then select **4: 2-SampTInt**. Enter the sample means, sample standard deviations, sample sizes (or list names (list3 & list4) and Freq1:1 & Freq2:1), confidence level. Highlight the **Yes** option under Pooled. Press the [ENTER] key to calculate. The calculator returns the confidence interval. If you have the raw data, select Data and enter the list names.

Example 8-13: Non-rechargeable alkaline batteries and nickel metal hydride (NiMH) batteries are tested, and their voltage is compared. The data follow. Test to see if there is a difference in the means using a 95% confidence interval. Assume that both variables are normally distributed and the population variances are equal.

Alkaline	NiMH
$\bar{x}_1 = 9.2$ Volts	$\bar{x}_2 = 8.8$ Volts
$s_1 = 0.3$ Volts	$s_2 = 0.1$ Volts
$n_1 = 27$	$n_2 = 30$

Solution: First, set up the hypotheses $H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2$.

Next, find the $t_{\alpha/2}$ critical value for a 95% confidence interval.

From the question we are given $n_1 = 27$, $\bar{x}_1 = 9.2$, $s_1 = 0.3$, $n_2 = 30$, $\bar{x}_2 = 8.8$, and $s_2 = 0.1$.

Use technology to calculate the critical value for $df = n_1 + n_2 - 2 = 27 + 30 - 2 = 55$ and left tail area of the t-distribution $\alpha/2 = (1 - 0.95)/2 = 0.025$.

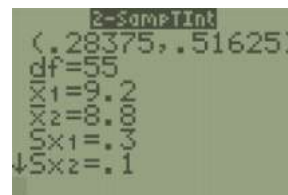
For a TI Calculator use the $\text{invT}(0.025, 55) = -2.004045$. For Excel use $\text{T.INV}(0.025, 55) = -2.004045$. The critical values are $t_{\alpha/2} = \pm 2.004045$.

Calculate the interval estimate (confidence interval):

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\Rightarrow (9.2 - 8.8) \pm 2.004045 \sqrt{\left(\frac{(26 \cdot 0.3^2 + 29 \cdot 0.1^2)}{(27+30-2)} \right) \left(\frac{1}{27} + \frac{1}{30} \right)}$$

$$\Rightarrow 0.4 \pm 0.11625.$$



Use interval notation $(0.2838, 0.5163)$ or standard notation $0.2838 < \mu_1 - \mu_2 < 0.5163$.

For an interpretation, if we were to use the same sampling techniques, approximately 95 out of 100 times the confidence interval $(0.2838, 0.5163)$ would contain the population mean difference in voltage between alkaline and NiMH batteries. Since both endpoints are positive, we can reject H_0 . We can be 95% confident that the population mean voltage for alkaline batteries is between 0.28 and 0.52 volts higher than nickel metal hydride batteries.

Trillian punched up the figures. They showed two-to-the power-of-Infinity-minus-one (an irrational number that only has a conventional meaning in Improbability physics).
 "... it's pretty low," continued Zaphod with a slight whistle.
 "Yes," agreed Trillian, and looked at him quizzically.
 "That's one big whack of Improbability to be accounted for.
 Something pretty improbable has got to show up on the balance sheet if it's all going to add up into a pretty sum."
 Zaphod scribbled a few sums, crossed them out and threw the pencil away.
 "Bat's dots, I can't work it out."
 "Well?"
 Zaphod knocked his two heads together in irritation and gritted his teeth.
 "OK," he said. "Computer!"
 (Adams, 2002)



Chapter 8 Formulas

<p>Hypothesis Test for 2 Proportions $H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$ $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p} \cdot \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ $\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)} = \frac{(\hat{p}_1 n_1 + \hat{p}_2 n_2)}{(n_1 + n_2)}$ $\hat{q} = 1 - \hat{p} \quad \hat{p}_1 = \frac{x_1}{n_1} \quad \hat{p}_2 = \frac{x_2}{n_2}$ TI-Calculator: 2-PropZTest</p>	<p>Confidence Interval for 2 Proportions $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\left(\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2} \right)}$ $\hat{p}_1 = \frac{x_1}{n_1} \quad \hat{p}_2 = \frac{x_2}{n_2}$ $\hat{q}_1 = 1 - \hat{p}_1 \quad \hat{q}_2 = 1 - \hat{p}_2$ TI-Calculator: 2-PropZInt</p>
<p>Hypothesis Test for 2 Dependent Means $H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$ $t = \frac{\bar{D} - \mu_D}{\left(\frac{s_D}{\sqrt{n}} \right)}$ TI-Calculator: T-Test</p>	<p>Confidence Interval for 2 Dependent Means $\bar{D} \pm t_{\alpha/2} \left(\frac{s_D}{\sqrt{n}} \right)$ TI-Calculator: TInterval</p>
<p>Hypothesis Test for 2 Independent Means $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ T-Test: Assume variances are unequal $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}}$ TI-Calculator: 2-SampTTest $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\left(\frac{s_1^2}{n_1} \right)^2 \left(\frac{1}{n_1 - 1} \right) + \left(\frac{s_2^2}{n_2} \right)^2 \left(\frac{1}{n_2 - 1} \right) \right)}$</p>	<p>Confidence Interval for 2 Independent Means T-Interval: Assume variances are unequal $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}$ TI-Calculator: 2-SampTInt $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\left(\frac{s_1^2}{n_1} \right)^2 \left(\frac{1}{n_1 - 1} \right) + \left(\frac{s_2^2}{n_2} \right)^2 \left(\frac{1}{n_2 - 1} \right) \right)}$</p>
<p>Hypothesis Test for 2 Independent Means $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ T-Test: Assume variances are equal $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ $df = n_1 + n_2 - 2$</p>	<p>T-Interval: Assume variances are equal $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$ $df = n_1 + n_2 - 2$</p>

The following flow chart in Figure 8-10 can help you decide which formula to use. Start on the left, ask yourself is the question about proportions (%), means (averages), standard deviations or variances? Are there 1 or 2 samples? Was the population standard deviation given? Are the samples dependent or independent? Are you asked to test a claim? If yes then use the test statistic (TS) formula. Are you asked to find a confidence interval? If yes then use the confidence interval (CI) formula. In each box is the null hypothesis and the corresponding TI calculator shortcut key.

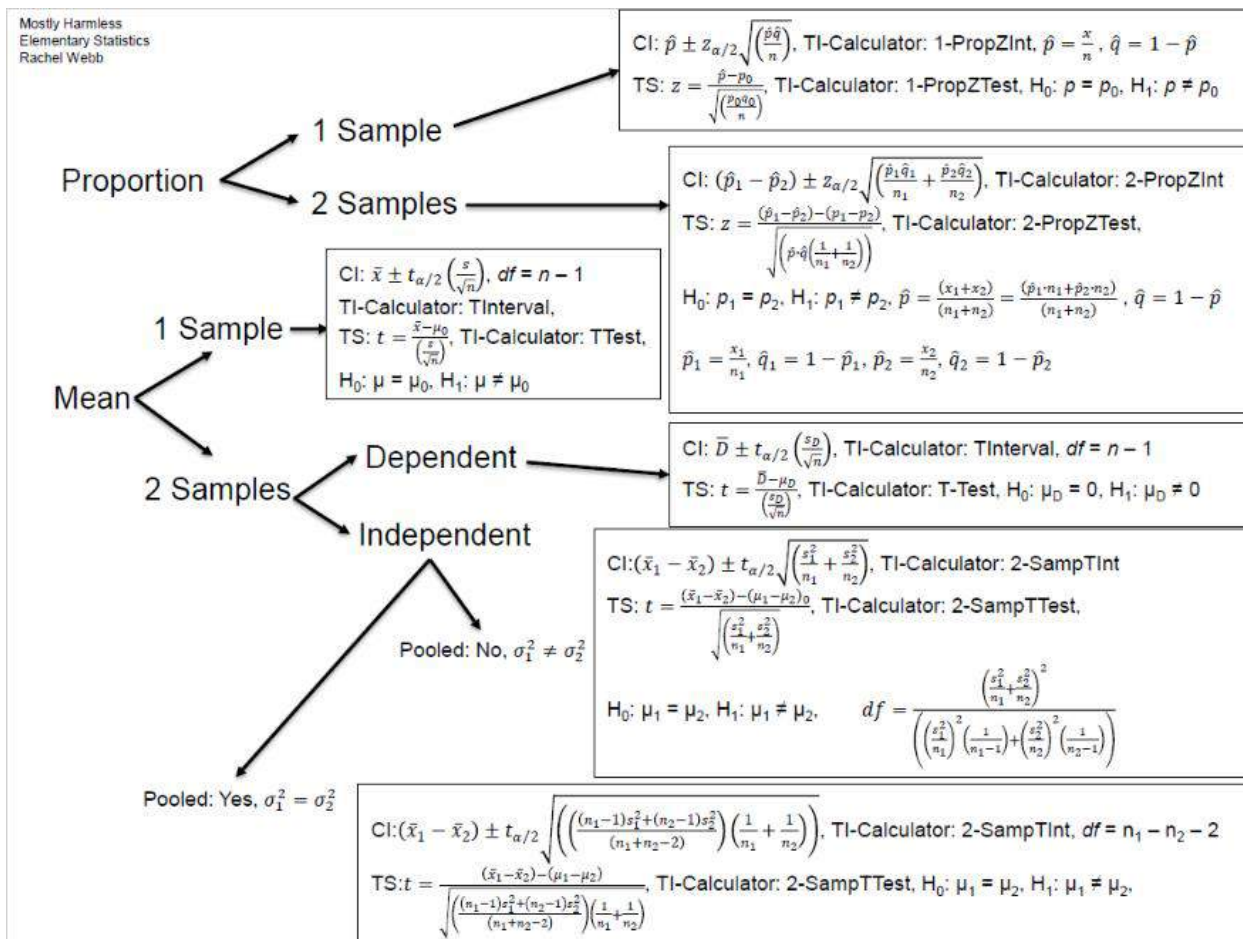


Figure 8-10

Download a .pdf version of the flowchart at: <http://MostlyHarmlessStatistics.com>.

The same steps are used in hypothesis testing for a one sample test. Use technology to find the p-value or critical value. A clue with many of these questions of whether the samples are dependent is the term “paired” is used, or the same person was being measured before and after some applied experiment or treatment. The p-value will always be a positive number between 0 and 1.

The same three methods to hypothesis testing, critical value method, p-value method and the confidence interval method are also used in this section. The p-value method is used more often than the other methods. The rejection rules for the three methods are:

- P-value method: reject H_0 when the p-value $\leq \alpha$.
- Critical value method: reject H_0 when the test statistic is in the critical tail(s).
- Confidence Interval method, reject H_0 when the hypothesized value (0) found in H_0 is outside the bounds of the confidence interval.

The most important step in any method you use is setting up your null and alternative hypotheses.

Chapter 8 Exercises

For exercises 1-5, show all 5 steps for hypothesis testing:

- State the hypotheses.
 - Compute the test statistic.
 - Compute the critical value or p-value.
 - State the decision.
 - Write a summary.
- A random sample of 406 college freshmen found that 295 bought most of their textbooks from the college's bookstore. A random sample of 772 college seniors found that 537 bought their textbooks from the college's bookstore. You wish to test the claim that the proportion of all freshmen who purchase most of their textbooks from the college's bookstore is greater than the proportion of all seniors at a significance level of $\alpha = 0.01$.
 - A researcher wants to see if there is a difference in the proportion of on-time flights for two airlines. Test the claim using $\alpha = 0.10$.

	United Airlines	Alaska Airlines
Number of flights	350	400
Number of on-time flights	215	280

- TDaP is a booster shot that prevents Diphtheria, Tetanus, and Pertussis in adults and adolescents. The shot should be administered every 8 years in order for it to remain effective. A random sample of 500 people living in a town that experienced a pertussis outbreak this year were divided into two groups. Group 1 was made up of 132 individuals who had not had the TDaP booster in the past 8 years, and Group 2 consisted of 368 individuals who had. In Group 1, 15 individuals caught pertussis during the outbreak, and in Group 2, 11 individuals caught pertussis. Is there evidence to suggest that the proportion of individuals who caught pertussis and were not up to date on their booster shot is significantly higher than those that were? Test at the 0.05 level of significance.
- To determine whether various antismoking campaigns have been successful, annual surveys are conducted. Randomly selected individuals are asked whether they smoke. The responses for this year had 163 out of 662 who smoked. Ten years ago, the survey found 187 out of 695 who smoked. Can we infer that the proportion of smokers has declined from 10 years ago? Use $\alpha = 0.10$.
- The makers of a smartphone have received complaints that the facial recognition tool often does not work, or takes multiple attempts to finally unlock the phone. The company upgraded to a new version and are claiming the tool has improved. To test the claim, a critic takes a random sample of 75 users of the old version (Group 1) and 80 users of the new version (Group 2). They find that the facial recognition tool works on the first try 56% of the time in the old version and 70% of the time in the new version. Can it be concluded that the new version is performing better? Test at $\alpha=0.10$.
- A random sample of 54 people who live in a city were selected and 16 identified as a "dog person." A random sample of 84 people who live in a rural area were selected and 34 identified as a "dog person." Test the claim that the proportion of people who live in a city and identify as a "dog person" is significantly different from the proportion of people who live in a rural area and identify as a "dog person" at the 10% significance level. Use the confidence interval method.
- In a sample of 80 faculty from Portland State University, it was found that 90% were union members, while in a sample of 96 faculty at University of Oregon, 75% were union members. Find the 95% confidence interval for the difference in the proportions of faculty that belong to the union for the two universities.

For exercises 8-9, show all 5 steps for hypothesis testing:

- State the hypotheses.
 - Compute the test statistic.
 - Compute the critical value or p-value.
 - State the decision.
 - Write a summary.
8. A manager wishes to see if the time (in minutes) it takes for their workers to complete a certain task will decrease when they are allowed to wear earbuds at work. A random sample of 20 workers' times was collected before and after. Test the claim that the time to complete the task has decreased at a significance level of $\alpha = 0.01$. For the context of this problem, $\mu_D = \mu_{\text{before}} - \mu_{\text{after}}$ where the first data set represents before measurement and the second data set represents the after measurement. Assume the population is normally distributed. You obtain the following sample data.

Before	After	Before	After
69	62.3	61.7	56.8
71.5	61.6	55.9	44.7
39.3	21.4	56.8	50.6
67.7	60.4	71	63.4
38.3	47.9	80.6	68.9
85.9	77.6	59.8	35.5
67.3	75.1	72.1	77
59.8	46.3	49.9	38.4
72.1	65	56.2	55.4
79	83	63.3	51.6

9. An adviser is testing out a new online learning module for a placement test. Test the claim that on average the new online learning module increased placement scores at a significance level of $\alpha = 0.05$. For the context of this problem, $\mu_D = \mu_{\text{before}} - \mu_{\text{after}}$ where the first data set represents the after test scores and the second data set represents before test scores. Assume the population is normally distributed. You obtain the following paired sample of 19 students that took the placement test before and after the learning module.

Before	After	Before	After	Before	After
55.8	57.1	30.6	35.2	26.8	28.6
51.7	58.3	53	46.7	11.4	14.5
76.6	83.6	21	22.5	56.3	43.7
47.5	49.5	58.5	47.7	46.1	57
48.6	51.1	42.6	51.5	72.8	66.1
11.4	20.6	61.2	76.6	42.2	38.1
				51.3	42.4

10. A veterinary nutritionist developed a diet for overweight dogs. The total volume of food consumed remains the same, but half of the dog food is replaced with a low-calorie “filler” such as green beans. Ten overweight dogs were randomly selected from her practice and were put on this program. Their initial weights were recorded, and then the same dogs were weighed again after 4 weeks. Compute the 99% confidence interval. Can it be concluded that the dogs lost weight? Assume the populations are normally distributed and the groups are dependent.

Before	41	25	50	12	57	62	39	15	52	31
After	35	28	50	11	53	58	36	14	48	27

11. Doctors developed an intensive intervention program for obese patients with heart disease. Subjects with a BMI of 30 kg/m^2 or more with heart disease were assigned to a three-month lifestyle change of diet and exercise. Patients' Left Ventricle Ejection Fraction (LVEF) are measured before and after intervention. Assume that LVEF measurements are normally distributed.

Before	After
44	56
49	58
50	64
49	60
57	63
62	71
39	49
41	51
52	60
42	55

- Find the 95% confidence interval for the mean of the differences.
- Using the confidence interval answer, did the intensive intervention program significantly increase the mean LVEF? Explain why.

For exercises 12-14, assume the populations are normally distributed and show all 5 steps for hypothesis testing:

- State the hypotheses.
 - Find the test statistic.
 - Find the critical value or p-value.
 - State the decision.
 - Write a summary.
12. A physician wants to see if there was a difference in the average smoker's daily cigarette consumption after wearing a nicotine patch. The physician sets up a study to track daily smoking consumption. They give the patients a placebo patch that did not contain nicotine for 4 weeks, then a nicotine patch for the following 4 weeks. Use the following computer output to test to see if there was a difference in the average smoker's daily cigarette consumption using $\alpha = 0.01$.

t-Test: Paired Two Sample for Means		
	<i>Placebo</i>	<i>Nicotine</i>
Mean	16.75	10.3125
Variance	64.46667	33.29583
Observations	16	16
Pearson Correlation	0.6105	
Hypothesized Mean Difference	0	
df	15	
t Stat	4.0119	
P(T<=t) one-tail	0.0006	
t Critical one-tail	2.6025	
P(T<=t) two-tail	0.0011	
t Critical two-tail	2.9467	

13. A researcher is testing reaction times between the dominant and non-dominant hand. They randomly start with different hands for 20 subjects and their reaction times for both hands are recorded in milliseconds. Use the following computer output to test to see if the reaction time is faster for the dominant hand using a 5% level of significance.

t-Test: Paired Two Sample for Means		
	<i>Non-Dominant</i>	<i>Dominant</i>
Mean	63.33	56.28
Variance	218.9643158	128.7522105
Observations	20	20
Pearson Correlation	0.9067	
Hypothesized Mean Difference	0	
df	19	
t Stat	4.7951	
P(T<=t) one-tail	0.0001	
t Critical one-tail	1.7291	
P(T<=t) two-tail	0.0001	
t Critical two-tail	2.0930	

14. A manager wants to see if it is worth going back for an MBA degree. They randomly sample 18 managers' salaries before and after undertaking an MBA degree and record their salaries in thousands of dollars. Assume salaries are normally distributed. Use the following computer output to test the claim that the MBA degree, on average, increases a manager's salary. Use a 10% level of significance.

t-Test: Paired Two Sample for Means		
	<i>Before</i>	<i>After</i>
Mean	52.66667	59.82778
Variance	112.7012	175.5551
Observations	18	18
Pearson Correlation	0.7464	
Hypothesized Mean Difference	0	
df	17	
t Stat	-3.4340	
P(T<=t) one-tail	0.0016	
t Critical one-tail	1.3334	
P(T<=t) two-tail	0.0033	
t Critical two-tail	1.7396	

For exercises 15-28, show all 5 steps for hypothesis testing:

- State the hypotheses.
 - Compute the test statistic.
 - Compute the critical value or p-value.
 - State the decision.
 - Write a summary.
15. A liberal arts college in New Hampshire implemented an online homework system for their introductory math courses and wanted to know whether the system improved test scores. In the Fall semester, homework was completed with pencil and paper, checking answers in the back of the book. In the Spring semester, homework was completed online – giving students instant feedback on their work. The results are summarized below. Is there evidence to suggest that the online system improves test scores? Use $\alpha = 0.05$. Assume the population variances are unequal.

	Fall Semester	Spring Semester
Number of Students	127	144
Mean Test Score	73.4	77.4
Sample Standard Deviation	10.2	11.1

16. Researchers conducted a study to measure the effectiveness of the drug Methylphenidate on patients diagnosed with attention-deficit hyperactivity disorder (ADHD). A total of 112 patients with ADHD were randomly split into two groups. Group 1 included 56 patients and they were each given a dose of 20 mg. of Methylphenidate daily. The 56 patients in Group 2 were given a daily placebo. The effectiveness of the drug was measured by testing the patients score on a behavioral test. Higher scores indicate more ADHD symptoms. Group 1 was found to have a mean improvement of 9.3 points with a standard deviation of 6.5 points. Group 2 had a mean improvement of 11.7 points with a standard deviation of 8.7 points. Is there evidence to suggest the patients taking Methylphenidate have improved the mean ADHD symptoms? Test at the 0.01 level of significance. Assume the population variances are unequal.
17. In Major League Baseball, the American League (AL) allows a designated hitter (DH) to bat in place of the pitcher, but in the National League (NL), the pitcher has to bat. However, when an AL team is the visiting team for a game against an NL team, the AL team must abide by the home team's rules and thus, the pitcher must bat. A researcher is curious if an AL team would score more runs for games in which the DH was used. She samples 20 games for an AL team for which the DH was used, and 20 games for which there was no DH. The data are below. Assume the population is normally distributed with unequal population variances. Is there evidence to suggest that AL team would score more runs for games in which the DH was used? Use $\alpha = 0.10$.

With Designated Hitter			
0	5	4	7
1	2	7	6
6	4	2	10
1	2	7	5
8	4	11	0

Without Designated Hitter			
3	6	5	2
12	4	0	1
6	3	7	8
4	0	5	1
2	4	6	4

18. A movie theater company wants to see if there is a difference in the average movie ticket sales in San Diego and Portland per week. They sample 20 sales from San Diego and 20 sales from Portland over a week. Test the claim using a 5% level of significance. Assume the population variances are unequal, the samples are independent, and that movie sales are normally distributed.

San Diego			
223	243	231	235
221	182	217	211
206	229	219	239
215	214	234	221
226	233	239	232

Portland			
233	228	209	214
219	212	214	222
226	216	223	220
226	219	221	223
219	211	218	224

19. A national food product company believes that it sells more frozen pizza during the winter months than during the summer months. Weekly samples of sales found the following statistics in volume of sales (in hundreds of pounds). Use $\alpha = 0.10$ to test the company's claim. Assume the populations are approximately normally distributed with unequal variances.

Season	n	\bar{x}	s
Winter	24	312.34	135
Summer	22	224.75	84.42

20. You are testing the claim that the mean GPA of students who take evening classes is less than the mean GPA of students who only take day classes. You sample 20 students who take evening classes, and the sample mean GPA is 2.74 with a standard deviation of 0.86. You sample 25 students who only take day classes, and the

sample mean GPA is 2.86 with a standard deviation of 0.54. Test the claim using a 10% level of significance. Assume the population standard deviations are unequal and that GPAs are normally distributed.

21. "Durable press" cotton fabrics are treated to improve their recovery from wrinkles after washing. "Wrinkle recovery angle" measures how well a fabric recovers from wrinkles. Higher scores are better. Here are data on the wrinkle recovery angle (in degrees) for a random sample of fabric specimens. Assume the populations are approximately normally distributed with unequal variances. A manufacturer believes that the mean wrinkle recovery angle for Hylite is better. A random sample of 20 Permafresh (group 1) and 25 Hylite (group 2) were measured. Test the claim using a 10% level of significance.

Permafresh			
124	139	164	142
144	102	131	118
136	127	137	148
117	137	147	129
133	137	148	135

Hylite				
139	146	139	139	146
131	138	138	132	142
133	142	138	137	134
146	137	138	138	133
139	140	141	140	141

22. A large fitness center manager wants to test the claim that the mean delivery time for REI is faster than the delivery time for Champs Sports. The manager randomly samples 30 REI delivery times and finds a mean of 3.05 days with a standard deviation of 0.75 days. The manager randomly selects 30 Champs Sports delivery times and finds a mean delivery time of 3.262 days with a standard deviation of 0.27 days. Test the claim using a 5% level of significance. Assume the populations variances are unequal.
23. A new over-the-counter medicine to treat a sore throat is to be tested for effectiveness. The makers of the medicine take two random samples of 25 individuals showing symptoms of a sore throat. Group 1 receives the new medicine and Group 2 receives a placebo. After a few days on the medicine, each group is interviewed and asked how they would rate their comfort level 1-10 (1 being the most uncomfortable and 10 being no discomfort at all). The results are below. Is there sufficient evidence to conclude the mean scores from Group 1 is more than Group 2? Test at $\alpha = 0.01$. Assume the populations are normally distributed and have unequal variances.

Group 1				
3	5	6	7	5
3	4	5	7	7
3	2	5	8	8
7	7	8	4	8
4	8	3	9	10

Group 2				
4	5	8	3	5
2	7	8	2	4
1	2	2	3	2
1	3	5	5	1
6	4	7	8	1

24. Two competing fast food restaurants advertise that they have the *fastest* wait time from when you order to when you receive your meal. A curious critic takes a random sample of 40 customers at each restaurant to test the claim. They find that Restaurant A has a sample mean wait time of 2.25 minutes with a standard deviation of 0.35 minutes and Restaurant B has a sample mean wait time of 2.15 minutes with a standard deviation of 0.57 minutes in wait time. Can they conclude that the mean wait time is significantly different for the two restaurants? Test at $\alpha = 0.05$. Assume the population variances are unequal.
25. In a random sample of 60 pregnant women with preeclampsia, their systolic blood pressure was taken right before beginning to push during labor. The mean systolic blood pressure was 174 with a standard deviation of 12. In another random sample of 80 pregnant women without preeclampsia, there was a mean systolic blood pressure of 133 and a standard deviation of 8 when the blood pressure was also taken right before beginning to push. Is there sufficient evidence to conclude that women with preeclampsia have a higher mean blood pressure in the late stages of labor? Test at the 0.01 level of significance. Assume the population variances are unequal.
26. The manager at a pizza place has been getting complaints that the auto-fill soda machine is either under filling or over filling their cups. The manager took a random sample of 20 fills from her machine, and a random sample of 20 fills from another branch of the restaurant that has not been having complaints. From her machine,

she found a sample mean of 11.5 oz. with a standard deviation of 1.3 oz. and from the other restaurant's machine she found a sample mean of 10.95 oz. with a standard deviation of 0.65 oz. At the 0.05 level of significance, does it seem her machine has a significantly different mean than the other machine? Use the confidence interval method. Assume the populations are normally distributed with unequal variances.

27. In a study that followed a group of students who graduated from high school in 2015, each was monitored in progress made toward earning a bachelor's degree. The group was divided in two – those who started at community college and later transferred to a four-year college, and those that started out in a four-year college as freshmen. That data below summarizes the findings. Is there evidence to suggest that community college transfer students take longer to earn a bachelor's degree? Use $\alpha = 0.05$. Assume the population variances are unequal.

	Community College Transfers	Non-Transfers
Number of Students	317	1,297
Mean Time to Graduate (in years)	5.09	4.68
Sample Standard Deviation	1.896	1.097

28. In a random sample of 50 Americans five years ago, the average credit card debt was \$5,798 with a standard deviation of \$1,154. In a random sample of 50 Americans in the present day, the average credit card debt is \$6,511, with a standard deviation of \$1,645. Using a 0.05 level of significance, test if there is a difference in credit card debt today versus five years ago. Assume the population variances are unequal.
29. A researcher is curious what year in college students make use of the gym at a university. They take a random sample of 30 days and count the number of sophomores and seniors who use the gym each day. Is there evidence to suggest that a difference exists in gym usage based on year in college? Construct a 90% confidence interval for the data below to decide. Assume the population variances are unequal.

Sophomores				
189	203	167	154	217
209	198	143	208	220
188	197	165	207	231
201	177	186	193	201
190	165	180	245	200
199	155	165	188	187

Seniors				
209	199	186	210	221
204	214	230	170	197
190	201	165	183	235
187	199	189	194	197
192	195	200	211	205
200	190	218	210	229

30. A researcher takes sample temperatures in Fahrenheit of 17 days from New York City and 18 days from Phoenix. Test the claim that the mean temperature in New York City is different from the mean temperature in Phoenix. Assume the populations are approximately normally distributed with unequal variances. You obtain the following two samples of data. Compute the 90% confidence interval for the difference in the mean temperatures.

New York City		
98	85.4	87.7
95.5	75.4	86.1
92.2	79.5	74.3
102	82.4	85.2
85.4	64.3	82.8
80	65.5	

Phoenix		
106.8	82	120.1
98.6	72	114.4
91.5	115.2	93.7
82	94.2	89.7
97.7	72	104.7
64.9	86.8	76.6

31. An employee at a large company is told that the mean starting salary at her company differs based on level of experience. The employee is skeptical and randomly samples 30 new employees with less than 5 years of experience and categorizes them as Group 1 and 30 new employees with 5 years of experience or more and categorizes them as Group 2. In Group 1, she finds the sample mean starting salary to be \$50,352 with a standard deviation of \$4,398.10. Group 2 has a sample mean starting salary of \$52,391 with a standard

deviation of \$7,237.32. Compute the 90% confidence interval for the difference in the mean starting salaries. Assume the populations are normally distributed with unequal variances.

32. Two random samples are taken from private and public universities (out-of-state tuition) around the nation. The yearly tuition is recorded from each sample and the results can be found below. Find the 95% confidence interval for the mean difference between private and public institutions. Assume the populations are normally distributed and have unequal variances.

Private Institutions			
43,120	34,750	29,498	30,129
28,190	44,897	31,980	33,980
34,490	32,198	22,764	47,909
20,893	18,432	54,190	32,200
42,984	33,981	37,756	38,120

Public Institutions			
25,469	21,871	22,650	28,745
19,450	24,120	29,143	30,120
18,347	27,450	25,379	21,190
28,560	29,100	23,450	21,540
32,592	21,870	23,871	26,346

For exercises 33-39, show all 5 steps for hypothesis testing:

- State the hypotheses.
 - Compute the test statistic.
 - Compute the critical value or p-value.
 - State the decision.
 - Write a summary.
33. A professor wants to know if there is a difference in comprehension of a lab assignment among students depending if the instructions are given all in text, or if they are given primarily with visual illustrations. She randomly divides her class into two groups of 15, gives one group instructions in text and the second group instructions with visual illustrations. The following data summarizes the scores the students received on a test given after the lab. Assume the populations are normally distributed with equal variances. Is there evidence to suggest that a difference exists in the mean comprehension of the lab based on the test scores? Use $\alpha = 0.10$.

Text		
57.3	87.3	67.2
45.3	75.2	54.4
87.1	88.2	93.0
61.2	67.5	89.2
43.1	86.2	52.0

Visual Illustrations		
59.0	76.7	88.2
57.6	78.2	43.8
72.9	64.4	97.1
83.2	89.0	95.1
64.0	72.9	84.1

34. The manager at a local coffee shop is trying to decrease the time customers wait for their orders. He wants to find out if keeping multiple registers open will make a difference. He takes a random sample of 30 customers when only one register is open and finds that they wait an average of 6.4 minutes to reach the front with a standard deviation of 1.34 minutes. He takes another random sample of 35 customers when two registers are open and finds that they wait an average of 4.2 minutes to reach the front with a standard deviation of 1.21 minutes. He takes both his samples during peak hours to maintain consistency. Can it be concluded at the 0.05 level of significance that mean wait time is less with two registers open? Assume the population variances are equal.
35. The CEO of a large manufacturing company is curious if there is a difference in productivity level of her warehouse employees based on the region of the country the warehouse is located in. She randomly selects 35 employees who work in warehouses on the East Coast and 35 employees who work in warehouses in the Midwest and records the number of parts shipped out from each for a week. She finds that East Coast group ships an average of 1,287 parts and a standard deviation of 348. The Midwest group ships an average of 1,449 parts and a standard deviation of 298. Using a 0.01 level of significance, test if there is a difference in mean productivity level. Assume the population variances are equal.

36. The mean speeds (mph) of fastball pitches from two different left-handed baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. Assume the populations are approximately normal distributed with equal variances. Scouts believe that Brandon Eisert pitches a speedier fastball. Test the scouts' claim that Eisert's mean speed is faster at the 5% level of significance?

Pitcher	Sample Mean	Sample Standard Deviation
Nico Tellache	87	5
Brandon Eisert	91	7

37. A large shoe company is interested in knowing if the amount of money a customer is willing to pay on a pair of shoes is different depending on location. They take a random sample of 50 single-pair purchases from Southern states and another random sample of 50 single-pair purchases from Midwestern states and record the cost for each. The results can be found below. At the 0.05 level of significance, is there evidence that the mean cost differs between the Midwest and the South? Assume the population variances are equal.

Midwest (cost in dollars)					South (cost in dollars)				
70	43	21	62	45	73	75	34	59	81
60	23	15	37	66	80	17	18	65	27
65	38	30	46	64	62	32	60	60	56
71	54	51	61	82	30	28	31	17	34
33	79	28	84	63	36	54	48	85	54
68	72	78	43	84	46	55	30	41	53
78	80	24	80	16	80	16	67	36	39
82	45	38	84	84	22	16	38	46	50
73	23	36	69	78	16	49	43	54	27
76	71	38	46	18	83	50	57	51	68

38. A physical therapist believes that at 30 years old adults begin to decline in flexibility and agility. To test this, they randomly sample 35 of their patients who are less than 30 years old and 32 of their patients who are 30 years or older and measure each patient's flexibility in the Sit-and-Reach test. The results are below, higher numbers indicate more flexibility. Is there evidence to suggest that adults under the age of 30 years are more flexible? Use $\alpha = 0.05$. Assume the population variances are equal.

	Less Than 30 Years	30 Years or Older
<i>n</i>	35	32
Mean Sit-and-Reach Score	20.46	18.84
Sample Standard Deviation	2.237	2.118

39. In a random sample of 100 college students, 47 were sophomores and 53 were seniors. The sophomores reported spending an average of \$37.03 per week going out for food and drinks with a standard deviation of \$7.23, while the seniors reported spending an average of \$52.94 per week going out for food and drinks with a standard deviation of \$12.33. Compute the 90% confidence interval for difference in the mean amount spent on food and drinks between sophomores and seniors? Assume the population variances are equal.
40. A pet store owner believes that dog owners, on average spend a different amount on their pets compared to cat owners. The owner randomly records the sales of 40 customers who said they only owned dogs and found the mean of the sales of \$56.07 with a standard deviation of \$24.50. The owner randomly records the sales of 40 customers who said they only owned cats and found a mean of the sales of \$52.92 with a standard deviation of \$23.53. Compute the 95% confidence interval to test the pet store owner's claim. Assume the population variances are equal.

41. Two groups of students are given a problem-solving test, and the results are compared. Assume the populations are normally distributed with equal variances. Compute the 95% confidence interval for the difference of the mean scores.

Mathematics Majors	Computer Science Majors
$\bar{x}_1 = 83.6$	$\bar{x}_2 = 79.2$
$s_1 = 4.3$	$s_2 = 3.8$
$n_1 = 16$	$n_2 = 20$

42. A survey found that the average daily cost to rent a car in Los Angeles is \$103.24 and in Las Vegas is \$97.24. The data were collected from two random samples of 40 in each of the two cities and the sample standard deviations are \$5.98 for Los Angeles and \$4.21 for Las Vegas. At the 0.05 level of significance, construct a confidence interval for the difference in the means and then decide if there is a significant difference in the rates between the two cities using the confidence interval method. Assume the population variances are equal.

It is an important and popular fact that things are not always what they seem. For instance, on the planet Earth, man had always assumed that he was more intelligent than dolphins because he had achieved so much – the wheel, New York, wars and so on – whilst all the dolphins had ever done was muck about in the water having a good time. But conversely, the dolphins had always believed that they were far more intelligent than man – for precisely the same reasons.
(Adams, 2002)

Chapter 9

Chi-Square Tests



- 9.1 Introduction to the Chi-Square Distribution
- 9.2 Goodness of Fit Test
- 9.3 Test for Independence

9.1 Introduction to the Chi-Square Distribution

This chapter covers statistical methods used to determine whether an observed data set follows an expected distribution or if there is a relationship between two categorical variables. The goodness of fit test involves comparing the observed data with the expected data under a specific hypothesis. The hypothesis represents the assumption or belief about the distribution or pattern that the data should follow. By conducting this test, we can evaluate whether the observed data deviates significantly from the expected data, indicating that the observed frequencies are unlikely to have occurred by chance. The test for independence is used to determine whether there is a relationship or association between two categorical variables. These tests also assess whether the observed frequencies of the variables in a sample differ significantly from the expected frequencies under the assumption of independence.

These tests are commonly used in various fields, including biology, epidemiology, finance, genetics, market research, quality control, and social sciences. They allow researchers to examine relationships between variables and identify patterns or dependencies that may exist. For example, a chi-square test for independence can be used to investigate whether there is an association between gender and voting preference, or to explore whether there is a relationship between smoking status and the development of a particular disease.

These tests use a statistical measure known as the chi-square statistic, which quantifies the standard deviation between the observed and expected frequencies.

A χ^2 -distribution (chi-square, pronounced “ki-square”) is another special type of distribution for a continuous random variable. The sampling distribution for a variance and standard deviation follows a chi-square distribution.

Properties of the χ^2 -distribution density curve:

1. Right skewed starting at zero.
2. The center and spread of a χ^2 -distribution are determined by the degrees of freedom with a mean = df and standard deviation = $\sqrt{2df}$.
3. Chi-square variables cannot be negative.
4. As the degrees of freedom increase, the χ^2 -distribution becomes normally distributed for $df > 50$. Figure 9-1 shows χ^2 -distributions for df of 2, 4, 10, and 30.
5. The total area under the curve is equal to 1, or 100%.

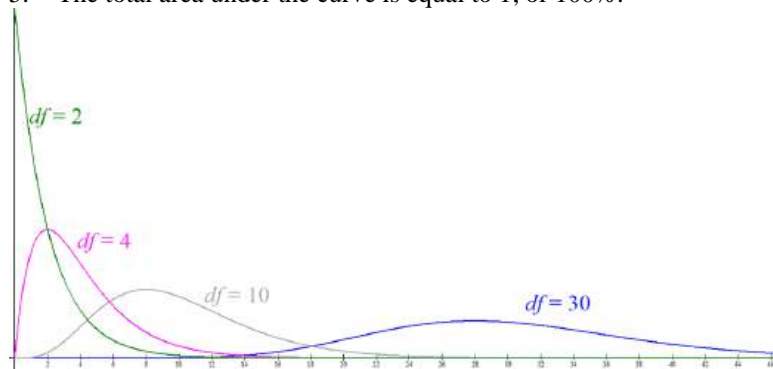


Figure 9-1

We will use the χ^2 -distribution for hypothesis testing later in this chapter. For now, we are just learning how to find a critical value χ^2_α .

The symbol χ^2_α is the critical value on the χ^2 -distribution curve with area $1 - \alpha$ to the left of the critical value and area α to the right of the critical value, as shown below in Figure 9-2.

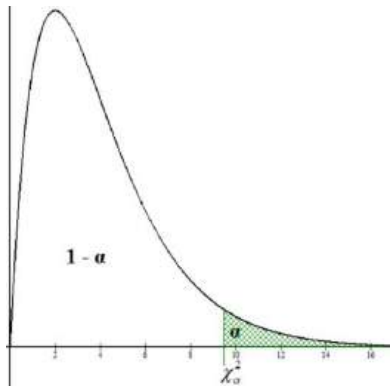


Figure 9-2

Use technology to compute the critical value for the χ^2 -distribution.

TI-84: Use the INVCCHI2 program downloaded at Rachel Webb's website: <http://MostlyHarmlessStatistics.com>. Start the program and enter the area α and the df when prompted.

TI-89: Go to the [Apps] **Stat/List Editor**, then select F5 [DISTR]. This will get you a menu of probability distributions. Arrow down to **Inverse > Inverse Chi-Square** and press [ENTER]. Enter the area $1 - \alpha$ to the left of the χ value and the df into each cell. Press [ENTER].

Excel: =CHISQ.INV($1 - \alpha$, df) or =CHISQ.INV.RT(α , df)

Alternatively, use the following online calculator: <https://homepage.divms.uiowa.edu/~mbogner/applets/chisq.html>.

Example 9-1: Compute the critical value χ_α^2 for a $\alpha = 0.05$ and $df = 6$.

Solution: Start by drawing the curve and determining the area in the right-tail as shown in Figure 9-3. Then use technology to find the critical value.

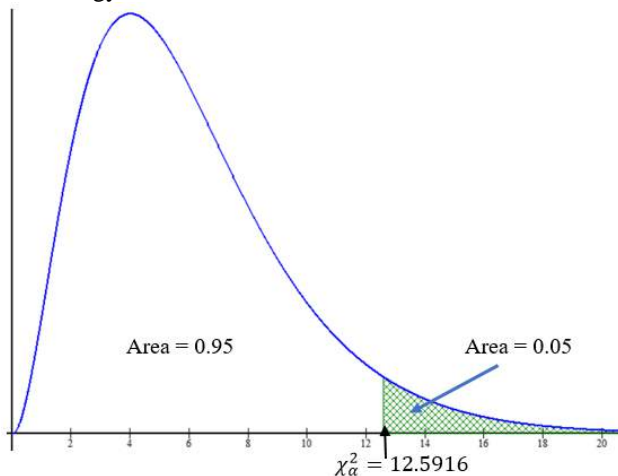
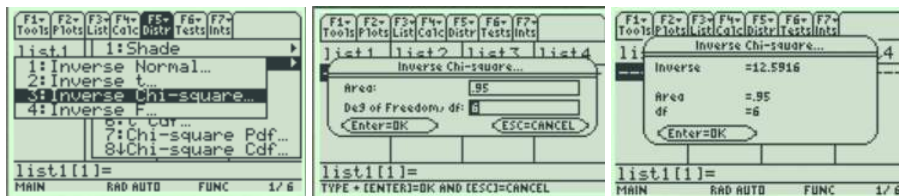


Figure 9-3

In Excel there are two options. Use =CHISQ.INV(area in left-tail, df) or right-tail =CHISQ.INV.RT(area in right-tail, df). For this example, then we would have $\chi_\alpha^2 = \text{CHISQ.INV}(0.95,6)$ or =CHISQ.INV.RT(0.05,6) = 12.5916.

TI-89 use Distr > Inverse Chi-square with area 0.95 and $df = 6$.



9.2 Goodness of Fit Test

The χ^2 **goodness-of-fit test** can be used to test the distribution of three or more proportions. A hypothesis test for three or more proportions is a statistical analysis used to compare the proportions of multiple groups or categories and determine if there are significant differences between them. This type of test uses categorical data and investigates whether there are variations in proportions across different groups or populations.

The primary objective of a goodness of fit test for three or more proportions is to assess whether the observed differences in proportions are statistically significant or merely due to chance. By conducting this test, researchers can gain insights into whether there are meaningful disparities among the groups being compared. Hypothesis testing for three or more proportions allows researchers to explore whether there are significant differences in proportions across multiple groups, aiding in decision-making, policy formulation, and identifying areas for further investigation or intervention.

To perform the test, a null hypothesis is established, assuming that there is no difference in the proportions of interest across the groups. The alternative hypothesis, on the other hand, asserts that at least one of the proportions differs significantly from the others. The test involves collecting data from each group, recording the observed frequencies, and assess the degree of discrepancy between the observed and expected frequencies.

The χ^2 -test is a statistical test for testing the goodness-of-fit of a variable. It can be used when the data are obtained from a random sample and when the expected frequency (E) from each category is 5 or more.

$$\text{The formula for the } \chi^2\text{-test statistic is: } \chi^2 = \sum \frac{(O-E)^2}{E}.$$

Use a right-tailed χ^2 -distribution with $df = k - 1$ where k = the number of categories.

With

O = the observed frequency (what was observed in the sample) and

E = the expected frequency (based on H_0 and the sample size) = $n \cdot p_0$ for each group.

H_0 : $p_1 = p_0, p_2 = p_0, \dots, p_k = p_0$.

H_1 : At least one proportion is different.

Example 9-2: An instructor claims that their student's grade distribution is different than the department's grade distribution. The department's grades have the following proportion of students who get A's is 35%, B's is 23%, C's is 25%, D's is 10% and F's is 7% in introductory statistics courses. For a sample of 250 introductory statistics students with this instructor, there were 80 A's, 50 B's, 58 C's, 38 D's, and 24 F's. Test the instructor's claim at the 5% level of significance.

Solution: This is a test for three or more proportions within a single population, so use the goodness-of-fit test. We will always use a right-tailed χ^2 -test. The hypotheses for this example would be:

H_0 : $p_A = 0.35, p_B = 0.23, p_C = 0.25, p_D = 0.10, p_F = 0.07$

H_1 : At least one proportion is different.

Even though there is an inequality in H_1 , the goodness-of-fit test is always a right-tailed test. This is because we are testing to see if there is a large variation between the observed versus the expected values. If the variance between the observed and expected values is large, then there is a difference in the proportions.

Also note that we do not write the alternative hypothesis as $p_A \neq 0.35, p_B \neq 0.23, p_C \neq 0.25, p_D \neq 0.10, p_F \neq 0.07$ since it could be that any one of these proportions is different. All of the proportions not equal to their hypothesized values is just one case.

There are $k = 5$ categories that we are comparing: A's, B's, C's, D's and F's.

The observed counts are the actual number of A's, B's, C's, D's and F's from the sample.

We must compute the expected count for each of the five categories. Find the expected counts by multiplying the expected proportion of A's, B's, C's, D's and F's by the sample size.

It will be helpful to make a table to organize the work.

Grade	A's	B's	C's	D's	F's
Observed (O)	80	50	58	38	24
Expected (E)	$0.35 \cdot 250 = 87.5$	$0.23 \cdot 250 = 57.5$	$0.25 \cdot 250 = 62.5$	$0.10 \cdot 250 = 25$	$0.07 \cdot 250 = 17.5$
$\frac{(O - E)^2}{E}$	$\frac{(80-87.5)^2}{87.5} = 0.6429$	$\frac{(50-57.5)^2}{57.5} = 0.9738$	$\frac{(58-62.5)^2}{62.5} = 0.324$	$\frac{(38-25)^2}{25} = 6.76$	$\frac{(24-17.5)^2}{17.5} = 2.4143$

The test statistic is the sum of this last row: $\chi^2 = \sum \frac{(O-E)^2}{E} = 0.6429 + 0.9783 + 0.324 + 6.76 + 2.4143 = 11.1195$.

The critical value for a right-tailed χ^2 -test with $df = k - 1 = 5 - 1 = 4$ is found by finding the area in the χ^2 -distribution using your calculator or Excel. Use $\alpha = 0.05$ area in the right-tail, to get the critical value of $\chi^2_{\alpha} = \text{CHISQ.INV.RT}(0.05,4) = 9.4877$. Draw and label the curve as shown in Figure 9-4.

The test statistic of $\chi^2 = 11.1195 > \chi^2_{\alpha} = 9.4877$ and is in the rejection area, so our decision is: Reject H_0 .

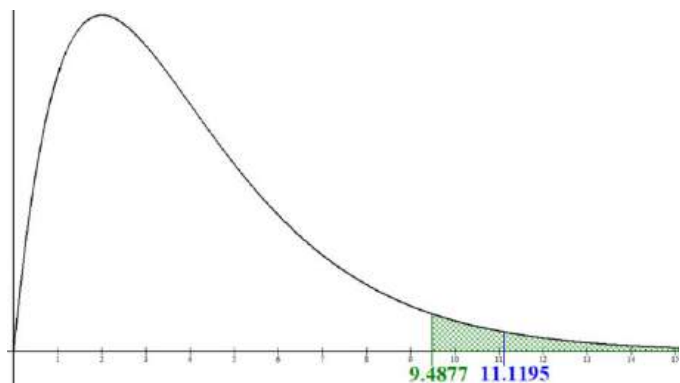


Figure 9-4

There is sufficient evidence to support the claim that the proportion of students who get A's, B's, C's, D's and F's in introductory statistics courses for this instructor is different than the department's proportions of 35%, 23%, 25%, 10% and 7% respectively.

If we were asked to find the p-value, you would just find the area to right of the test statistic (always a right-tailed test) using your calculator or Excel $=\text{CHISQ.DIST.RT}(11.1195,4) = 0.0253$.

```
χ²cdf(11.1195,1E
99,4)
.0252533235
```

This gives a p-value = 0.0252 which is less than $\alpha = 0.05$, therefore reject H_0 .

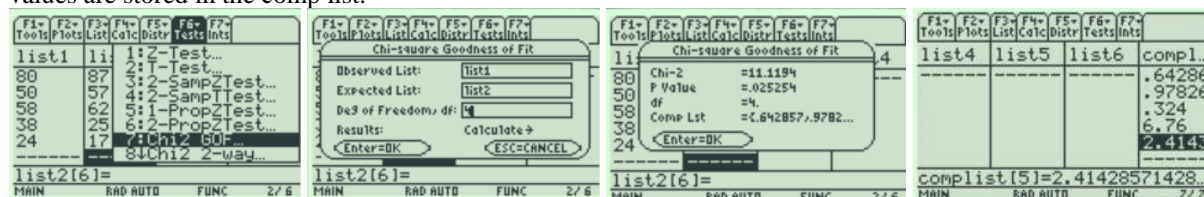
You can use the GOF shortcut function on your calculator to get a p-value; see directions below. If you get the program from your instructor or the website for your TI-83, you can also have the calculator find the $\frac{(O-E)^2}{E}$ values. The TI-84 and 89 already does this.

TI-84: Note: For the TI-83 download a GOF program from <http://MostlyHarmlessStatistics.com>. Only newer TI-84 operating systems have a calculator shortcut key for GOF. Use the same GOF program as for the TI-83 if your 84 does not have the χ^2 GOF-Test.

Before you start, write down your observed and expected values. Select Stat, then Calc. Type in the observed values into list 1, and the expected values into list 2. Select Stat, then Tests. Go down to option D: χ^2 GOF-Test. Choose L₁ for the Observed category and L₂ for the Expected category, type in your degrees of freedom ($df = k - 1$), and then select Calculate. The calculator returns the χ^2 -test statistic and the p-value. Right arrow to see the rest of the (O-E)²/E values.



TI-89: Go to the [Apps] Stat/List Editor, then type in the observed values into list 1, and the expected values into list 2. Press [2nd] then F6 [Tests], then select 7: Chi-2GOF. Type in the list names and the degrees of freedom ($df = k - 1$). Then press the [ENTER] key to calculate. The calculator returns the χ^2 -test statistic and the p-value. The (O-E)²/E values are stored in the comp list.



Example 9-3: A research company is looking to see if the proportion of consumers who purchase a cereal is different based on shelf placement. They have four locations: Bottom Shelf, Middle Shelf, Top Shelf, and Aisle End Shelf. Test to see whether there is a preference among the four shelf placements. Use the p-value method with $\alpha = 0.05$.

Shelf	Bottom	Middle	Top	End
Observed	45	67	55	73

Solution: The hypotheses can be written as a sentence or as proportions. If you use proportions, note that there are no percentages given. We would expect that each shelf placement be the same if there was no preference. There are 4 categories, so $p_0 = \frac{1}{4} = 0.25$ or 25% for each placement.

$$H_0: p_B = 0.25, p_M = 0.25, p_T = 0.25, p_E = 0.25$$

$$H_1: \text{At least one proportion is different.}$$

It is also acceptable to write the hypotheses as a sentence.

$$H_0: \text{Proportion of cereal sales is equally distributed across the four shelf placements.}$$

$$H_1: \text{Proportion of cereal sales is not equally distributed across the four shelf placements.}$$

Find the expected values. Total all the observed values to get the sample size: $n = 45 + 67 + 55 + 73 = 240$. Then take the $n \cdot p_0$ to get $240 \cdot (0.25) = 60$. The expected value for each group is 60.

$$\begin{aligned} \text{Compute the test statistic: } \chi^2 &= \sum \frac{(O-E)^2}{E} = \frac{(45-60)^2}{60} + \frac{(67-60)^2}{60} + \frac{(55-60)^2}{60} + \frac{(73-60)^2}{60} \\ &= 3.75 + 0.816667 + 0.416667 + 2.816667 = 7.8 \end{aligned}$$

Check your work using technology and find the p-value.

TI-84: The degrees of freedom are the number of groups minus one = $df = k - 1 = 3$.



You can scroll right by selecting the right arrow button to see the rest of the contribution values.

Excel: In Excel the p-value is found by the formula =CHISQ.DIST.RT(7.8,3). On the TI-Calculator, use the χ^2 GOF-Test shortcut. The p-value = 0.05033 which is larger than $\alpha = 0.05$, therefore do not reject H_0 .

There is not enough evidence to support the claim that cereal shelf placement makes a statistically significant difference in the proportion of sales at the 5% level of significance.

9.3 Test for Independence

Use the chi-square test for independence to test the independence of two categorical variables. Remember, qualitative data is collected on individuals that are categories or names. Then you would count how many of the individuals had particular qualities. An example is that there is a theory that there is a relationship between breastfeeding and having autism spectrum disorder (ASD). To determine if there is a relationship, researchers could collect the time-period that a mother breastfed her child and if that child was diagnosed with ASD. Then you would have a table containing this information. Now you want to know if each cell is independent of each other cell. Remember, independence says that one event does not affect another event. Here it means that having ASD is independent of being breastfed. What you really want is to see if they are dependent (not independent). In other words, does one affect the other? If you were to do a hypothesis test, this is your alternative hypothesis and the null hypothesis is that they are independent. There is a hypothesis test for this and it is called the **chi-square test for independence**.

There is only a right-tailed test for testing the independence between two variables:

H_0 : Variable 1 and Variable 2 are independent (unrelated).

H_1 : Variable 1 and Variable 2 are dependent (related).

Finding the test statistic involves several steps. First, the data is collected, counted, and then organized into a contingency table. These values are known as the observed frequencies, which the symbol for an observed frequency is O . Total each row and column.

The null hypothesis is that the two variables are independent. If two events are independent then $P(B) = P(B | A)$ and we can use the multiplication rule for independent events, to calculate the probability that variable A and B as the $P(A \text{ and } B) = P(A) \cdot P(B)$. Remember in a hypothesis test, you assume that H_0 is true, the two variables are assumed to be independent.

$P(A \text{ and } B) = P(A) \cdot P(B)$ if A and B are independent.

$$P(A) \cdot P(B) = \frac{\text{Number of ways A can happen}}{\text{Total number of individuals}} \cdot \frac{\text{Number of ways B can happen}}{\text{Total number of individuals}} = \frac{\text{Row Total}}{n} \cdot \frac{\text{Column Total}}{n}$$

	Variable 1		
Variable 2	B	Not B	Total
A	P(A and B)	P(A and Not B)	Row Total for A
Not A	P(Not A and B)	P(Not A and Not B)	Row Total for Not A
Total	Column Total for B	Column Total for Not B	n

Now you want to find out how many individuals you expect to be in a certain cell. To find the expected frequencies, you just need to multiply the probability of that cell times the total number of individuals. Do not round the expected frequencies.

$$\text{Expected frequency (cell A and B)} = E(A \text{ and } B) = n \left(\frac{\text{Row Total}}{n} \cdot \frac{\text{Column Total}}{n} \right) = \frac{\text{Row Total} \cdot \text{Column Total}}{n}$$

If the variables are independent, the expected frequencies and the observed frequencies should be the same.

The test statistic here will involve looking at the difference between the expected frequency and the observed frequency for each cell. Then you want to find the “total difference” of all of these differences. The larger the total, the smaller the chances that you could find that test statistic given that the assumption of independence is true. That means that the assumption of independence is not true.

How do you find the test statistic? First, compute the differences between the observed and expected frequencies. Because some of these differences will be positive and some will be negative, you need to square these differences. These squares could be large just because the frequencies are large, so you need to divide by the expected frequencies to scale them. Then finally add up all of these fractional values. This process finds the variance and we use a chi-square distribution to find the critical value or p-value. Hence, sometimes this test is called a chi-square test.

The χ^2 -test is a statistical test for testing the independence between two variables. It can be used when the data are obtained from a random sample, and when the expected value (E) from each cell is 5 or more.

$$\text{The formula for the } \chi^2\text{-test statistic is: } \chi^2 = \sum \frac{(O-E)^2}{E}$$

Use χ^2 -distribution with degrees of freedom

$$df = (\text{the number of rows} - 1)(\text{the number of columns} - 1), \text{ that is, } df = (R - 1)(C - 1).$$

Where, O = the observed frequency (sample results) and
 E = the expected frequency (based on H_0 and the sample size).

Example 9-4: Is there a relationship between autism spectrum disorder (ASD) and breastfeeding? To determine if there is, a researcher asked mothers of ASD and non-ASD children to say what time-period they breastfed their children. Does the data provide enough evidence to show that breastfeeding and ASD are independent? Test at the 1% level.

ASD	Length of Breastfeeding				Total
	None	Less than 2 months	2 to 6 months	Over 6 months	
Yes	241	198	164	215	818
No	20	25	27	44	116
Total	261	223	191	259	934

(Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006.)

Solution: The question is asking if breastfeeding and ASD are independent. The correct hypothesis is:

H_0 : Autism spectrum disorder and length of breastfeeding are independent.

H_1 : Autism spectrum disorder and length of breastfeeding are dependent.

There are 2 rows and 4 columns of data. We must compute the Expected count for each of the $2 \times 4 = 8$ cells.

The expected counts for each cell are found by the formula:

$$\text{Expected Value} = \frac{\text{Row Total} \cdot \text{Column Total}}{\text{Grand Total}}$$

It will be helpful to make a table for the expected counts and another one for each of the $\frac{(O-E)^2}{E}$ values to aid in computing the test statistic.

Observed	ASD	None	Below 2	2-6	Over 6	Total
	Yes	241	198	164	215	818
	No	20	25	27	44	116
	Total	261	223	191	259	934

Expected	ASD	None	Below 2	2-6	Over 6	Total
	Yes	$\frac{818 \cdot 261}{934} = 228.585$	$\frac{818 \cdot 223}{934} = 195.304$	$\frac{818 \cdot 191}{934} = 167.278$	$\frac{818 \cdot 259}{934} = 226.833$	818
	No	$\frac{116 \cdot 261}{934} = 32.415$	$\frac{116 \cdot 223}{934} = 27.696$	$\frac{116 \cdot 191}{934} = 23.722$	$\frac{116 \cdot 259}{934} = 32.167$	116
	Total	261	223	191	259	934

$\frac{(O-E)^2}{E}$	ASD	None	Below 2	2-6	Over 6	Total
	Yes	$\frac{(241-228.585)^2}{228.585} = 0.6743$	$\frac{(198-195.304)^2}{195.304} = 0.0372$	$\frac{(164-167.278)^2}{167.278} = 0.0642$	$\frac{(215-226.833)^2}{226.833} = 0.6173$	11.217
	No	$\frac{(20-32.415)^2}{32.415} = 4.7552$	$\frac{(25-27.696)^2}{27.696} = 0.2624$	$\frac{(27-23.722)^2}{23.722} = 0.4531$	$\frac{(44-32.167)^2}{32.167} = 4.3529$	

The test statistic is the sum of all eight $\frac{(O-E)^2}{E}$ values: $\chi^2 = \sum \frac{(O-E)^2}{E} = 11.217$.

The critical value for a right-tailed χ^2 -test with degrees of freedom $df = (R - 1)(C - 1) = (2 - 1)(4 - 1) = 3$ is found using a χ^2 -distribution $\alpha = 0.01$ right-tail area. The critical value is $\chi^2 = \text{CHISQ.INV.RT}(0.01, 3) = 11.3449$. See Figure 9-5.

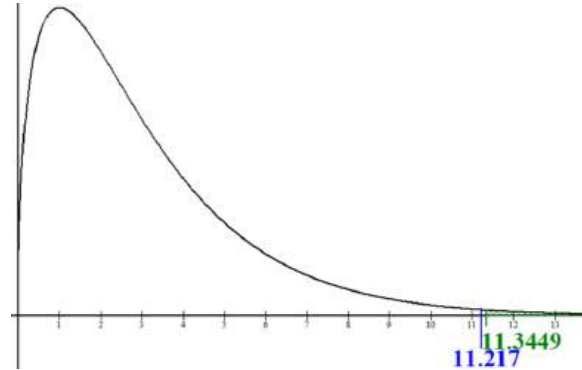
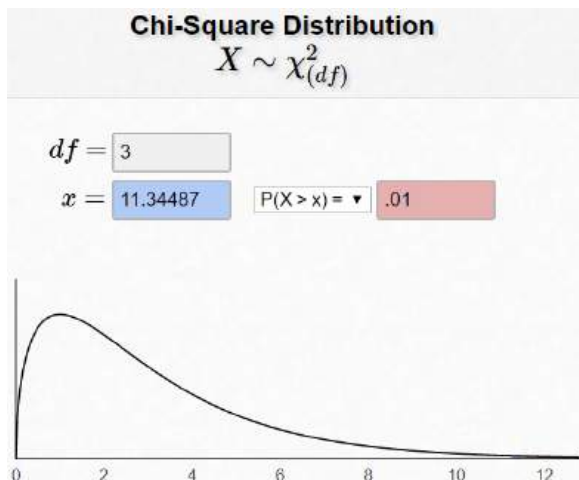


Figure 9-5

Alternatively, use the online calculator: <https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html>.



Since the test statistic $\chi^2 = 11.217$ is not in the rejection area our decision is to fail to reject H_0 .

There is not enough evidence to show a relationship between autism spectrum disorder and breastfeeding.

If we were asked to find the p-value, you would just find the area to right of the test statistic (always a right-tailed test) using your calculator or Excel. This gives a p-value = 0.0106 which is more than $\alpha = 0.01$, therefore we do not reject H_0 .



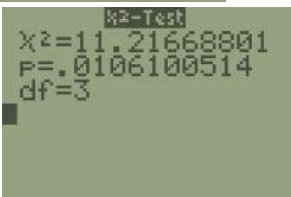
You can also use the χ^2 -Test shortcut keys on your calculator to get a p-value, see directions below.

TI-84: Press the [2nd] then [MATRX] key. Arrow over to the EDIT menu and 1:[A] should be highlighted, press the [ENTER] key. For a $m \times n$ contingency table, type in the number of rows(m) and the number of columns(n) at the top of the screen so that it looks like this: MATRIX[A] $m \times n$. For a 2×4 contingency table, the top of the screen would look like this MATRIX[A] 2×4 . As you press [ENTER] the table will automatically widen to the size you put in. Now enter all of the observed values in their proper positions. Then press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [C: χ^2 -Test] and press the [ENTER] key. Leave the default as Observed:[A] and Expected:[B], arrow down to [Calculate] and press the [ENTER] key.



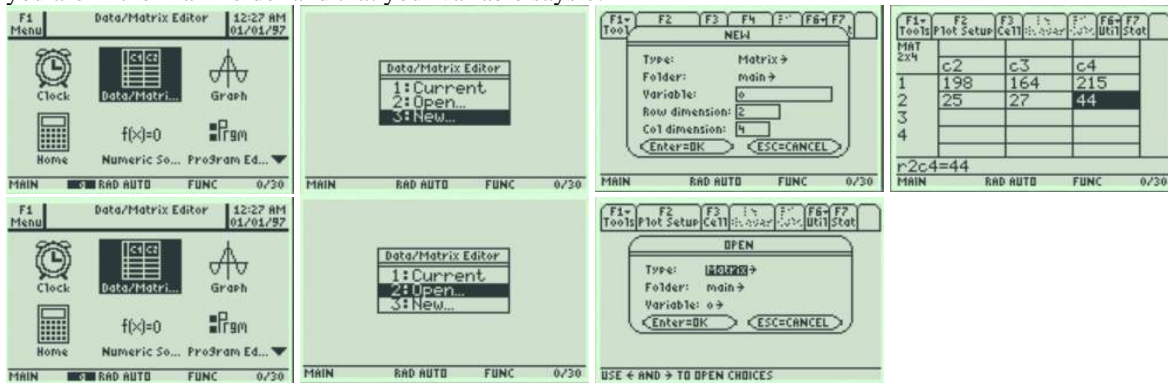
The calculator returns the χ^2 -test statistic and the p-value.

If you go back to the matrix menu [2nd] then [MATRX] key, arrow over to EDIT and choose 2:[B], you will see all of the expected values.

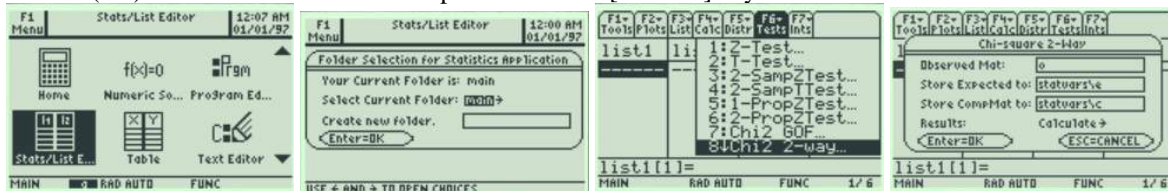


TI-89: First you need to create the matrix for the observed values: Press [Home] to return to the Home screen, press [Apps] and select Data/Matrix Editor. A menu is displayed, select 3:New. The New dialog box is displayed. Press the right arrow key to highlight 2:Matrix, and press [ENTER] to choose Matrix type. Press the down arrow key to highlight 1:Main, and press [ENTER], to choose main folder. Press the down arrow key, and then enter the letter o for the name in the Variable field. Enter 2 for Row dimension and 4 for Column dimension. Press [ENTER] to display the matrix editor. Enter the observed value (do not include total row or column). Important: Next time you use this test instead of option 3:New, choose 2: Open. The open dialog box is displayed. Press the right

arrow key to highlight 2:Matrix, and press [ENTER] to choose Matrix type. Press the down arrow key to make sure you are in the Main folder and that your variable says o.



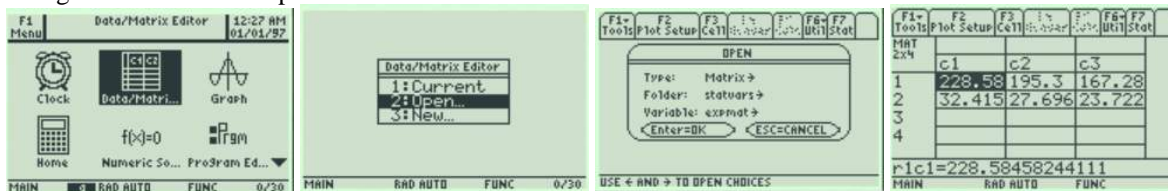
Press [Apps], and then select Stats/List Editor. To display the Chi-square 2-Way dialog box, press 2nd then F6 [Tests], then select 8: Chi-2 2-way. Enter in in the Observed Mat: o; leave the other rows alone: Store Expected to: statvars\c; Store CompMat to: statvars\c. This will store the expected values in the matrix folder statvars with the name expmat, and the $(o-e)^2/e$ values in the matrix compmat. Press the [ENTER] key to calculate.



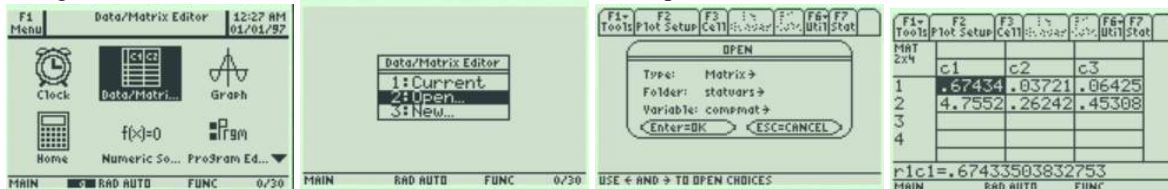
The calculator returns the χ^2 -test statistic and the p-value. If you go back to the matrix menu, you will see some of the expected and $(o-e)^2/e$ values.



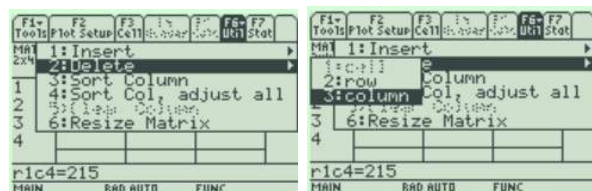
To see all the expected values, select [APPS] and select Data/Matrix Editor. Select 2:Open, change the Type to Matrix, change the Folder to statvars, and change the Variable to expmat.



To see all the $(o-e)^2/e$ values, select [APPS] and select Data/Matrix Editor. Select 2:Open, change the Type to Matrix, change the Folder to statvars, and change the Variable to compmat.



If you need to delete a row or column, move the cursor to the row or column that you want to delete, then select F6 Util, then 2:Delete, then choose row or column, then enter. To add a row or column, just arrow over to the new row or column and type in the observed values. Or, use the option 6: Resize Matrix.



Example 9-5: The sample data below show the number of companies providing dental insurance for small, medium and large companies. Test to see if there is a relationship between dental insurance coverage and company size. Use $\alpha = 0.05$.

	Size of the Company		
Dental Insurance	Small	Medium	Large
Yes	21	25	19
No	46	39	10

Solution: State the hypotheses.

H_0 : Dental insurance coverage and company size are independent.

H_1 : Dental insurance coverage and company size are dependent.

Compute the expected values by taking each row total times column total, divided by grand total.

$O = \text{observed}$	Size of the Company			
Dental Insurance	Small	Medium	Large	Total
Yes	21	25	19	65
No	46	39	10	95
Total	67	64	29	160

For the small companies with dental insurance: $(65 \cdot 67) / 160 = 27.21875$,
 small companies without dental insurance: $(95 \cdot 67) / 160 = 39.78125$,
 medium companies with dental insurance: $(65 \cdot 64) / 160 = 26$, etc. See table below.

$E = \text{expected}$	Size of the Company			
Dental Insurance	Small	Medium	Large	Total
Yes	27.21875	26	11.78125	65
No	39.78125	38	17.21875	95
Total	67	64	29	160

Compute the test statistic.

$\frac{(O - E)^2}{E}$	Size of the Company		
Dental Insurance	Small	Medium	Large
Yes	$\frac{(21 - 27.21875)^2}{27.21875} = 1.42082$	$\frac{(25 - 26)^2}{26} = 0.03846$	$\frac{(19 - 11.78125)^2}{11.78125} = 4.42316$
No	$\frac{(46 - 39.78125)^2}{39.78125} = 0.97214$	$\frac{(39 - 38)^2}{38} = 0.02632$	$\frac{(10 - 17.21875)^2}{17.21875} = 3.02637$

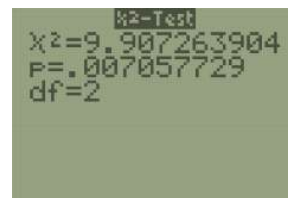
Test statistic is $\chi^2 = \sum \frac{(O-E)^2}{E} = 1.42082 + 0.03846 + 4.42316 + 0.97214 + 0.02632 + 3.02637 = 9.9073$.

Use technology to find the p-value using the chi-square cdf with
 $df = (R - 1)(C - 1) = (2 - 1)(3 - 1) = 2$.

Using the TI-Calculator, we find the p-value = 0.0071.

The p-value is less than α therefore reject H_0 .

There is enough evidence to support the claim that there is a relationship between dental insurance coverage and company size.



Chapter 9 Formulas

<p>Goodness of Fit Test $H_0: p_1 = p_0, p_2 = p_0, \dots, p_k = p_0$. H_1: At least one proportion is different. $\chi^2 = \sum \frac{(O-E)^2}{E}$ $df = k - 1, p_0 = 1/k$ or given % TI-84: χ^2 GOF-Test</p>	<p>Test for Independence H_0: <u>Variable 1</u> and <u>Variable 2</u> are independent. H_1: <u>Variable 1</u> and <u>Variable 2</u> are dependent. $\chi^2 = \sum \frac{(O-E)^2}{E}$ $df = (R - 1)(C - 1)$ TI-84: χ^2-Test</p>
--	--

One of the major selling point of that wholly remarkable travel book, the Hitchhiker's Guide to the Galaxy, apart from its relative cheapness and the fact that it has the words DON'T PANIC written in large friendly letters on its cover, is its compendious and occasionally accurate glossary. The statistics relating to the geo-social nature of the Universe, for instance, are deftly set out between pages nine hundred and thirty-eight thousand and twenty-four and nine hundred and thirty-eight thousand and twenty-six; and the simplistic style in which they are written is partly explained by the fact that the editors, having to meet a publishing deadline, copied the information off the back of a packet of breakfast cereal, hastily embroidering it with a few footnoted in order to avoid prosecution under the incomprehensibly tortuous Galactic Copyright laws.
(Adams, 2002)

Chapter 9 Exercises

- The shape of the χ^2 -distribution is usually:
 - Normal
 - Bell-shaped
 - Skewed left
 - Skewed right
 - Uniform
- Why are Goodness of Fit tests always right-tailed?
 - Because the test checks for a large variance between observed and expected values.
 - Because the χ^2 -distribution is skewed right.
 - Because χ^2 values can never be negative.
 - Because they test a variance and variance is always positive.
- What are the requirements to be satisfied before using a Goodness of Fit test? Check all that apply.
 - The data are obtained using systematic sampling.
 - The data are obtained from a simple random sample.
 - The expected frequency from each category is 5 or more.
 - The observed frequency from each category is organized from largest to smallest.
 - The degrees of freedom are less than 30.
- Calculate the critical value for a right-tailed test using a χ^2 -distribution with $\alpha = 0.01$ and $df = 5$.
- Calculate the critical value for a right-tailed test using a χ^2 -distribution with $\alpha = 0.05$ and $df = 17$.
- Calculate the critical value for a right-tailed test using a χ^2 -distribution with $\alpha = 0.05$ and $df = 6$.
- Calculate the critical value for a right-tailed test using a χ^2 -distribution with $\alpha = 0.10$ and $df = 7$.
- What is the mean of a Chi Square distribution with 7 degrees of freedom?
- What is the mean of a Chi Square distribution with 6 degrees of freedom?
- Skittles candy have 5 colors: green, orange, purple, red, and yellow. You buy a bag of skittles that has 56 pieces of candy. You are curious if all 5 colors are equally likely to appear in the bag or whether certain colors were more likely. If all 5 colors were equally likely to appear, what would be the expected value of skittles of each color?
- Suppose you are trying to determine whether a 20-sided dice is fair or if it tends to land on certain numbers more than others. You roll the dice 200 times, what is the expected value for each of the 20 sides?

For exercises 12-27, show all 5 steps for hypothesis testing:

- State the hypotheses.
- Compute the test statistic.
- Compute the critical value or p-value.
- State the decision.
- Write a summary.

12. Pamplona, Spain, is the home of the festival of San Fermin – The Running of the Bulls. The town is in festival mode for a week and a half every year at the beginning of July. There is a running joke in the city that Pamplona has a baby boom every April – 9 months after San Fermin. To test this claim, a resident takes a random sample of 200 birthdays from native residents and finds the following. At the 0.05 level of significance, can it be concluded that births in Pamplona are not equally distributed throughout the 12 months of the year?

January	17
February	19
March	16
April	24
May	15
June	16

July	13
August	16
September	18
October	16
November	14
December	16

13. A professor using an open-source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the term, she asks her students to complete a survey where they indicate what format of the book they used. Of the 126 students, 45 said they bought a hard copy of the book, 25 said they printed it out from the web, and 56 said they read it online. Run a Goodness of Fit test at $\alpha = 0.05$ to see if the distribution is different than expected.
14. The proportion of final grades for an anatomy class for the whole department are distributed as 10% A's, 23% B's, 45% C's, 14% D's, and 8% F's. A department chair is getting quite a few student complaints about a particular professor. The department chair wants to check to see if the professor's students' grades have a different distribution compared to the rest of the department. At the end of the term, the students have the following grades. Use $\alpha = 0.05$.

Grades	A	B	C	D	F
Observed	12	18	25	8	5

15. You might think that if you looked at the first digit in randomly selected numbers that the distribution would be uniform. Actually, it is not! Simon Newcomb and later Frank Benford both discovered that the digits occur according to the following distribution.

Digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

A forensic accountant can use Benford's Law to detect fraudulent tax data. Suppose you work for the IRS and are investigating an individual suspected of embezzling. The first digit of 192 checks to a supposed company are as follows.

Digit	1	2	3	4	5	6	7	8	9
Observed Frequency	56	23	19	20	16	19	17	10	12

Run a complete Goodness of Fit test to see if the individual is likely to have committed tax fraud. Use $\alpha = 0.05$. Should law enforcement officials pursue the case? Explain.

16. A college professor is curious if location of seat in class affects grade in the class. She is teaching in a lecture hall to 200 students. The lecture hall has 10 rows, so she splits it into 5 categories – Rows 1-2, Rows 3-4, Rows 5-6, Rows 7-8, and Rows 9-10. At the end of the course, she determines the top 25% of grades in the class, and if location of seat makes no difference, she would expect that these top 25% of students would be equally dispersed throughout the classroom. Her observations are recorded below. Run a Goodness of Fit test to determine whether location has an impact on grade. Let $\alpha = 0.05$.

# In Top 25%	Rows 1-2	Rows 3-4	Rows 5-6	Rows 7-8	Rows 9-10
Observed	12	7	10	8	13

17. Consumer panel preferences for four store displays follow. Test to see whether there is a preference among the four display designs. Use $\alpha = 0.05$.

Display	A	B	C	D
Observed	43	60	47	50

18. The manager of a coffee shop wants to know if his customers' drink preferences have changed in the past year. He knows that last year the preferences followed the following proportions – 34% Americano, 21% Cappuccino, 14% Espresso, 11% Latte, 10% Macchiato, 10% Other. In a random sample of 300 customers, he finds that 90 ordered Americanos, 65 ordered Cappuccinos, 52 ordered Espressos, 35 ordered Lattes, 34 ordered Macchiatos, and the rest ordered something in the Other category. Run a Goodness of Fit test to determine whether drink preferences have changed at his coffee shop. Use a 0.05 level of significance.
19. The director of a Driver's Ed program is curious if the time of year has an impact on number of car accidents in the United States. They assume that weather may have a significant impact on the ability of drivers to control their vehicles. They take a random sample of 100 car accidents and record the season each occurred in. They found that 20 occurred in the spring, 31 in the summer, 23 in the fall, and 26 in the winter. Can it be concluded at the 0.05 level of significance that car accidents are not equally distributed throughout the year?
20. A college prep school advertises that their students are more prepared to succeed in college than other schools. To show this, they categorize GPAs into 4 groups and look up the proportion of students at a state college in each category. They find that 7% have a 0-0.99, 21% have a 1-1.99, 37% have a 2-2.99, and 35% have a 3-4.00 in GPA. They then take a random sample of 150 of their graduates at the state college and find that 5 has a 0-0.99, 18 have a 1-1.99, 67 have a 2-2.99, and 60 have a 3-4.00. Can they conclude that the grades of their graduates are distributed differently than the general population at the school? Test at the 0.05 level of significance.
21. The permanent residence of adults aged 18-25 in the United States was examined in a survey from the year 2000. The survey revealed that 27% of these adults lived alone, 32% lived with a roommate(s), and 41% lived with their parents/guardians. In 2008, during an economic recession in the country, another such survey of 1,500 people revealed that 378 lived alone, 452 lived with a roommate(s), and 670 lived with their parents. Is there a significant difference in where young adults lived in 2000 versus 2008? Test with a Goodness of Fit test at $\alpha = 0.05$.
22. A color code personality test categorizes people into four colors – Red (Power), Blue (Intimacy), Green (Peace), and Yellow (Fun). In general, 25% of people are Red, 35% Blue, 20% Green, and 20% Yellow. An art class of 33 students is tested at a university and 4 are found to be Red, 14 Blue, 7 Green, and 8 Yellow. Can it be concluded that personality type has an impact on students' areas of interest and talents, such as artistic students? Test at a 0.05 level of significance.
23. An urban economist is curious if the distribution in where Oregon residents live is different today than it was in 1990. She observes that today there are approximately 3,050 thousand residents in NW Oregon, 907 thousand residents in SW Oregon, 257 thousand in Central Oregon, and 106 thousand in Eastern Oregon. She knows that in 1990 the breakdown was as follows: 72.7% NW Oregon, 19.7% SW Oregon, 4.8% Central Oregon, and 2.8% Eastern Oregon. Can she conclude that the distribution in residence is different today at a 0.05 level of significance?
24. A large department store is curious what sections of the store make the most sales. The manager has data from 10 years prior that show 30% of sales come from Clothing, 25% Home Appliances, 18% Houseware, 13% Cosmetics, 12% Jewelry, and 2% Other. In a random sample of 500 current sales, 176 came from Clothing, 150 Home Appliances, 75 Houseware, 42 Cosmetics, 51 Jewelry, and 6 Other. At $\alpha = 0.10$, can the manager conclude that the distribution of sales among the departments has changed?
25. Students at a high school are asked to evaluate their experience in a class at the end of each school year. The courses are evaluated on a 1-4 scale – with 4 being the best experience possible. In the History Department, the

courses typically are evaluated at 10% 1's, 15% 2's, 34% 3's, and 41% 4's. A new history teacher, Mr. Mendoza, sets a goal to outscore these numbers. At the end of the year, he takes a random sample of his evaluations and finds 11 1's, 14 2's, 47 3's, and 53 4's. At the 0.05 level of significance, can Mr. Mendoza claim that his evaluations are significantly different from the History Department's?

26. A company manager believes that a person's ability to be a leader is directly related to their zodiac sign. He never selects someone to chair a committee without first evaluating their zodiac sign. An irate employee sets out to show her manager is wrong. She claims that if zodiac sign truly makes a difference in leadership, then a random sample of 200 CEOs in our country would reveal a difference in zodiac sign distribution. She finds the following zodiac signs for her random sample of 200 CEOs. Can the employee conclude that there is a difference in the proportion of CEOs for the twelve zodiac signs? Use $\alpha = 0.05$.

Births	Signs
23	Aries
12	Taurus
16	Gemini
20	Cancer
14	Leo
16	Virgo

Births	Signs
15	Libra
14	Scorpio
20	Sagittarius
11	Capricorn
17	Aquarius
22	Pisces

27. A company that develops over-the-counter medicines is working on a new product that is meant to shorten the length of sore throats. To test their product for effectiveness, they take a random sample of 100 people and record how long it took for their symptoms to completely disappear. The results are in the table below. The company knows that on average (without medication) it takes a sore throat 6 days or less to heal 42% of the time, 7-9 days 31% of the time, 10-12 days 16% of the time, and 13 days or more 11% of the time. Can it be concluded at the 0.01 level of significance that the patients who took the medicine healed at a different rate than these percentages?

Duration of Sore Throat	6 days or less	7-9 days	10-12 days	13 or more days
Observed	47	38	10	5

28. What are the requirements to be satisfied before using the χ^2 Independence Test? Check all that apply.
- The sample sizes must be greater than 30.
 - The data are obtained from a random sample.
 - The data are obtained using stratified sampling.
 - The expected frequency from each category is 5 or more.
 - The observed frequency from each category is organized from largest to smallest.
 - The population is normally distributed.
29. The null hypothesis for the χ^2 Independence Test always states that _____.
- the two values are equal.
 - one variable is dependent on another variable.
 - one variable is independent of another variable.
 - the expected values and observed values are the same.
30. A contingency table has 4 rows and 5 columns, what would be the degrees of freedom used in the χ^2 Independence Test?

31. What are the degrees of freedom used in the χ^2 Independence Test?

- a) $n - 1$
- b) Rows + Columns
- c) n
- d) $(\text{Rows} - 1) * (\text{Columns} - 1)$
- e) $n - 2$

For exercises 32-42, show all 5 steps for hypothesis testing:

- a) State the hypotheses.
- b) Compute the test statistic.
- c) Compute the critical value or p-value.
- d) State the decision.
- e) Write a summary.

32. The sample data below show the number of companies providing health insurance for small, medium and large companies. Test to see whether health insurance coverage and company size are dependent. Use $\alpha = 0.01$.

	Size of the Company		
Health Insurance	Small	Medium	Large
Yes	40	70	90
No	20	20	10

33. A restaurant chain that has 3 locations in Portland is trying to determine which of their 3 locations they should keep open on New Year's Eve. They survey a random sample of customers at each location and ask each whether they plan to go out to eat on New Year's Eve. The results are below. Run a test for independence to decide if the proportion of customers who will go out to eat on New Year's Eve is dependent on location. Use $\alpha = 0.05$.

	NW Location	NE Location	SE Location
Will Go Out	45	33	36
Won't Go Out	23	29	25

34. The following sample was collected during registration at a large middle school. At the 0.05 level of significance, can it be concluded that level of math is dependent on grade level?

	Honors Math	Regular Math	General Math
6 th Grade	34	45	15
7 th Grade	37	49	13
8 th Grade	29	45	17

35. A high school offers math placement exams for incoming freshmen to place students into the appropriate math class during their freshman year. Three middle schools were sampled and the following pass/fail results were found. Test to see if the math placement exam and where students are placed are dependent at the 0.10 level of significance.

	School A	School B	School C
Pass	42	29	45
Fail	57	35	61

36. A public opinion poll surveyed a simple random sample of 500 voters in Oregon. The respondents were asked which political party they identified with most and were categorized by residence. Results are shown below. Decide if voting preference is dependent of location of residence. Let $\alpha = 0.05$.

	Republican	Democrat	Independent
NW Oregon	72	104	24
SW Oregon	31	59	10
Central Oregon	42	47	11
Eastern Oregon	59	33	8

37. A university changed to a new learning management system (LMS) during the past school year. The school wants to find out how it is working for the different departments – the results in preference found from a survey are below. Test to see if the department and LMS preference are dependent at $\alpha = 0.05$.

	Prefers Old LMS	Prefers New LMS	No Preference
School of Business	15	24	6
College of Liberal Arts & Science	34	7	19
College of Education	21	19	5

38. The medal count for the 2018 winter Olympics is recorded below. Run an independence test to find out if the medal won is dependent on country. Use $\alpha = 0.10$.

	Gold	Silver	Bronze
Norway	14	14	11
Germany	14	10	7
Canada	11	8	10
United States	9	8	6

39. An electronics store has 4 branches in a large city. They are curious if sales in any particular department are different depending on location. They take a random sample of purchases throughout the 4 branches – the results are recorded below. Test to see if the type of electronic device and store branch are dependent at the 0.05 level of significance.

	Appliances	TV	Computers	Cameras	Cellphones
Branch 1	54	28	61	24	81
Branch 2	44	21	55	23	92
Branch 3	49	18	49	30	72
Branch 4	51	29	65	29	102

40. A high school runs a survey asking students if they participate in sports. The results are found below. Test to see if there is a relationship between participating in sports and year in school at $\alpha = 0.05$.

	Freshmen	Sophomores	Juniors	Seniors
Yes	82	89	54	42
No	29	25	38	37

41. A manufacturing company knows that their machines produce parts that are defective on occasion. They have 4 machines producing parts and want to test if defective parts are dependent on the machine that produced it. They take a random sample of 300 parts and find the following results. Test at the 0.05 level of significance.

	Machine 1	Machine 2	Machine 3	Machine 4
Defective	9	12	15	6
Non-Defective	63	70	68	57

42. A 4-year college is curious which of their students hold down a job while also attending school. They poll the students and find the results below. Test to see if there is a relationship between college students having a job and year in school. Use $\alpha = 0.05$.

	Freshmen	Sophomores	Juniors	Seniors
Works	61	45	31	72
Doesn't Work	45	33	38	19

Chapter 10

Analysis of Variance



- 10.1 Introduction to the F-Distribution
- 10.2 One-Way ANOVA
- 10.3 Pairwise Comparison of Means
(Post Hoc Tests)

10.1 Introduction to the F-Distribution

Simple one-way Analysis of Variance (ANOVA) is a statistical test used to assess if there are significant differences between the means of 3 or more independent groups. In this test, data is collected from each group, typically in the form of continuous numerical measurements. The goal is to analyze the variation within each group as well as the variation between the group means.

To conduct a one-way ANOVA test, we calculate the F-statistic, which represents the ratio of between-group variation to within-group variation. This statistic is then compared to a critical value derived from the F-distribution to determine the statistical significance of the results. If the calculated F-statistic exceeds the critical value, or the p-value $\leq \alpha$, the test suggests that there is a significant difference among the group means, indicating that at least one mean differs from the others.

Following a statistically significant ANOVA result, post-hoc tests such as Tukey's test or the Bonferroni test can be employed to identify the specific group means that significantly differ from each other. These tests provide further insights into pairwise comparisons and help pinpoint which groups contribute to the observed differences.

An **F-distribution** is another special type of distribution for a continuous random variable.

Properties of the F-distribution:

- Right skewed.
- F-scores cannot be negative.
- The spread of an F-distribution is determined by the degrees of freedom of the numerator, and by the degrees of freedom of the denominator. The df are usually determined by the sample sizes of the two populations or number of groups.
- The total area under the curve is equal to 1 or 100%.

The shape of the distribution curve changes when the degrees of freedom change. Figure 10-1 shows examples of F-distributions with different degrees of freedom.

We will use the F-distribution in several types of hypothesis testing. For now, we are just learning how to find the critical value and probability using the F-distribution.

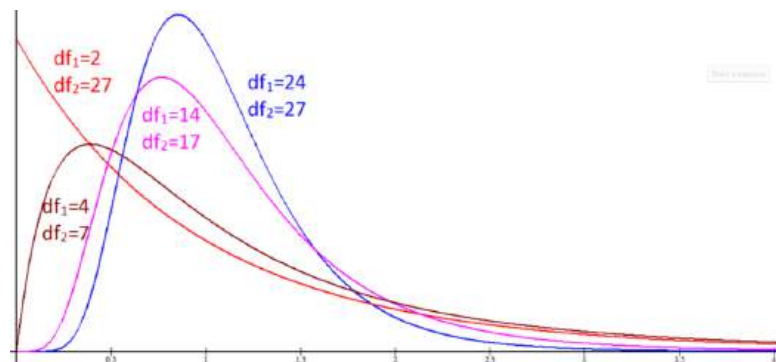


Figure 10-1

Use technology to calculate F values. The TI-89 calculator's function inverse F distribution menu or in Excel use the F.INV function will calculate the F-score critical value for the F-distribution. When finding a probability given an F-score, use the calculator Fcdf function under the DISTR menu or in Excel use F.DIST. Note that the TI-83 and TI-84 do not come with the INVF function, but you may be able to find the program online or from your instructor.

Alternatively, use the calculator at <https://homepage.divms.uiowa.edu/~mbognar/applets/f.html> which will also graph the distribution for you and shade in one tail at a time. You will see the shape of the F-distribution change in the following examples depending on the degrees of freedom used. For your own sketch just make sure you have a positively skewed distribution starting at zero.

The critical values $F_{\alpha/2}$ and $F_{1-\alpha/2}$ are for a two-tailed test on the F-distribution curve with area $1 - \alpha/2$ between the critical values as shown in Figure 10-2. Note that the distribution starts at zero, is positively skewed, and never has negative F-scores.

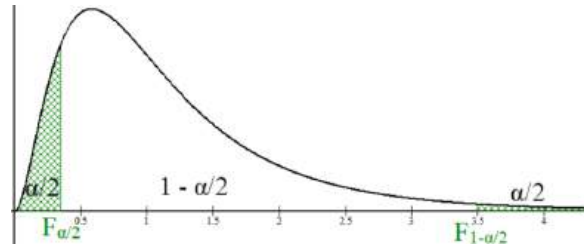


Figure 10-2

Example 10-1: Compute the critical values $F_{\alpha/2}$ and $F_{1-\alpha/2}$ with $df_1 = 6$ and $df_2 = 14$ for a two-tailed test, $\alpha = 0.05$.

Solution: Start by drawing the curve and finding the area in each tail. For this case, it would be an area of $\alpha/2$ in each tail. Then use technology to find the F-scores. Most technology only asks for the area to the left of the F-score you are trying to find. In Excel the function for $F_{\alpha/2}$ is $F.INV(\text{area in left-tail}, df_1, df_2)$.

There is only one function, so use areas 0.025 and 0.975 in the left tail. For this example, we would have critical values $F_{0.025} = F.INV(0.025, 6, 14) = 0.1888$ and $F_{0.975} = F.INV(0.975, 6, 14) = 3.5014$. See Figure 10-3.

We have to calculate two distinct F-scores unlike symmetric distribution where we could just do $\pm z$ -score or $\pm t$ -score.

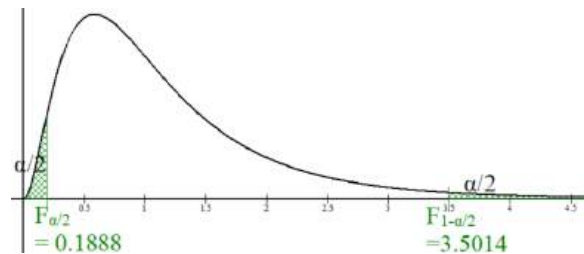


Figure 10-3

Note if you were doing a one-tailed test then do not divide alpha by two and use area = α for a left-tailed test and area = $1 - \alpha$ for a right-tailed test.

Example 10-2: Find the critical value for a right-tailed test with denominator degrees of freedom of 12 and numerator degrees of freedom of 2 with a 5% level of significance.

Solution: Draw the curve and shade in the top 5% of the upper tail since $\alpha = 0.05$, see Figure 10-4. When using technology, you will need the area to the left of the critical value that you are trying to find. This would be $1 - \alpha = 0.95$. Then identify the degrees of freedom. The first degrees of freedom are the numerator df , therefore $df_1 = 2$. The second degrees of freedom are the denominator df , therefore $df_2 = 12$. Using Excel, we would have $F_{1-\alpha} = F.INV(0.95, 2, 12) = 3.8853$.

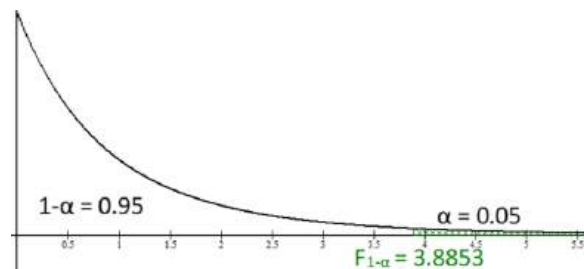


Figure 10-4

Example 10-3: Compute $P(F > 3.894)$, with $df_1 = 3$ and $df_2 = 18$.

Solution: In Excel, use the function $F.DIST(x, \text{deg_freedom1}, \text{deg_freedom2}, \text{cumulative})$. Always use TRUE for the cumulative. The $F.DIST$ function will find the probability (area) below F . Since we want the area above F , we would need to also use the complement rule. The formula would be $=1 - F.DIST(3.894, 3, 18, TRUE) = 0.0263$.

TI-84: The TI-84 calculator has a built in F-distribution. Press $[2^{nd}]$ $[DISTR]$ (this is F5: DISTR in the STAT app in the TI-89), then arrow down until you get to the Fcdf and press $[Enter]$. Depending on your calculator, you may not get a prompt for the boundaries and df . If you just see Fcdf(then you will need to enter each the lower boundary,

upper boundary, df_1 , and df_2 with a comma between each argument. The lower boundary is the 3.394 and the upper boundary is infinity (TI-83 and 84 use a really large number instead of ∞), then enter the two degrees of freedom. Press [Paste] and then [Enter], this will put the $Fcdf(3.894, 1E99, 3, 18)$ on your screen and then press [Enter] again to calculate the value.



Using the online calculator at <https://homepage.divms.uiowa.edu/~mbognar/applets/f.html> we get the same answer, see Figure 10-5.

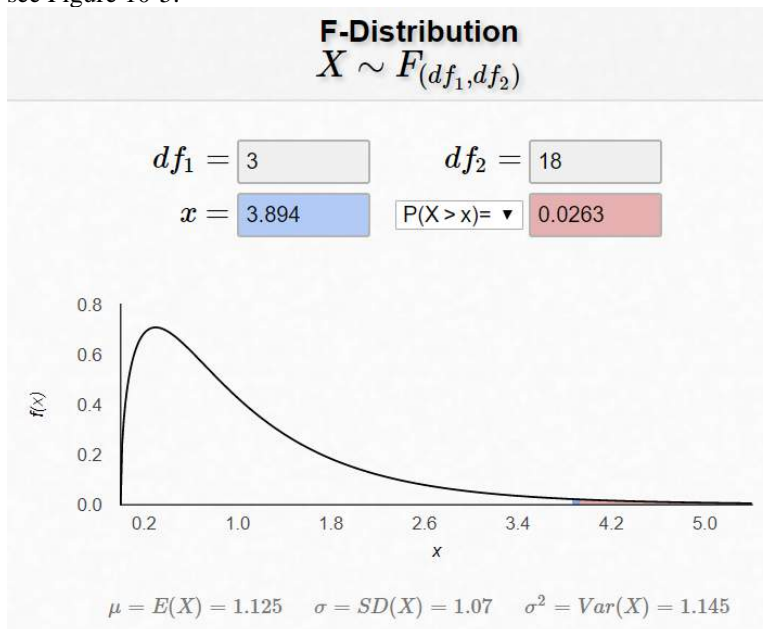


Figure 10-5

10.2 One-Way ANOVA

The z - and t -tests can be used to test the equality between two population means μ_1 and μ_2 . When we have more than two groups, we would inflate the probability of making a type I error if we were to compare just two at a time and make a conclusion about all the groups together. To account for this P(Type I Error) inflation, we instead will do an analysis of variance (ANOVA) to test the equality between 3 or more population means $\mu_1, \mu_2, \mu_3, \dots, \mu_k$.

The F-test (for ANOVA) is a statistical test for testing the equality of k population means.

The one-way ANOVA F-test is a statistical test for testing the equality of k population means from 3 or more groups within one variable or factor. There are many different types of ANOVA; for now, we are going to start with what is commonly referred to as a one-way ANOVA, which has one main effect or factor that is split up into three or more independent treatment levels. In more advanced courses you would learn about dependent groups or two or more factors.

Assumptions:

- The populations are normally distributed continuous variables with equal variances.
- The observations are independent.

The hypotheses for testing the equality of k population means (ANOVA) are set up with all the means equal to one another in the null hypothesis and at least one mean is different in the alternative hypothesis.

$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$
 $H_1: \text{At least one mean is different.}$

Even though there is equality in H_0 , the ANOVA test is testing if the variance **between** groups is significantly greater than the variance **within** groups, hence this will always be set up as a right-tailed test.

We will be using abbreviations for many of the numbers found in this section.

B = Between, W = Within

MS = Mean Square (This is a variance)

MSB = Mean Square (Variance) Between groups

MSW = Mean Square (Variance) Within groups.

The formula for the F-test statistic is $F = \frac{MSB}{MSW}$.

Use the F-distribution with degrees of freedom from the between and within groups. The numerator degrees of freedom are equal to the number of groups minus one, that is numerator degrees of freedom are $df_B = k - 1$. The denominator degrees of freedom are equal to the total of all the sample sizes minus the number of groups, that is denominator degrees of freedom are $df_W = N - k$.

The sum of squares, degrees of freedom and mean squares are organized in a table called an ANOVA table.

Figure 10-6 below is a template for an ANOVA table.

Source	SS = Sum of Squares	df	MS = Mean Square	F
Between (Factor)	SSB	$k - 1$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within (Error)	SSW	$N - k$	$MSW = \frac{SSW}{N-k}$	
Total	SST	$N - 1$		

Figure 10-6

Where:

\bar{x}_i = sample mean from the i^{th} group

s_i^2 = sample variance from the i^{th} group

n_i = sample size of the i^{th} group

k = number of groups

$N = n_1 + n_2 + \dots + n_k$ = sum of the individual sample sizes for groups

Grand mean from all groups = $\bar{x}_{GM} = \frac{\sum x_i}{N}$

Sum of squares between groups = $SSB = \sum n_i (\bar{x}_i - \bar{x}_{GM})^2$

Sum of squares within groups = $SSW = \sum (n_i - 1) s_i^2$

Mean squares between groups (or the between-groups variance s_B^2) = $MSB = \frac{SSB}{k-1}$

Mean squares within-group (or error within-groups variance s_W^2) = $MSW = \frac{SSW}{N-k}$

$F = \frac{MSB}{MSW}$ is the test statistic.

These calculations can be time-consuming to do by hand so use technology to find the ANOVA table values, critical value and/or p-value.

Different textbooks and computer software programs use different labels in the ANOVA tables.

- The TI-calculators use the word Factor for Between Groups and Error for Within Groups.
- Some software packages use Treatment instead of Between groups. You may see different notation depending on which textbook, software or video you are using.
- For between groups $SSB = SS_B = SSTR = SST = SSF$ and for within groups $SSW = SS_W = SSE$.
- One thing that is consistent within the ANOVA table is that the between = factor = treatment always appears on the first row of the ANOVA table, and the within = error always is in the second row of the ANOVA table.
- The df column usually is in the second column since we divide the sum of squares by the df to find the mean squares. However, some software packages will put the df column before the sum of squares column.
- The F test statistic is under the column labeled F.
- Many software packages will give an extra column for the p-value (sometimes labeled sig.) and some software packages give a critical value too.

Assumption: The population we are sampling from must be approximately normal with equal variances. If these assumptions are not met there are more advanced statistical methods that should be used.

Example 10-4: An educator wants to see if there is a difference in the average grades given to students for the 4 different instructors who teach intro statistics courses. They randomly choose courses that each of the 4 instructors taught over the last few years and perform an ANOVA test. What would be the correct hypotheses for this test?

Solution: There are 4 groups (the 4 instructors) so there will be 4 means in the null hypothesis,
 $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$.

It is tempting to write $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ for the alternative hypothesis, but the test is testing the opposite of all equal is that at least one mean is different. We could for instance have just the mean for group 2 be different and groups 1, 3 and 4 have equal means. If we wanted to write out all the way to have unequal groups, you would have a combination problem with ${}^4C_3 + {}^4C_2 + {}^4C_1$ ways of getting unequal means.

Instead of all of these possibilities, we just write a sentence “at least one mean is different.”

They hypotheses are: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
 $H_1: \text{At least one mean is different.}$

Example 10-5: A researcher claims that there is a difference in the average age of assistant professors, associate professors, and full professors at her university. Faculty members are selected randomly and their ages are recorded. Assume faculty ages are normally distributed. Test the claim at the $\alpha = 0.01$ significance level. The data are listed below.

Assistant Prof	28	32	36	42	50	33	38
Associate Prof	44	61	52	54	62	45	46
Prof	54	56	55	65	52	50	46

Solution: The claim is that there is a difference in the average age of assistant professors (μ_1), associate professors (μ_2), and full professors (μ_3) at her university.

The correct hypotheses are: $H_0: \mu_1 = \mu_2 = \mu_3$
 $H_1: \text{At least one mean differs.}$

We need to compute all of the necessary parts for the ANOVA table and F-test.

Compute the descriptive stats for each group with your calculator using 1-Var Stats L_1 .

1-Var Stats	1-Var Stats	1-Var Stats
$\bar{x}=37$	$\bar{x}=52$	$\bar{x}=54$
$\Sigma x=259$	$\Sigma x=364$	$\Sigma x=378$
$\Sigma x^2=9901$	$\Sigma x^2=19262$	$\Sigma x^2=20622$
$S_x=7.280109889$	$S_x=7.461009762$	$S_x=5.916079783$
$\sigma_x=6.740072064$	$\sigma_x=6.907552802$	$\sigma_x=5.477225575$
$n=7$	$n=7$	$n=7$

Record the sample size, sample mean, sum of x, and sample variances. Take the standard deviation s_x and square it to find the variance s_x^2 for each group.

Assistant Prof	$n_1 = 7$	$\bar{x}_1 = 37$	$\Sigma x_1 = 259$	$s_1^2 = 53$
Associate Prof	$n_2 = 7$	$\bar{x}_2 = 52$	$\Sigma x_2 = 364$	$s_2^2 = 55.66667$
Prof	$n_3 = 7$	$\bar{x}_3 = 54$	$\Sigma x_3 = 378$	$s_3^2 = 35$

Compute the grand mean: $N = n_1 + n_2 + n_3 = 7 + 7 + 7 = 21$, $\bar{x}_{GM} = \frac{\Sigma x_i}{N} = \frac{(259+364+378)}{21} = 47.66667$.

Compute the sum of squares for between groups: $SSB = \Sigma n_i(\bar{x}_i - \bar{x}_{GM})^2 = n_1(\bar{x}_1 - \bar{x}_{GM})^2 + n_2(\bar{x}_2 - \bar{x}_{GM})^2 + n_3(\bar{x}_3 - \bar{x}_{GM})^2 = 7(37 - 47.66667)^2 + 7(52 - 47.66667)^2 + 7(54 - 47.66667)^2 = 1208.66667$.

Compute the sum of squares within groups:

$SSW = \Sigma(n_i - 1)s_i^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 = 6 \cdot 53 + 6 \cdot 55.66667 + 6 \cdot 35 = 862$.

Place the sum of squares into your ANOVA table and add them up to get the total.

Source	SS	df	MS	F
Between	1208.66667			
Within	862			
Total	2070.66667			

Next, find the degrees of freedom: $k = 3$ since there are 3 groups, so $df_B = k - 1 = 2$; $df_W = N - k = 21 - 3 = 18$. Add the degrees of freedom to the table to get the total df .

Source	SS	df	MS	F
Between	1208.66667	2		
Within	862	18		
Total	2070.66667	20		

Compute the mean squares by dividing the sum of squares by their corresponding df then add these numbers to the table:

$$MSB = \frac{SSB}{k-1} = \frac{1208.66667}{2} = 604.3333 \quad MSW = \frac{SSW}{N-k} = \frac{862}{18} = 47.8889.$$

The test statistic is the ratio of these two mean squares: $F = \frac{MSB}{MSW} = \frac{604.3333}{47.8889} = 12.6195$. Add the test statistic to the table under F.

Source	SS	df	MS	F
Between	1208.66667	2	604.3333	12.6195
Within	862	18	47.8889	
Total	2070.66667	20		

All ANOVA tests are right-tailed tests, so the critical value for a right-tailed F-test is found the F-distribution.

Use $\alpha = 0.01$ area in the right-tail. The degrees of freedom are $dfN = 2$, and $dfD = 18$. In Excel the critical value formula would be $F_{1-\alpha} = F.INV.RT(0.01,2,18) = 6.0129$. Since the critical value is 6.0129, see the sampling distribution curve in Figure 10-7.

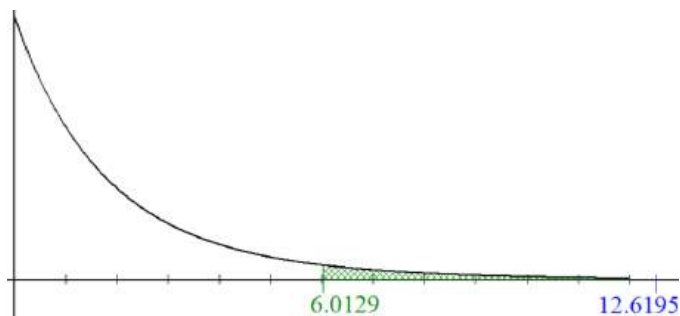


Figure 10-7

Note that the F-distribution starts at zero, and is skewed to the right. The test statistic of 12.6295 is larger than the critical value of 6.0129 so the decision would be to reject the null hypothesis.

Decision: Reject H_0 .

Summary: At the 1% significance level, there is enough evidence to support the claim that there is a difference in the average age of assistant professors, associate professors, and full professors at her university.

If we were using the p-value method, then we would use the calculator or computer for a right-tailed test. The p-value is the probability of observing the F-statistic, or more extreme, given the null hypothesis is true

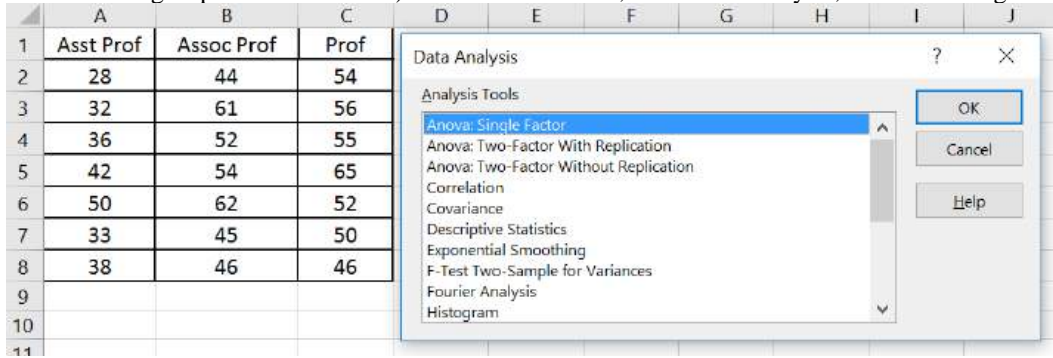
The p-value = 0.0003755 is less than $\alpha = 0.01$ which leads to the same decision of reject H_0 that we found using the critical value method.

Alternatively, use technology to compute the ANOVA table and p-value.

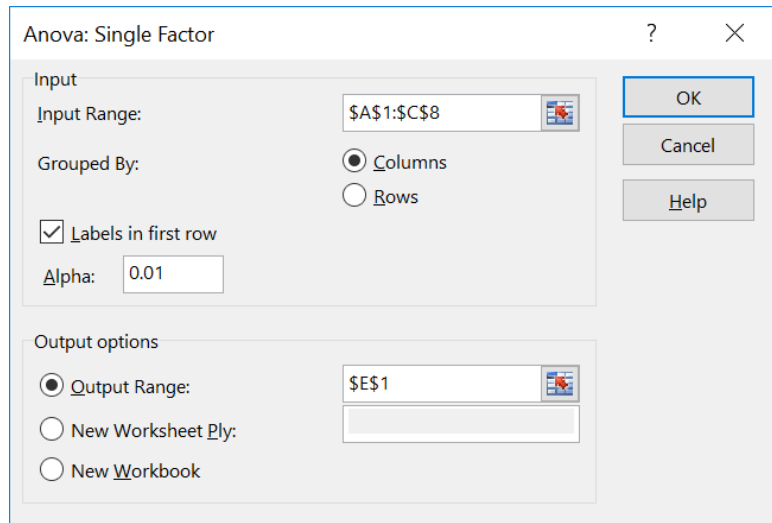
TI-84: ANOVA, hypothesis test for the equality of k population means. Note you have to have the actual raw data to do this test on the calculator. Press the [STAT] key and then the [EDIT] function, type the three lists of data into list one, two and three. Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [F:ANOVA()] and press the [ENTER] key. This brings you back to the regular screen where you should now see ANOVA(. Now press the [2nd] [L₁] [,] [2nd] [L₂] [,] [2nd] [L₃][)] keys in that order. You should now see ANOVA(L₁,L₂,L₃), if you had 4 lists you would then have an additional list. Press the [ENTER] key. The calculator returns the F-test statistic, the p-value, Factor (Between) df , SS and MS, Error (Within) df , SS and MS. The last value Sxp is the square root of the MSE.

TI-89: ANOVA, hypothesis test for the equality of k population means. Go to the [Apps] **Stat/List Editor**, then type in the data for each group into a separate list (or if you don't have the raw data, enter the sample size, sample mean and sample variance for group 1 into list1 in that order, repeat for list2, etc.). Press [2nd] then F6 [Tests], then select **C:ANOVA**. Select the input method **data** or **stats**. Select the number of groups. Press the [ENTER] key to calculate. The calculator returns the F-test statistic, the p-value, Factor (Between) df , SS and MS, Error (Within) df , SS and MS. The last value Sxp is the square root of the MSE.

Excel: Type the labels and data into adjacent columns (it is important not to have any blank columns or this will be an additional group counted as zeros). Select the Data tab, then Data Analysis, ANOVA: Single-Factor, then OK.



Next, select all three columns of data at once for the input range. Check the box that says Labels in first row (only select this if you actually selected the labels in the input range). Change your value of alpha and output range will be one cell reference where you want your output to start, see below.



You get the following output:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Asst Prof	7	259	37	53		
Assoc Prof	7	364	52	55.6667		
Prof	7	378	54	35		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1208.6667	2	604.3333	12.6195	0.0004	6.0129
Within Groups	862	18	47.8889			
Total	2070.6667	20				

Excel gives both the p-value and critical value so you can use either method when making your decision, but make sure you are comfortable with both.

Example 10-6: A manager wants to see if there is a difference in the average production rate for 4 manufacturing plants. The partially filled out ANOVA table is provided, but not the actual data. Fill out the missing pieces of the ANOVA table, and test the manager's claim using $\alpha = 5\%$.

Source	SS	df	MS	F
Between				
Within			6.2776	
Total	160.92	20		

Solution: The claim is that there is a difference in the average production rate for 4 manufacturing plants.

The correct hypotheses are: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
 $H_1: \text{At least one mean differs.}$

To find the test statistic, we need to fill out the ANOVA table. Start with the df . Since there are 4 groups, $k = 4$, so the $df_B = k - 1 = 3$. Since the total df equals 20, then $df_W = 20 - 3 = 17$. This means that the total $N = 21$.

Source	SS	df	MS	F
Between		3		
Within		17	6.2776	
Total	160.92	20		

Next, find the missing sum of squares. Recall that $MSW = \frac{SSW}{df_W}$, this means that the $SSW = df_W * MSW = 17 * 6.2776 = 106.7192$. The total of the sum of squares must equal 160.92, so subtract the $SSW = 106.7192$ from the total to get $160.92 - 106.7192 = 54.2008$.

Source	SS	df	MS	F
Between	$160.92 - 106.7192 = 54.2008$	3		
Within	$17 * 6.2776 = 106.7192$	17	6.2776	
Total	160.92	20		

Next, calculate the $MSB = SSB/df_B = 54.2008/3 = 18.066933$. Then calculate the test statistic $F = MSB/MSW = 18.066933/6.2776 = 2.878$.

Source	SS	df	MS	F
Between	54.2008	3	$54.2008/3 = 18.066933$	$18.066933/6.2776 = 2.878$
Within	106.7192	17	6.2776	
Total	160.92	20		

Compute the critical value using the inverse F program with right tail area = 0.05, $df_B=3$, $df_W=17$, to get the critical value $F_{1-\alpha} = 3.1968$.

Draw and label the chi-square distribution curve. Label zero and the critical region. See Figure 10-8.

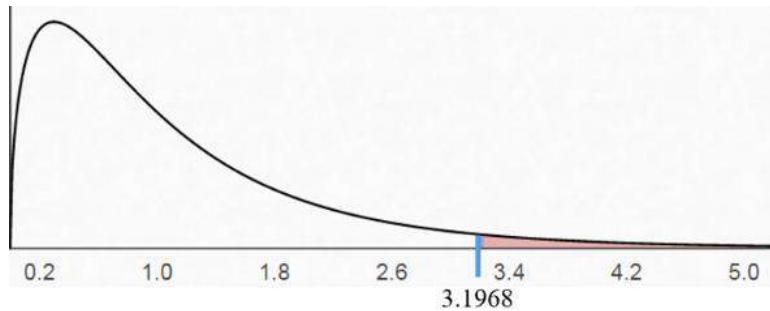


Figure 10-8

Decision: Fail to reject H_0 , since the test statistic $F = 2.878$ is not in the critical region.

Summary: At the 5% significance level, there is not enough evidence to support the claim that there is a difference in the average production rate for 4 manufacturing plants.

The ANOVA test gives evidence that there is a difference between three or more means. The null hypothesis will always have the means equal to one another versus the alternative hypothesis that at least one mean is different. The F-test results are about the difference in means, but the test is actually testing if the variation between the groups is larger than the variation within the groups. If this between group variation is significantly larger than the within groups then we can say there is a statistically significant difference in the population means. Hence, we are always performing a right-tailed F-test for ANOVA. Make sure to only compare the p-value with α and the test statistic to the critical value.

10.3 Pairwise Comparisons of Means (Post-Hoc Tests)

If you reject H_0 , then you know that at least two of the means are different. The ANOVA test does not tell which of those means are different, only that a difference exists. Most likely your sample means will be different from each other, but how different do they need to be for there to be a statistically significant difference in the population means?

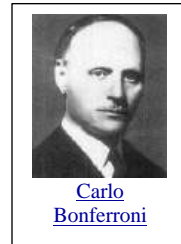
To determine which means are significantly different, you need to conduct further tests. These post-hoc tests include the range test, multiple comparison tests, Duncan test, Student-Newman-Keuls test, Tukey test, Scheffé test, Dunnett test, and the Bonferroni test to name a few. There are more and there is no consensus on which test to use. These tests are available in statistical software packages such as R, Minitab and SPSS.

One should never use two-sample t-tests from the previous chapter. This would inflate the type I error.

The probability of at least one type I error increases exponentially with the number of groups you are comparing. Let us assume that $\alpha = 0.05$, then the probability that an observed difference between two groups that does not occur by chance is $1 - \alpha = 0.95$. If two comparisons are made, the probability that the observed difference is true is no longer 0.95. The probability is $(1 - \alpha)^2 = 0.9025$, and the $P(\text{Type I Error}) = 1 - 0.9025 = 0.0975$. Therefore, the $P(\text{Type I Error})$ occurs if m comparisons are made is $1 - (1 - \alpha)^m$.

For instance, if we are comparing the means of four groups: There would be $m = {}_4C_2 = 6$ different ways to compare the 4 groups, groups (1,2), (1,3), (1,4), (2,3), (2,4), and (3,4). The $P(\text{Type I Error}) = 1 - (1 - \alpha)^m = 1 - (1 - 0.05)^6 = 0.2649$. This is why a researcher should use ANOVA for comparing means instead of independent t-tests.

There are many different methods to use. Many require special tables or software. We could actually just start with post-hoc tests, but they are a lot of work. If we run an ANOVA and we fail to reject the null hypothesis, then there is no need for further testing and it will save time if you were doing these steps by hand. Most statistical software packages give you the ANOVA table followed by the pairwise comparisons with just a change in the options menu. Keep in mind that Excel is not a statistical software and does not give pairwise comparisons.



We will use are the Bonferroni Test named after the mathematician Carlo Bonferroni. The Bonferroni Test uses the t-distribution table and is similar to previous t-tests that we used earlier but adjusts α to the number of comparisons being made.

The Bonferroni test is a statistical test for testing the difference between two population means (only done after an ANOVA test shows not all means are equal).

$$\text{The formula for the Bonferroni test statistic is } t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\left(MSW \left(\frac{1}{n_i} + \frac{1}{n_j} \right) \right)}}$$

where \bar{x}_i and \bar{x}_j are the means of the samples being compared, n_i and n_j are the sample sizes, and MSW is the within-group variance from the ANOVA table.

The Bonferroni test critical value or p-value is found by using the t-distribution with within degrees of freedom $df_w = N - k$, using an adjusted $\frac{\alpha}{m}$ two-tail area under the t-distribution, where $k =$ number of groups and $m = {}_kC_2$, all the combinations of pairs out of k groups.

Critical Value Method

Example 10-7: According to the ANOVA test that we previously performed there does appear to be a difference in the average age of assistant professors (μ_1), associate professors (μ_2), and full professors (μ_3) at her university.

Source	SS	df	MS	F
Between	1208.6667	2	604.3333	12.6195
Within	862	18	47.8889	
Total	2070.6667	20		

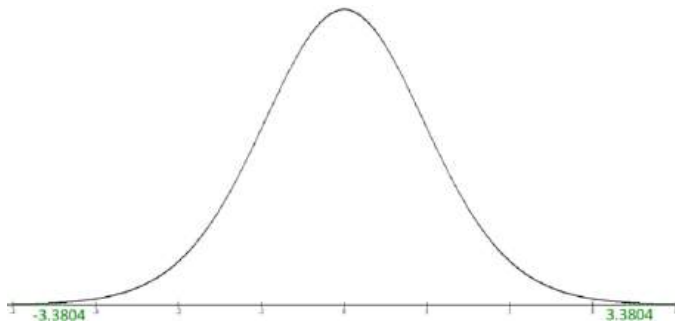
The hypotheses were: $H_0: \mu_1 = \mu_2 = \mu_3$
 $H_1: \text{At least one mean differs.}$

The decision was to reject H_0 , which means there is a significant difference in the mean age. The ANOVA test does not tell us, though, where the differences are. Determine which of the difference between each pair of means is significant. That is, test if $\mu_1 \neq \mu_2$, if $\mu_1 \neq \mu_3$, and/or if $\mu_2 \neq \mu_3$.

Solution: The alternative hypothesis for the ANOVA was “at least one mean is different.” There will be ${}_3C_2 = 3$ subsequent hypothesis tests to compare all the combinations of pairs (Group 1 vs. Group 2, Group 1 vs. Group 3, and Group 2 vs. Group 3). Note that if you have 4 groups then you would have to do ${}_4C_2 = 6$ comparisons, etc.

Use the t-distribution to find the critical value for the Bonferroni test. The total of all the individual sample sizes $N = 21$ and $k = 3$ then $m = {}_3C_2 = 3$, then the area for both tails would be $\frac{\alpha}{m} = \frac{0.01}{3} = 0.003333$.

This is a two-tailed test so the area in one tail is $\frac{0.003333}{2}$ with $df_w = N - k = 21 - 3 = 18$ gives C.V. = ± 3.3804 . The critical values are really far out in the tail so it is hard to see the shaded area. See Figure 10-9.



```
invT(0.003333/2,
18)
-3.380407352
```

Figure 10-9

Compare μ_1 and μ_2 :
 $H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2$

$$\text{The test statistic is } t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(MSW\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}} = \frac{37 - 52}{\sqrt{\left(47.8889\left(\frac{1}{7} + \frac{1}{7}\right)\right)}} = -4.0552.$$

Compare the test statistic to the critical value. Since the test statistic = $-4.0552 < \text{critical value} = -3.3804$, we reject H_0 .

There is enough evidence to conclude that there is a difference in the average age of assistant and associate professors.

Compare μ_1 and μ_3 :
 $H_0: \mu_1 = \mu_3$
 $H_1: \mu_1 \neq \mu_3$

$$\text{The test statistic is } t = \frac{\bar{x}_1 - \bar{x}_3}{\sqrt{\left(MSW\left(\frac{1}{n_1} + \frac{1}{n_3}\right)\right)}} = \frac{37 - 54}{\sqrt{\left(47.8889\left(\frac{1}{7} + \frac{1}{7}\right)\right)}} = -4.5958.$$

Compare the test statistic to the critical value. Since the test statistic = $-4.5958 < \text{critical value} = -3.3804$, we reject H_0 .

Reject H_0 , since the test statistic is in the lower tail. There is enough evidence to conclude that there is a difference in the average age of assistant and full professors.

Compare μ_2 and μ_3 :
 $H_0: \mu_2 = \mu_3$
 $H_1: \mu_2 \neq \mu_3$

$$\text{The test statistic is } t = \frac{\bar{x}_2 - \bar{x}_3}{\sqrt{\left(MSW\left(\frac{1}{n_2} + \frac{1}{n_3}\right)\right)}} = \frac{52 - 54}{\sqrt{\left(47.8889\left(\frac{1}{7} + \frac{1}{7}\right)\right)}} = -0.5407.$$

Compare the test statistic to the critical value. Since the test statistic is between the critical values $-3.3804 < -0.5407 < 3.3804$, we fail to reject H_0 .

Do not reject H_0 , since the test statistic is between the two critical values. There is enough evidence to conclude that there is not a difference in the average age of associate and full professors.

Note: you should get at least one group that has a reject H_0 , since you only do the Bonferroni test if you reject H_0 for the ANOVA. Also, note that the transitive property does not apply. It could be that group 1 = group 2 and group 2 = group 3, this does not mean that group 1 = group 3.

P-Value Method

Example 10-8: A research organization tested microwave ovens. At $\alpha = 0.10$, is there a significant difference in the average prices of the three types of ovens?

1000-watts	270	245	190	215	250	230		
900-watts	240	135	160	230	250	200	200	210
800-watts	180	155	200	120	140	180	140	130

Solution: The ANOVA was run in Excel.

SUMMARY				
Groups	Count	Sum	Average	Variance
1000-watts	6	1400	233.3333	796.6667
900-watts	8	1625	203.125	1549.554
800-watts	8	1245	155.625	795.9821

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	21729.73	2	10864.87	10.1182	0.001019	2.61
Within Groups	20402.08	19	1073.794			
Total	42131.82	21				

To test if there is a significant difference in the average prices of the three types of ovens, the hypotheses are:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_1 : At least one mean differs.

Use the Excel output to find the p-value in the ANOVA table of 0.001019, which is less than α so Reject H_0 , there is at least one mean that is different in the average oven prices.

There is a statistically significant difference in the average prices of the three types of ovens. Use the Bonferroni test p-value method to see where the differences are.

Compare μ_1 and μ_2 :

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(MSW\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}} = \frac{233.3333 - 203.125}{\sqrt{\left(1073.794\left(\frac{1}{6} + \frac{1}{8}\right)\right)}} = 1.7070$$

To find the p-value, find the area in both tails and multiply this area by m . The area to the right of the test statistic $t = 1.707$, using $df_w = 19$ is 0.0520563.

Remember these are always two-tail tests, so multiply this area by 2, to get both tail areas of 0.104113. Then multiply this area by $m = {}_3C_2 = 3$ to get a p-value = 0.3123.

```
tcdf(1.707, 1e99, 19)
.0520563372
Ans*2
.1041126743
Ans*3
.312338023
```

Since the p-value = 0.3123 > $\alpha = 0.10$ we do not reject H_0 .

There is not a statistically significant difference in the average price of the 1,000- and 900-watt ovens.

Compare μ_1 and μ_3 :

$$H_0: \mu_1 = \mu_3$$

$$H_1: \mu_1 \neq \mu_3$$

$$t = \frac{\bar{x}_1 - \bar{x}_3}{\sqrt{\left(MSW\left(\frac{1}{n_1} + \frac{1}{n_3}\right)\right)}} = \frac{233.3333 - 155.625}{\sqrt{\left(1073.794\left(\frac{1}{6} + \frac{1}{8}\right)\right)}} = 4.3910$$

```
tcdf(4.391, 1E99,
1.570414446E-4
Ans*2
3.140828892E-4
Ans*3
9.422486676E-4
```

Use $df_W = 19$ and the test statistic $t = 4.3910$ to find the p-value. Since the p-value = (tail areas)*3 = 0.00094 < $\alpha = 0.10$ we reject H_0 .

There is a statistically significant difference in the average price of the 1,000- and 800-watt ovens.

Compare μ_2 and μ_3 :

$$H_0: \mu_2 = \mu_3$$

$$H_1: \mu_2 \neq \mu_3$$

$$t = \frac{\bar{x}_2 - \bar{x}_3}{\sqrt{\left(MSW\left(\frac{1}{n_2} + \frac{1}{n_3}\right)\right)}} = \frac{203.125 - 155.625}{\sqrt{\left(1073.794\left(\frac{1}{8} + \frac{1}{8}\right)\right)}} = 2.8991$$

```
tcdf(2.8991, 1E99,
.0045984289
Ans*2
.0091968577
Ans*3
.0275905732
```

Use $df_W = 19$ and the test statistic $t = 2.8991$ to find the p-value (remember these are always two-tail tests).

Since the p-value = 0.0276 < $\alpha = 0.10$ we reject H_0 ,

There is a statistically significant difference in the average price of the 900- and 800-watt ovens.

There is a chance that after we multiply the area by the number of comparisons that the p-value would be greater than one. However, since the p-value is a probability, we would cap the probability at one.

This is a lot of math! The calculators and Excel do not have post-hoc pairwise comparisons shortcuts, but we can use the statistical software called SPSS to get the following results. We will look specifically at interpreting the SPSS output for Example 10-8.

Descriptives

Price

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
800	8	155.6250	28.21316	9.97486	132.0382	179.2118	120.00	200.00
900	8	203.1250	39.36437	13.91741	170.2156	236.0344	135.00	250.00
1000	6	233.3333	28.22528	11.52292	203.7127	262.9540	190.00	270.00
Total	22	194.0909	44.79148	9.54958	174.2315	213.9503	120.00	270.00

ANOVA

Price

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	21729.735	2	10864.867	10.118	.001
Within Groups	20402.083	19	1073.794		
Total	42131.818	21			

Multiple Comparisons

Dependent Variable: Price

Bonferroni

(I) Watts	(J) Watts	Mean Difference		Sig.	90% Confidence Interval	
		(I-J)	Std. Error		Lower Bound	Upper Bound
1000-Watts	900-Watts	30.20833	17.69717	.312	-10.3959	70.8126
	800-Watts	77.70833*	17.69717	.001	37.1041	118.3126
900-Watts	1000-Watts	-30.20833	17.69717	.312	-70.8126	10.3959
	800-Watts	47.50000*	16.38440	.028	9.9078	85.0922
800-Watts	1000-Watts	-77.70833*	17.69717	.001	-118.3126	-37.1041
	900-Watts	-47.50000*	16.38440	.028	-85.0922	-9.9078

*. The mean difference is significant at the 0.10 level.

Figure 10-10

The first table in Figure 10-4 gives descriptive statistics, the second table in Figure 10-10 is the ANOVA table, note the p-value is in the column labeled Sig. The Multiple Comparisons table is the third table in Figure 10-10, and this is where we want to look. There are repetitive pairs in the last rows of the table, just in a different order.

The first two rows in the Multiple Comparisons table are comparing group 1 with groups 2 and 3. If we follow the first row across under the Sig. column, this gives the p-value = 0.312 for comparing the 1,000- and 900-watt ovens.

1000-Watts	900-Watts	30.20833	17.69717	.312	-10.3959	70.8126
------------	-----------	----------	----------	------	----------	---------

The second row in the Multiple Comparisons table compares the 1,000- and 800-watt ovens, p-value = 0.001.

1000-Watts	900-Watts	30.20833	17.69717	.312	-10.3959	70.8126
	800-Watts	77.70833*	17.69717	.001	37.1041	118.3126

The third row in the Multiple Comparisons table compares the 900- and 1000-watt ovens in the reverse order as the first row, note the difference in the means is negative but the p-value is the same.

900-Watts	1000-Watts	-30.20833	17.69717	.312	-70.8126	10.3959
-----------	------------	-----------	----------	------	----------	---------

The fourth row in the Multiple Comparisons table compares the 900- and 800-watt ovens, the p-value = 0.028.

900-Watts	1000-Watts	-30.20833	17.69717	.312	-70.8126	10.3959
	800-Watts	47.50000*	16.38440	.028	9.9078	85.0922

The last set of rows in the Multiple Comparisons table are again repetitive and give the 800 compared to the 900- and 1000-watt ovens.

Keep in mind that post-hoc is defined as occurring after an event. A post-hoc test is done after an ANOVA test shows that there is a statistically significant difference. You should get at least one group that has reject H_0 , since you only do the Bonferroni test if you reject H_0 for the ANOVA.

Chapter 10 Formulas

One-Way ANOVA				
H ₀ : $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$				
H ₁ : At least one mean is different.				
Source	SS = Sum of Squares	df	MS = Mean Square	F
Between (Factor)	$\sum n_i(\bar{x}_i - \bar{x}_{GM})^2$	$k - 1$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within (Error)	$\sum(n_i - 1)s_i^2$	$N - k$	$MSW = \frac{SSW}{N-k}$	
Total	SST	$N - 1$		
\bar{x}_i = sample mean from the i^{th} group s_i^2 = sample variance from the i^{th} group n_i = sample size of the i^{th} group $N = n_1 + n_2 + \dots + n_k$ k = number of groups $\bar{x}_{GM} = \frac{\sum x_i}{N}$				
Bonferroni test statistic: $t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\left(MSW \left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right)}}$ H ₀ : $\mu_i = \mu_j$ H₁: $\mu_i \neq \mu_j$				
Multiply p-value by $m = {}_kC_2$, or divide area for critical value by $m = {}_kC_2$.				

Chapter 10 Exercises

- What is the primary purpose of conducting a one-way ANOVA?
 - To examine the relationship between two continuous variables.
 - To compare the means of multiple independent groups.
 - To analyze the difference between paired samples.
 - To determine the correlation coefficient between two variables.
- What does the p-value indicate in a one-way ANOVA?
 - The probability of observing a significant result if the null hypothesis is false.
 - The strength of the relationship between the independent and dependent variables.
 - The variability of the data within each group.
 - The probability of observing the F-statistic, or more extreme, given the null hypothesis is true.
- What does the acronym ANOVA stand for?
 - Analysis of Variance
 - Analysis of Means
 - Analyzing Various Means
 - Anticipatory Nausea and Vomiting
 - Average Noise Variance
- What would the test statistic equal if $MSB = MSW$?
 - 1
 - 0
 - 1
 - 4
 - 1.96
- What is the alternative hypothesis in a one-way ANOVA?
 - There is no difference between the group means.
 - There is a difference between the group means.
 - There is a linear relationship between the variables.
 - There is no relationship between the variables.
- A researcher would like to test to see if there is a difference in the average profit between 5 different stores. Which are the correct hypotheses for an ANOVA?
 - $H_0: \mu_1 = \mu_2 = \mu_3$ $H_1: \text{At least one mean is different.}$
 - $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ $H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$
 - $H_0: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$ $H_1: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
 - $H_0: \sigma_B^2 \neq \sigma_W^2$ $H_1: \sigma_B^2 = \sigma_W^2$
 - $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ $H_1: \text{At least one mean is different.}$
- Which assumption is required for conducting a one-way ANOVA?
 - The dependent variable is normally distributed.
 - Equal variances across the groups.

- c) Independence of observations.
 d) All of the above
8. When conducting a one-way ANOVA, if the calculated F-statistic is smaller than the critical value, what is the appropriate action?
- a) Conduct post-hoc tests.
 b) Reject the null hypothesis.
 c) Fail to reject the null hypothesis.
 d) Assume equal variances.
9. What is the purpose of post-hoc tests in a one-way ANOVA?
- a) To identify the specific group means that are significantly different from each other.
 b) To determine the overall significance of the analysis.
 c) To estimate effect sizes for each group.
 d) To assess the normality assumption of the data.
10. Calculate the critical value for a right-tailed F-test with a 5% level of significance with $df_1 = 4$ and $df_2 = 33$.
11. Calculate the critical value for a right-tailed F-test with a 1% level of significance with $df_1 = 3$ and $df_2 = 55$.
12. Calculate the critical value for a left-tailed F-test with a 10% level of significance with $df_1 = 29$ and $df_2 = 20$.
13. Calculate the critical values for a two-tailed F-test with a 1% level of significance with $df_1 = 31$ and $df_2 = 10$.
14. An ANOVA was run to test to see if there was a significant difference in the average cost between three different brands of snow skis. Random samples for each of the three brands were collected from different stores. Assume the costs are normally distributed. At $\alpha = 0.05$, test to see if there is a difference in the means. State the hypotheses, fill in the ANOVA table to find the test statistic, compute the p-value, state the decision and summary.

Source	SS	df	MS	F
Between	25.3633	2		
Within	23.3617	15		
Total	48.725	17		

15. An ANOVA test was run for the per-pupil costs for private school tuition for three counties in the Portland, Oregon, metro area. Assume tuition costs are normally distributed. At $\alpha = 0.05$, test to see if there is a difference in the means.

SUMMARY				
Groups	<i>n</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Clackamas County	11	147215	13383.1818	36734231.36
Multnomah County	12	182365	15197.0833	33731956.63
Washington County	10	124555	12455.5	40409869.17

- a) State the hypotheses.
 b) Fill out the ANOVA table to find the test statistic.

ANOVA				
Source	SS	df	MS	F
Between Groups				
Within Groups				
Total				

- c) Compute the p-value.
- d) State the correct decision and summary.

16. Cancer is a terrible disease. Surviving may depend on the type of cancer the person has. To see if the mean survival times for several types of cancer are different, data was collected on the survival time in days of patients with one of these cancers in advanced stage. The data is from "Cancer survival story," 2013. (Please realize that this data is from 1978. There have been many advances in cancer treatment, so do not use this data as an indication of survival rates from these cancers.) Does the data indicate that there is a difference in the mean survival time for these types of cancer? Use a 1% significance level.

SUMMARY				
<i>Groups</i>	<i>n</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Stomach	13	3718	286	119930.333
Bronchus	17	3597	211.5882	44040.6324
Colon	17	7776	457.4118	182473.007
Ovary	6	5306	884.3333	1206875.47
Breast	11	15355	1395.9091	1535038.49

- a) State the hypotheses.
- b) Construct an ANOVA table and calculate the test statistic.
- c) Compute the p-value.
- d) State the correct decision and summary.

17. What does the Bonferroni comparison test for?

- a) The analysis of between and within variance.
- b) The difference between all the means at once.
- c) The difference between two pairs of mean.
- d) The sample size between the groups.

18. True or False: The Bonferroni test should only be done when you reject the null hypothesis F-test?

19. True or False: The Bonferroni post-hoc test compares each group mean to the grand mean.

20. True or False: The Bonferroni post-hoc test adjusts the p-values for each pairwise comparison.

21. When should you use a Bonferroni post-hoc test?

- a) When the overall ANOVA test is not statistically significant.
- b) When the overall ANOVA test is statistically significant.
- c) When you want to compare group means to a control group.
- d) When you have a large sample size.
- e) When the sample sizes are unequal.

22. Which of the following statements is true about the Bonferroni post-hoc test?

- a) It is only applicable when there are more than three groups in the analysis.
- b) It increases the risk of Type I errors.
- c) It is not suitable for datasets with unequal sample variances.
- d) It provides adjusted p-values for each pairwise comparison.
- e) It is not suitable for datasets with equal sample sizes.

23. A manufacturing company wants to see if there is a significant difference in three types of plastic for a new product. They randomly sample prices for each of the three types of plastic and run an ANOVA. Use $\alpha = 0.05$

to see if there is a statistically significant difference in the mean prices. Part of the computer output is shown below.

SUMMARY				
<i>Groups</i>	<i>n</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Plastic A	39	512	13.12821	15.48313
Plastic B	41	679	16.56098	1.302439
Plastic C	34	470	13.82353	22.08913

- State the hypotheses.
- Fill in the ANOVA table to find the test statistic.

ANOVA				
Source of Variation	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Between Groups				
Within Groups				
Total	1631.939	113		

- Compute the critical value.
- State the decision and summary.
- Which group(s) are significantly different based on the Bonferroni test?

24. A manager on an assembly line wants to see if they can speed up production by implementing a new switch for their conveyor belts. There are four switches to choose from and replacing all the switches along the assembly line will be quite costly. They test out each of the four designs and record assembly times. Use $\alpha = 0.05$ to see if there is a statistically significant difference in the mean times.

ANOVA				
Source of Variation	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Between Groups	80.3951	3		
Within Groups	875.9111	87		
Total	956.3062	90		

- State the hypotheses.
- Fill in the ANOVA table to find the test statistic.
- Compute the critical value.
- State the correct decision and summary.
- Should a post-hoc Bonferroni test be done? Why?
 - No, since the $p\text{-value} > \alpha$ there is no difference in the means.
 - Yes, we should always perform a post-hoc test after an ANOVA
 - No, since we already know that there is a difference in the means.
 - Yes, since the $p\text{-value} < \alpha$ we need to see where the differences are.
- All four new switches are significantly faster than the current switch method. Of the four new types of switches, switch 3 cost the least amount to implement. Which of the 4, if any, should the manager choose?
 - The manager should stay with the old switch method since we failed to reject the null hypothesis.
 - The manager should switch to any of the four new switches since we rejected the null hypothesis.
 - The manager should randomly pick from switch types 1, 2 or 4.
 - Since there is no statistically significant difference in the mean time, they should choose switch 3 since it is the least expensive.

For exercises 25-32, Assume that all distributions are normal with equal population standard deviations, and the data was collected independently and randomly. Show all 5 steps for hypothesis testing. If there is a significant difference is found, run a Bonferroni test to see which means are different.

- State the hypotheses.
- Compute the test statistic.
- Compute the critical value or p-value.
- State the decision.
- Write a summary.

25. Is a statistics class's delivery type a factor in how well students do on the final exam? The table below shows the average percent on final exams from several randomly selected classes that used the different delivery types. Use a level of significance of $\alpha = 0.10$.

Face-to-Face	Blended	Online
79	70	100
77	58	66
75	55	91
68	74	91
95	76	98
78	83	74
69	66	57
65		88
65		

26. The dependent variable is the number of times a photo gets a like on social media. The independent variable is the subject matter, selfie or people, landscape, meme, or a cute animal. The researcher is exploring whether the type of photo makes a difference on the mean number of likes. A random sample of photos were taken from social media. Test to see if there is a significant difference in the means using $\alpha = 0.05$.

Selfie or People	Landscape	Meme	Cute Animal
9	17	13	14
16	15	12	10
18	13	15	13
6	14	20	16
10	17	21	18
16	9	19	14
17	24	21	13
12	23	22	19
18	21	14	17
7	6	20	13
14	17	10	18

27. The dependent variable is movie ticket prices, and the groups are the geographical regions where the theaters are located (suburban, rural, urban). A random sample of ticket prices were taken from randomly chosen states. Test to see if there is a significant difference in the means using $\alpha = 0.05$.

Suburb	Rural	Urban
11.25	11.75	11.25
11	9.5	11.25
11	11.25	12.25
12.25	10.5	9.75
11.25	10	10.75

Suburb	Rural	Urban
10	10	11.75
8.75	11.5	12
11	10.75	12.5
10.75	10.25	11
10.75	9.25	10.75
11.5	10.75	12
9.75	10	12
12.25	13	10.75
9.75	11	10.5
9.25	12	12.75

28. Recent research indicates that the effectiveness of antidepressant medication is directly related to the severity of the depression (Khan, Brodhead, Kolts & Brown, 2005). Based on pre-treatment depression scores, patients were divided into four groups based on their level of depression. After receiving the antidepressant medication, depression scores were measured again and the amount of improvement was recorded for each patient. The following data are similar to the results of the study. Use a significance level of $\alpha = 0.05$. Test to see if there is a difference in the mean scores.

Low Moderate	High Moderate	Moderately Severe	Severe
1.3	2.3	1.7	2.6
0.8	0.3	2.7	4.4
2.2	1.2	2.1	3.6
1.6	2.1	2.7	3.1
0.5	1.3	1.7	2.1
1.4	1.7	1.8	2
2.1	2.4	3.1	2.4
1.3	1.4	2.1	2.9
0	3.7	1.5	3.2
4	3.8	3.5	1.9
2	2.8	1.9	3.1
2.4	2.5	1.5	3
3.8	1.5	2.9	2
2.2	2.7	1.6	2.2
2.3	1.2	2.1	3.7
1.5	0.5	3.4	3.5
0.5	2	2.5	3.4
2.1	3.4	2.9	3
0.5	0.9	2.2	4.8
1.7	1.6	4.4	2.8
0.8	3.5	1.8	1.5
1.3	2.5	1.5	6.1
1.3	2.8	3.4	1.7
0.8	3.7	2.9	0.2
2.5	1.8	2.7	1.9

29. An ANOVA was run to test to see if there was a significant difference in the average cost between three different types of fabric for a new clothing company. Random samples for each of the three fabric types was collected from different manufacturers. At $\alpha = 0.10$, run an ANOVA test to see if there is a difference in the means.

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
A	34	10608	312	204.1212
B	37	11655	315	97.5
C	32	9600	300	2019.548

30. An ANOVA was run to test see if there was a significant difference in the average response time for three different training levels for police dispatchers. At $\alpha = 0.05$, run an ANOVA test to see if there is a difference in the means.

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Level 1	18	93.69	5.205	0.316932
Level 2	20	98.9	4.945	0.312111
Level 3	16	60.88	3.805	0.134

31. A researcher is testing to see if there is a difference in the average per-pupil costs for private school tuition for three counties in the Portland, Oregon, metro area. Assume tuition costs are normally distributed. The following table shows random samples for the per-pupil costs for private school tuition in thousands of dollars. Using a significance level of 5%, test to see if there is a significant difference.

Clackamas	Multnomah	Washington
15.74	14.97	14.77
14.66	12.28	15.2
14.6	12.94	15.13
15.17	11.33	15.35
14.83	13.27	14.8
15.06	13.38	14.45
14.54	12.95	14.64
15.13	11.86	13.97
14.76	13.83	15.12
15.3	12.48	15.42
15.19	10.68	
14.94	12.29	
14.96	11.25	
15.1	12.31	
14.96		
15.2		

32. Three students, Linda, Tuan, and Javier, are given laboratory rats for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her 9 rats Formula A, Tuan feeds his 9 rats Formula B, and Javier feeds his 9 rats Formula C. At the end of a specified time-period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 0.10, test to see if there is a difference in the mean weight gain for the three formulas.

Formula A	44.2	39.3	29.5	47.9	37.3	53.6	31.2	49.9	45.1
Formula B	10.5	52.8	49.2	32.1	33.5	37.3	27.2	13.1	26.3
Formula C	42.8	41	32.8	36.5	50.2	39.7	43.6	47.7	44.2

Chapter 11

Regression Analysis



- 11.1 Correlation
- 11.2 Hypothesis Test for a Correlation
- 11.3 Hypothesis Testing for Linear Regression
- 11.4 Coefficient of Determination
- 11.5 Residual Analysis
- 11.6 Outliers
- 11.7 Prediction Interval
- 11.8 Linear Regression Analysis

11.1 Correlation

We are often interested in the relationship between two variables. This chapter determines whether a significant linear relationship exists between sets of quantitative data and making predictions for a population. For instance, the relationship between the number of hours of study time and an exam score, or smoking and heart disease.

A predictor variable (also called the independent or explanatory variable; usually we use the letter x) explains or causes changes in the response variable. The predictor variable can be manipulated or changed by the researcher.

A response variable (also called the dependent variable; usually we use the letter y) measures the outcome of a study. The different outcomes for a dependent variable are measured or observed by the researcher. For instance, suppose we are interested in how much time spent studying affects the scores on an exam. In this study, study time is the predictor variable, and exam score is the response variable.

In data from an experiment, it is much easier to know which variable we should use for the independent and dependent variables. This can be harder to distinguish in observational data. Think of the dependent variable as the variable that you are trying to learn about.

If we were observing the relationship between unemployment rate and economic growth rate, it may not be clear which variable should be x and y . Do we want to predict the unemployment rate or the economic growth rate? One should never jump to a cause and effect reasoning with observational data. Just because there is a strong relationship between unemployment rate and economic growth rate does not mean that one causes the other to change directly. There may be many other contributing factors to both of these rates changing at the same time, such as retirements or pandemics.

Calculating Correlation

As discussed in Chapter 3, to calculate the correlation coefficient by hand we would use the following formulas.

Sum of Squares (SS) for Linear Regression		
$SS_{xx} = (n - 1)s_x^2$	$SS_{yy} = (n - 1)s_y^2$	$SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y}$

Sample Correlation Coefficient
$r = \frac{\sum((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}} = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}}$

Example 11-1: Use the following data to calculate the correlation coefficient.

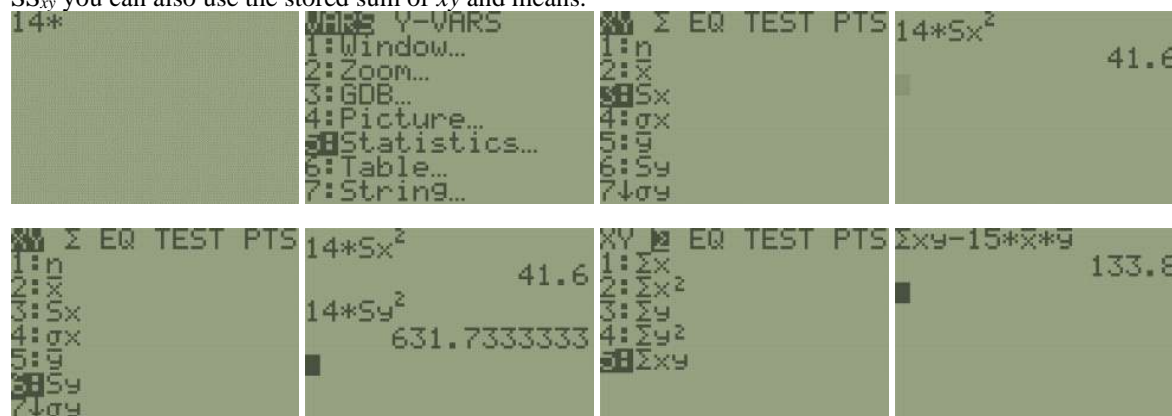
Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: We could show all the work the long way by hand using the shortcut formula. On the TI-84 press the [STAT] key and then the [EDIT] function, type the x values into L_1 and the y values into L_2 . Press the [STAT] key again and arrow over to highlight [CALC], select 2-Var Stats, then press [ENTER]. This will return the descriptive stats.

The TI calculator can run descriptive statistics and quickly get everything we need to find the sum of squares. Go to STAT > CALC > 2-Var Stats. For TI-83, you may need to enter your list names separated by a comma, for example 2-Var Stats L_1, L_2 then press enter. On the TI-89, open the Stats/List Editor. Enter all x -values in one list. Enter all corresponding y -values in a second list. Press F4, then select 2-Var Stats, then press [ENTER]. This will return the descriptive stats. Use the down arrow to see everything.



Once you do this the statistics are stored in your calculator so you can use the VARS key, go to Statistics, then select the standard deviation for x , repeat for the y -variable. This will reduce rounding errors by using exact values. For the SS_{xy} you can also use the stored sum of xy and means.



This gives:

$$SS_{xx} = (n - 1)s_x^2 = (15 - 1)1.723783215^2 = 41.6$$

$$SS_{yy} = (n - 1)s_y^2 = (15 - 1)6.717425811^2 = 631.7333$$

$$SS_{xy} = \Sigma(xy) - n \cdot \bar{x} \cdot \bar{y} = 20087 - (15 \cdot 16.6 \cdot 80.133333) = 133.8$$

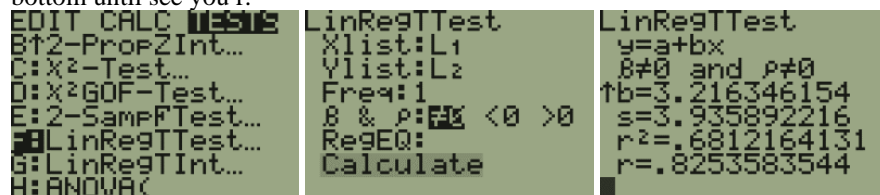
Note that both SS_{xx} and SS_{yy} will always be positive, but SS_{xy} could be negative or positive. For the TI-89, you will see the sum of squares at the very bottom of the descriptive statistics $\Sigma(x - \bar{x})^2 = 41.6$ and $\Sigma(y - \bar{y})^2 = 631.7333$.

To find the correlation, substitute the three sums of squares into the formula to get: $r = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}} = \frac{133.8}{\sqrt{(41.6 \cdot 631.7333)}}$
 $= 0.8524$. Try this now on your calculator to see if you are getting your order of operations correct.

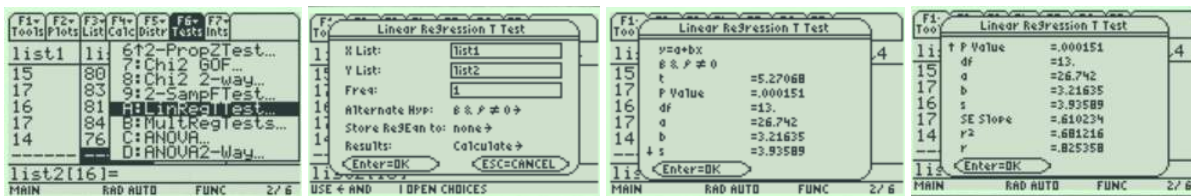
For our example, $r = 0.8254$ is close to 1, therefore it looks like there is positive linear relationship between the number of hours studying for an exam and the grade on the exam.

Most software has a built-in correlation function.

TI-84: Press the [STAT] key and then the [EDIT] function, type the x values into L_1 and the y values into L_2 . Press the [STAT] key again and arrow over to highlight [TEST], select LinRegTTest, then press [ENTER]. The default is Xlist: L_1 , Ylist: L_2 , Freq:1, β and $\rho: \neq 0$. Arrow down to Calculate and press the [ENTER] key. Scroll down to the bottom until see you r .



TI-89: On the TI-89, open the Stats/List Editor. Enter all x -values in one list. Enter all corresponding y -values in a second list. Press F6, then select LinRegTTest, then press [ENTER]. Scroll down to the bottom of the output to see r .



Excel:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14				
2	Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76		=CORREL(B1:P1,B2:P2)		

$$r = \text{CORREL}(\text{array1}, \text{array2}) = \text{CORREL}(B1:P1, B2:P2) = 0.8254$$

When is a correlation statistically significant? The next section shows how to run a hypothesis test for correlations.

11.2 Hypothesis Test for a Correlation

One should perform a hypothesis test to determine if there is a statistically significant correlation between the independent and the dependent variables. The **population correlation coefficient** ρ (this is the Greek letter rho, sounds like “row” not a p) is the correlation among all possible pairs of data values (x, y) taken from a population.

We will only be using the two-tailed test for a population correlation coefficient ρ .

The hypotheses are: $H_0: \rho = 0$
 $H_1: \rho \neq 0$.

The null-hypothesis of a two-tailed test states that there is no correlation (not a linear relation) between x and y . The alternative-hypothesis states that there is a significant correlation (there is a linear relation) between x and y .

The t-test is a statistical test for the correlation coefficient. It can be used when x and y are linearly related, the variables are random variables, and when the population of the variable y is normally distributed.

$$\text{The formula for the t-test statistic is } t = r \sqrt{\frac{n-2}{1-r^2}}.$$

Use the t-distribution with degrees of freedom equal to $df = n - 2$.

Note the $df = n - 2$ since we have two variables x and y .

Example 11-2: Test to see if the correlation for hours studied on the exam and grade on the exam is statistically significant. Use $\alpha = 0.05$.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution:

The hypotheses are $H_0: \rho = 0$
 $H_1: \rho \neq 0$.

Find the critical value using $df = n - 2 = 13$ for a two-tailed test $\alpha = 0.05$ inverse t function to get the critical values ± 2.160 . Draw the sampling distribution and label the critical values as shown in Figure 11-1.

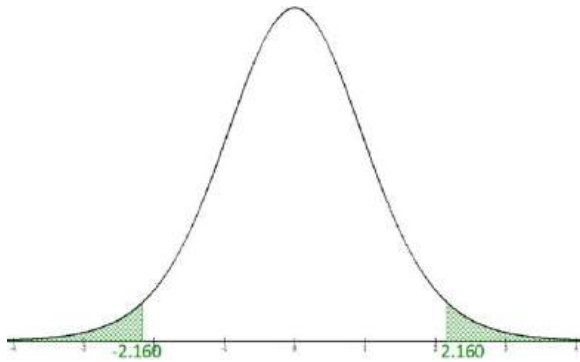


Figure 11-1

```
invT(.025,13)
-2.160368652
```

Next, find the test statistic $t = r \sqrt{\frac{n-2}{1-r^2}} = 0.8254 \sqrt{\frac{13}{(1-0.8254^2)}} = 5.271$, which is greater than 2.160 and in the rejection region.

The decision is to reject H_0 , since the test statistic is in the rejection area.

Summary: At the 5% significance level, there is enough evidence to support the claim that there is a statistically significant linear relationship (correlation) between the number of hours studied for an exam and exam scores.

The p-value method could also be used to find the same decision. We will use technology shortcuts for the p-value method. The p-value = $2 * P(t \geq 5.271 | H_0 \text{ is true}) = 0.000151$, which is less than $\alpha = 0.05$, therefore we reject H_0 .

Alternatively, we could test to see if the slope was equal to zero. If the slope is zero then the correlation will also be zero. The setup of a test is a little different, but we get the same results. Most software packages report the test statistic and p-value for a slope. This test is introduced in the next section.

TI-84: Enter the data in L_1 and L_2 . Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [LinRegTTest] and press the [ENTER] key. The default is Xlist: L_1 , Ylist: L_2 , Freq:1, β and $\rho: \neq 0$. Arrow down to Calculate and press the [ENTER] key. The calculator returns the t-test statistic, p-value and the correlation coefficient = r . Note the p-value = 0.0001513, is less than $\alpha = 0.05$, therefore reject H_0 , there is a significant correlation.

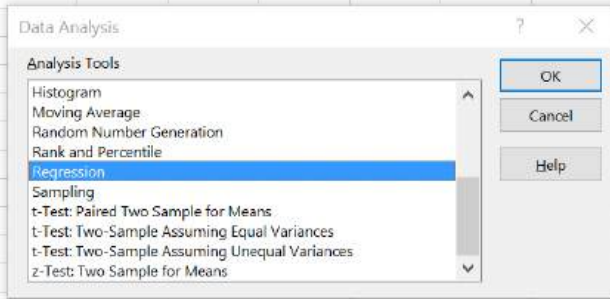
```
LinRegTTest
y=a+bx
b≠0 and ρ≠0
t=5.27068
P=1.5134617E-4
df=13
↓a=26.74198718
```

TI-89: Enter the data in List1 and List2. In the Stats/List Editor select F6 for the Tests menu. Use cursor keys to select A:LinRegTTest and press [Enter]. In the “X List” space type in the name of your list with the x variable without space, for our example “list1” or use [2nd] [Var-Link] and highlight list1. In the “Y List” space type in the name of your list with the y variable without space, for our example “list2” or use [2nd] [Var-Link] and highlight list2. Under the “Alternate Hyp” menu select the β and $\rho: \neq 0$ option, this is the same as the question’s alternative hypothesis statement, then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the t-test statistic, p-value, and the correlation = r .



Excel: Type the data into two columns in Excel. Select the Data tab, then Data Analysis, then choose Regression and select OK.

	A	B	C	D	E	F	G	H	I
1	Hours Studied for Exam	Grade on Exam							
2		20	89						
3		16	72						
4		20	93						
5		18	84						
6		17	81						
7		16	75						
8		15	70						
9		17	82						
10		15	69						
11		16	83						
12		15	80						
13		17	83						
14		16	81						
15		17	84						
16		14	76						
17									



Be careful here. The second column is the y range, and the first column is the x range. Only check the Labels box if you highlight the labels in the input range. The output range is one cell reference where you want the output to start, and then select OK.

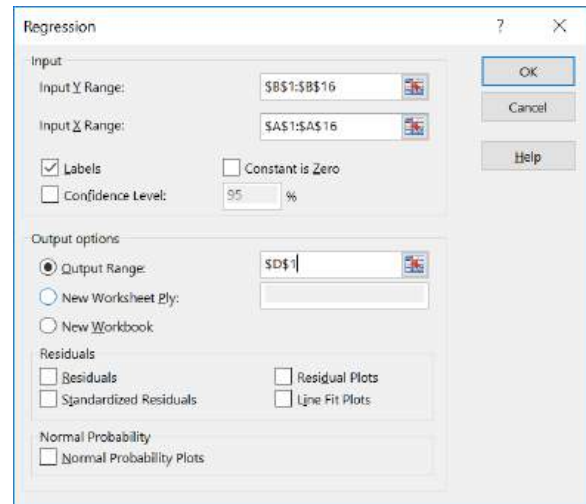


Figure 11-2 shows the regression output.

Regression Statistics	
Multiple R	0.825358
R Square	0.681216
Adjusted R Square	0.656695
Standard Error	3.935892
Observations	15

← Absolute value of the correlation coefficient $|r|$

← Sample size n

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	430.3471	430.3471	27.78002	0.000151
Residual	13	201.3862	15.49125		
Total	14	631.7333			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	26.74199	10.18074	2.626725	0.020917
Hours Studied for Exam	3.216346	0.610234	5.270675	0.000151

← t-test statistic ← p-value

Figure 11-2

When you reject H_0 , the slope is significantly different from zero. This means there is a significant relationship (correlation) between x and y and you can then find a regression line to use for prediction which we explore in the next section.

When looking at correlations, start with a scatterplot to see if there is a linear relationship prior to finding a correlation coefficient. If there is a linear relationship in the scatterplot then we can find the correlation coefficient to tell the strength and direction of the relationship. Clusters of dots forming a linear uphill pattern from left-to-right will have a positive correlation. The closer the dots in the scatterplot are to a straight line the closer r will be to 1. If the cluster of dots in the scatterplots go downhill from left-to-right in linear pattern, then there is a negative relationship. The closer those dots in the scatterplot are to a straight line going downhill, the closer r will be to -1 . Use a t -test to see if the correlation is statistically significant. As sample sizes get larger, smaller values of r become statistically significant. Be careful with outliers, which can heavily influence correlations. Most importantly, correlation is not causation. When x and y are significantly correlated, this does not mean that x causes y to change.

11.3 Hypothesis Testing for Linear Regression

A simple linear regression is a straight line that describes how the values of a response variable y change as the predictor variable x changes.

The equation of a line, relating x to y uses the slope-intercept form of a line, but with different letters than what you may be used to in a math class. We let b_0 represent the sample y -intercept (the value of y when $x = 0$), b_1 represent the sample slope (rise over run), and \hat{y} represent the predicted value of y for a specific value of x . The equation is written as $\hat{y} = b_0 + b_1x$.

Some textbooks and the TI calculators use the letter a to represent the y -intercept and b to represent the slope and the equation is written as $\hat{y} = a + bx$. These letters are just symbols representing the placeholders for the numeric values for the y -intercept and slope.

If we were to fit the best line that was closest to all the points on the scatterplot, we would get what we call the “line of best fit,” also known as the “regression equation” or “least squares regression line.”

Figure 11-3 is a scatterplot with just five points.

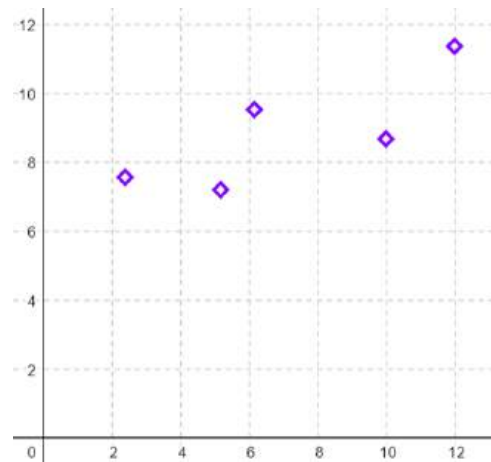


Figure 11-3

Figure 11-4 shows the **least-squares regression line** of y on x is the line that minimizes the squared vertical distance from all of the data. If we were to fit the line that best fits through the points, we would get the following picture.

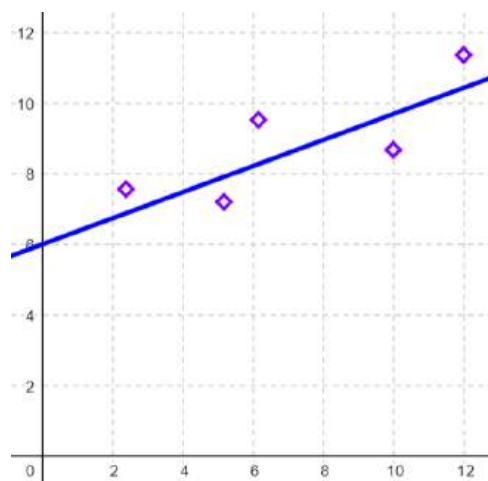


Figure 11-4

What we want to look for is the minimum of the squared vertical distance between each point and the regression equation, called a residual. This is where the name of least squares regression line comes from. Figure 11-5 shows the squared residuals.

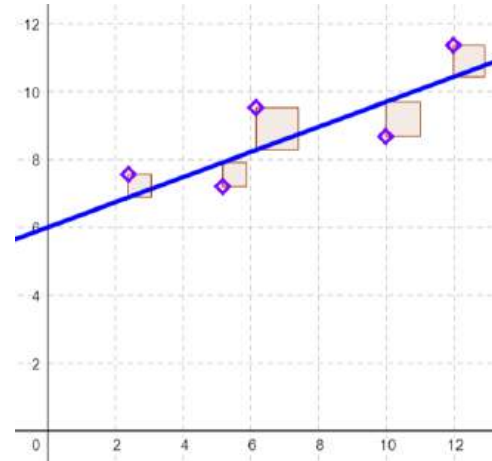


Figure 11-5

To compute the least squares regression line, you will need to first find the slope. Then substitute the slope into the following equation of the y-intercept: $b_0 = \bar{y} - b_1\bar{x}$, where \bar{x} = the sample mean of the x 's and \bar{y} = the sample mean of the y 's.

Regression Equation

To find the slope and y-intercept for the equation of the least-squares regression line $\hat{y} = b_0 + b_1x$ we use the following formulas: slope = $b_1 = \frac{SS_{xy}}{SS_{xx}}$, y-intercept: $b_0 = \bar{y} - b_1\bar{x}$.

Once we find the equation for the regression line, we can use it to estimate the response variable y for a specific value of the predictor variable x . Note: we would only want to use the regression equation for prediction if we reject H_0 and find that there is a significant correlation between x and y . Alternatively we could start with the regression equation and then test to see if the slope is significantly different from zero.

Example 11-3: Use the following data to find the line of best fit. Create a scatter plot with the regression equation.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: Start with finding the 2-Var Stats and sum of squares as shown in the steps for correlation.

$$SS_{xx} = (n - 1)s_x^2 = (15 - 1)1.723783215^2 = 41.6$$

$$SS_{yy} = (n - 1)s_y^2 = (15 - 1)6.717425811^2 = 631.7333$$

$$SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y} = 20087 - (15 \cdot 16.6 \cdot 80.133333) = 133.8$$

Calculate the slope: $b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{133.8}{41.6} = 3.216346$.

Calculate the y-intercept: $b_0 = \bar{y} - b_1 \cdot \bar{x} = 80.133333 - 3.216346 \cdot 16.6 = 26.742$.

Substitute these numbers back into the regression equation and write your answer as:
 $\hat{y} = 26.742 + 3.216346x$.

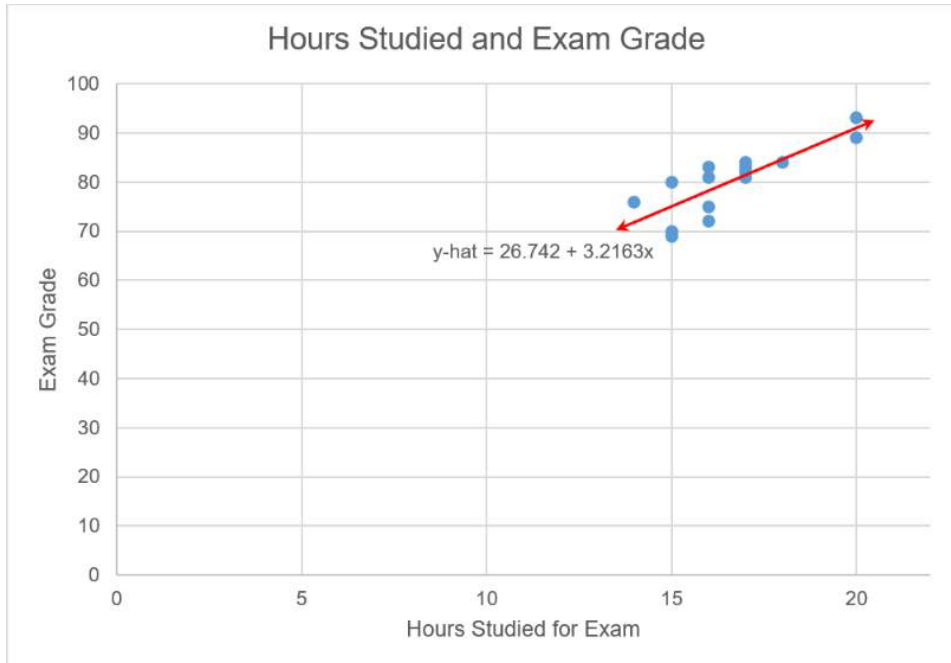
Interpreting the y-intercept coefficient: When $x = 0$, note that $\hat{y} = 26.742$, this means that we would expect a failing midterm score of 26.742 for students who had studied zero hours.

Interpreting the slope coefficient: For each additional hour studied for the exam, we would expect an increase in the midterm grade of 3.2163 points.

In general, when interpreting the slope coefficient, for each additional 1 unit increase in x , the predicted \hat{y} value will change by b_1 units.

Adding the Regression Line to the Scatterplot

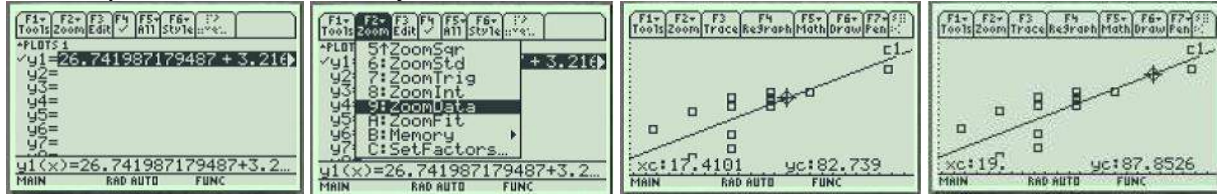
Excel: Follow the steps to create a scatter plot as shown in Chapter 2. Once the scatterplot is made, select the Add Chart Elements option, then select Trendline, then Linear.



TI-84: Make a scatterplot using the directions from the previous section. Turn your STAT scatter plot on. Press [Y=] and clear any equations that are in the y-editor. Into Y1, enter the least-squares regression equation manually as found above. Or, press the VARS key, go to option 5: Statistics, arrow over to EQ for equation, then choose the first option RegEQ. This will bring the equation over to the Y= menu without rounding error. Press [GRAPH]. You can press [TRACE] and use the arrow keys to scroll left or right. Pressing up or down on the arrow keys will change between tracing the scatterplot and the regression line. You can use the regression line to predict values of the response variable for a given value of the explanatory variable. While tracing the regression line type the value of the explanatory variable and press [ENTER]. For example, for $x = 19$ the value of $\hat{y} = 87.8526$.



TI-89: Make a scatterplot and find the regression line using the directions in the previous section. If you press [♦] then [F1] (Y=) you will notice the regression equation has been stored into y1 in the y-editor. Press [F2] **Trace** and use the left and right arrow keys to trace along the plot. Use the up and down arrow keys to toggle between the regression line and the scatterplot. You can use the regression line to predict values of the response variable for a given value of the explanatory variable. While tracing the regression line type the value of the explanatory variable and press [ENTER]. For example, for $x = 19$ the value of $\hat{y} = 87.8526$.



Hypothesis Test for Linear Regression

To test to see if the slope is significant, we will be doing a two-tailed test with hypotheses. The population least squares regression line would be $y = \beta_0 + \beta_1x + \varepsilon$ where β_0 (pronounced “beta-naught”) is the population y-intercept, β_1 (pronounced “beta-one”) is the population slope and ε is called the error term.

If the slope were horizontal (equal to zero), the regression line would give the same y value for every input of x and would be of no use. If there is a statistically significant linear relationship then the slope needs to be different from zero. We will only do the two-tailed test, but the same rules for hypothesis testing apply for a one-tailed test.

We will only be using the two-tailed test for a population slope.

The hypotheses are: $H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$.

The null-hypothesis of a two-tailed test states that there is not a linear relationship between x and y . The alternative-hypothesis of a two-tailed test states that there is a significant linear relationship between x and y .

Either a t-test or an F-test may be used to see if the slope is significantly different from zero. The population of the variable y must be normally distributed.

F-Test for Regression

An F-test can be used instead of a t-test. Both tests will yield the same results so it is a matter of preference and what technology is available.

Figure 11-6 is a template for a regression ANOVA table.

Source	SS = Sum of Squares	df	MS = Mean Square	F
Regression	$SSR = \frac{(SS_{xy})^2}{SS_{xx}}$	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error	$SSE = SS_{yy} - SSR$	$n - p - 1$	$MSE = \frac{SSE}{n - p - 1}$	
Total	$SST = SS_{yy}$	$n - 1$		

Figure 11-6

Where n is the number of pairs in the sample and p is the number of predictor (independent) variables, for now this is just $p = 1$. Use the F-distribution with degrees of freedom for regression = $df_R = p$, and degrees of freedom for error

$= dfE = n - p - 1$. This F-test is always a right-tailed test since ANOVA is testing the variation in the regression model is larger than the variation in the error.

Example 11-4: Use an F-test to see if there is a significant relationship between hours studied and grade on the exam, use $\alpha = 0.05$.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: The hypotheses are $H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$.

Compute the sum of squares.

$$SS_{xx} = 41.6, SS_{yy} = 631.7333, SS_{xy} = 133.8, n = 15 \text{ and } p = 1, SSR = \frac{(SS_{xy})^2}{SS_{xx}} = \frac{(133.8)^2}{41.6} = 430.3471154$$

$$SSE = SST - SSR = 631.7333 - 430.3471154 = 201.3862, SST = SS_{yy} = 631.7333$$

Compute the degrees freedom: $dfT = n - 1 = 14$ $dfE = n - p - 1 = 15 - 1 - 1 = 13$.

Compute the mean squares.

$$MSR = \frac{SSR}{p} = \frac{430.3471154}{1} = 430.3471154, MSE = \frac{SSE}{n-p-1} = \frac{201.3862}{13} = 15.4912$$

Compute the Test Statistic.

$$F = \frac{MSR}{MSE} = \frac{430.3471154}{15.4912} = 27.7801$$

Substitute the numbers into the ANOVA table:

Source	SS	df	MS	F
Regression	430.3471154	1	430.3471154	27.7801
Error	201.3862	13	15.4912	
Total	631.7333	14		

This is a right-tailed F-test with $df = 1, 13$ and $\alpha = 0.05$. In Excel we can find the critical value by using the function $=F.INV.RT(0.05,1,13) = 4.667$. The critical value of 4.667 is shown in Figure 11-7.

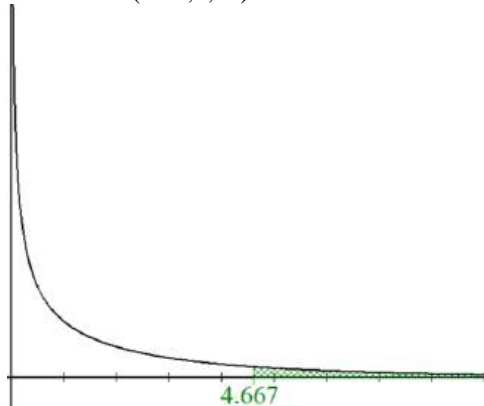


Figure 11-7

Or use the online calculator at <https://homepage.divms.uiowa.edu/~mbognar/applets/f.html>. It is hard to see the shaded tail in the applet in Figure 11-8, since the F-distribution is so close to the x-axis after 3, but it has the right-tail shaded from 4.667 and greater.

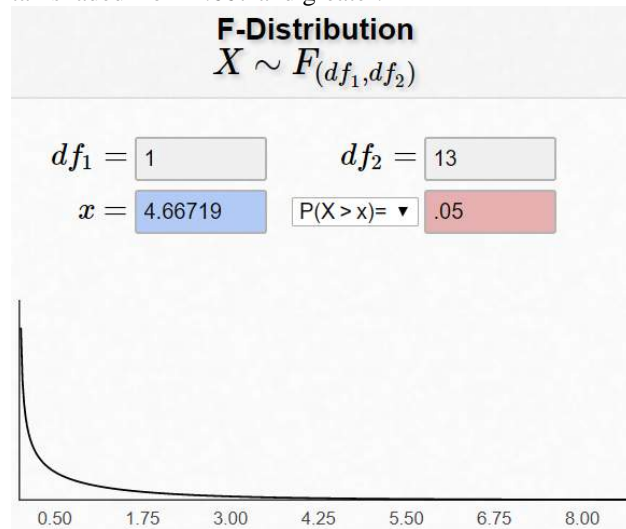


Figure 11-8

To visualize the critical value, refer back to Figure 11-7. The test statistic 27.78 is even further out in the tail than the critical value, so we would reject H_0 .

At the 5% level of significance, there is a statistically significant relationship between hours studied and grade on a student's final exam.

The p-value could also be used to make the decision. The p-value method would use the function =F.DIST.RT(27.78,1,13) = 0.00015 in Excel. The p-value is less than $\alpha = 0.05$ which also verifies that we reject H_0 .

The following is output from Excel and SPSS note the same ANOVA table information is shown but the columns are in a different order. Many statistical software packages will use the term significance to label where the p-value is located.

Excel

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	430.3471	430.3471	27.78002	0.000151
Residual	13	201.3862	15.49125		
Total	14	631.7333			

SPSS

		ANOVA ^a				
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	430.347	1	430.347	27.780	.000 ^b
	Residual	201.386	13	15.491		
	Total	631.733	14			

a. Dependent Variable: Exam Grade

b. Predictors: (Constant), Hours Studied

T-Test for Regression

If the regression equation has a slope of zero, then every x value will give the same y value and the regression equation would be useless for prediction. We should perform a t-test to see if the slope is significantly different from zero before using the regression equation for prediction. The numeric value of t will be the same as the t-test for a correlation. The two test statistic formulas are algebraically equal; however, the formulas are different and we use a different parameter in the hypotheses.

$$\text{The formula for the t-test statistic is } t = \frac{b_1}{\sqrt{\left(\frac{MSE}{SS_{xx}}\right)}}$$

Use the t-distribution with degrees of freedom equal to $df = n - p - 1$.

The t-test for slope has the same hypotheses as the F-test: $H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$.

Example 11-5: Use a t-test to see if there is a significant relationship between hours studied and grade on the exam, use $\alpha = 0.05$.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: The hypotheses are: $H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$.

Find the critical value using $df = n - p - 1 = 13$ for a two-tailed test $\alpha = 0.05$ inverse t-distribution to get the critical values ± 2.160 .

Draw the sampling distribution and label the critical values, see Figure 11-9.

The critical value is the same as we found using the t-test for correlation.

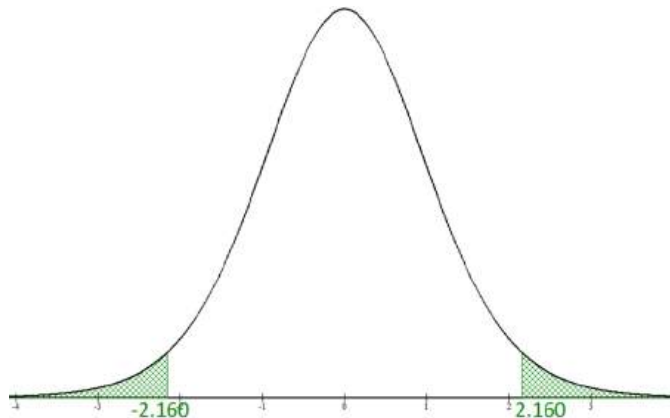


Figure 11-9

Next, find the test statistic $t = \frac{b_1}{\sqrt{\left(\frac{MSE}{SS_{xx}}\right)}} = \frac{3.216346}{\sqrt{\left(\frac{15.4912}{41.6}\right)}} = 5.2707$.

The test statistic value is the same value of the t-test for correlation even though they used different formulas. We look in the same place using technology as the correlation test. The test statistic is greater than the critical value of 2.160 and in the rejection region. The decision is to reject H_0 .

Summary: At the 5% significance level, there is enough evidence to support the claim that there is a significant linear relationship (correlation) between the number of hours studied for an exam and exam scores. The p-value method could also be used to find the same decision.

The p-value = 0.00015, the same as the previous tests. We will use technology for the p-value method. In the SPSS output, they use “Sig.” for the p-value and only rounds to three decimal places.

Excel

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	26.74199	10.18074	2.626725	0.020917
Hours Studied for Exam	3.216346	0.610234	5.270675	0.000151

TI-Calculator

```

LinRegTTest
y=a+bx
b≠0 and ρ≠0
t=5.270675167
P=1.5134617E-4
df=13
↓a=26.74198718
    
```

SPSS

		Coefficients^a				
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	26.742	10.181		2.627	.021
	Hours Studied	3.216	.610	.825	5.271	.000

a. Dependent Variable: Exam Grade

11.4 Coefficient of Determination

The **coefficient of determination** R^2 (or r^2) is the fraction (or percent) of the variation in the values of y that is explained by the least-squares regression of y on x . R^2 is a measure of how well the values of y are explained by x .

For example, there is some variability in the dependent variable values, such as grade. Some of the variation in student’s grades is due to hours studied and some is due to other factors. How much of the variation in a student’s grade is due to hours studied?

When considering this question, you want to look at how much of the variation in a student’s grade is explained by the number of hours they studied and how much is explained by other variables. Realize that some of the changes in grades have to do with other factors. You can have two students who study the same number of hours, but one student may have a higher grade. Some variability is explained by the model and some variability is not explained. Together, both of these give the total variability.

This is (Total Variation) = (Explained Variation) + (Unexplained Variation)

$$\sum(y - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y - \hat{y})^2$$

Coefficient of Determination

The proportion of the variation that is explained by the model is

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SSR}{SST}$$

We could also take the square of the correlation coefficient, $r^2 = R^2$ that we found in the previous section.

Example 11-6: Find and interpret the coefficient of determination for the hours studied and exam grade data.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: The coefficient of determination is this correlation coefficient squared. Note, when r is negative that when you square r the answer becomes positive. For the hours studied and exam grade $r = 0.825358$, so $r^2 = R^2 = 0.825358^2 = 0.6812$. Or use $SSR/SST = 430.3471154/631.7333 = 0.6812$.

Approximately 68% of the variation in a student's exam grade is explained by the least square regression equation and the number of hours a student studied.

You can also use the technology to find R^2 . Figure 11-10 shows the output for Excel, SPSS, and TI calculators.

Excel:

Regression Statistics	
Multiple R	0.825358
R Square	0.681216
Adjusted R Square	0.656695
Standard Error	3.935892
Observations	15

Coefficient of Determination R^2

TI-84:

```
LinRegTTest
y=a+bx
b≠0 and ρ≠0
↑b=3.216346154
s=3.935892216
r²=.6812164131
r=.8253583544
```

TI-89:

```
Linear Regression T Test
F1
Tool
11 P Value = .000151 .4
15 df = 13
17 a = 26.742
16 b = 3.21635
17 s = 3.93589
14 SE Slope = .610234
14 r² = .681216
14 r = .825358
112 Enter=OK
MAIN 8AP AUTO FUNC 27/8
```

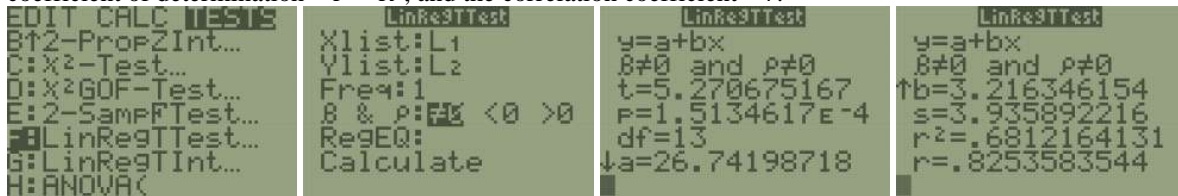
SPSS:

Model Summary

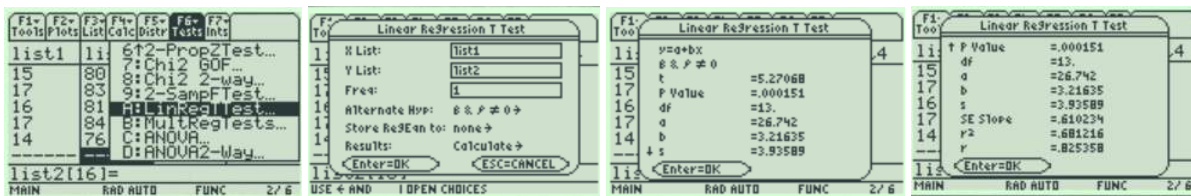
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825 ^a	.681	.657	3.93589

a. Predictors: (Constant), Hours Studied
Figure 11-10

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [LinRegTTest] and press the [ENTER] key. The default is Xlist:L1, Ylist:L2, Freq:1, β and $\rho \neq 0$. Arrow down to Calculate and press the [ENTER] key. The calculator returns the y-intercept = $a = b_0$, slope = $b = b_1$, the standard error of estimate = s , the coefficient of determination = $r^2 = R^2$, and the correlation coefficient = r .



TI-89: In the **Stats/List Editor** select F6 for the **Tests** menu. Use cursor keys to select **A:LinRegTTest** and press [Enter]. In the "X List" space type in the name of your list with the x variable without space, for our example "list1." In the "Y List" space type in the name of your list with the y variable without space, for our example "list2." Under the "Alternate Hyp" menu select the \neq sign that is the same as the problem's alternative hypothesis statement then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the y-intercept of the regression line = $a = b_0$, the slope of the regression line = $b = b_1$, the correlation = r , and the coefficient of determination = $r^2 = R^2$.



The coefficient of determination can take on any value between 0 and 1, or 0% to 100%. The closer R^2 is to 100% the better the regression equation models the data. Unlike r , which can only be used for simple linear regression, we can use R^2 for different types of regression. In more advanced courses, if we were to do non-linear or multiple linear regression, we could compare different models and pick the one that has the highest R^2 .

For instance, if we ran a linear regression on the scatterplot that showed an obvious curve pattern, we would get a regression equation with a zero slope and $R^2 = 0$. See Figure 11-11.

If we were to fit a parabola through the data, we get a perfect fit and $R^2 = 1$. See Figure 11-12.

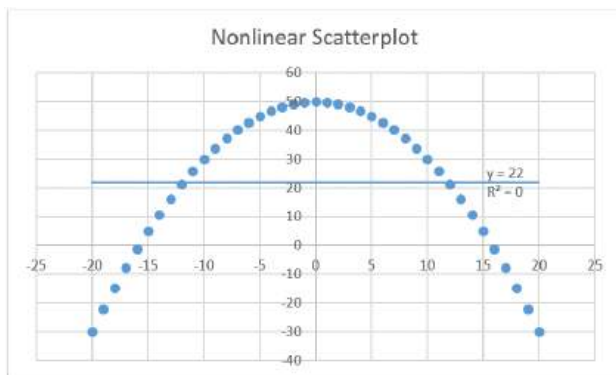


Figure 11-11

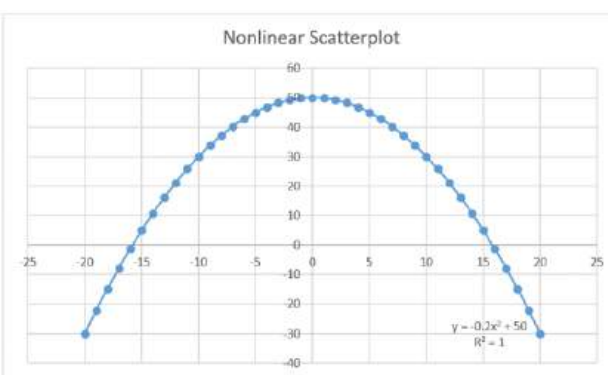


Figure 11-12

11.5 Residual Analysis

Residual analysis is a technique used in statistical analysis to assess the fit of a regression model and examine the assumptions underlying the model. The residuals represent the unexplained variation in the data, reflecting the discrepancies between the actual data points and the values predicted by the regression model. When we overlay the regression equation on a scatterplot, most of the time, the points do not lie on the line itself. The vertical distance between the actual value of y and the predicted value of \hat{y} is called the **residual**. The numeric value of the residual is found by subtracting the predicted value of y from the actual value of y , $y - \hat{y}$. When we find the line of best fit using least squares regression, this finds the regression equation with the smallest sum of the squared residuals $\sum(y - \hat{y})^2$.

When your residual is positive, then your data point is above the regression line, when the residual is negative, your data point is below the regression line. If you were to find the residuals for all the sample points and add them up you would get zero. The expected value of the residuals will always be zero. The regression equation is found so that there is just as much distance for the residuals above the line as there is below the line.

Example 11-7: Find the residual for the point (15, 80) for the exam data.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: Figure 11-13 is a scatterplot with the regression equation $\hat{y} = 26.742 + 3.216346x$ from the exam data.

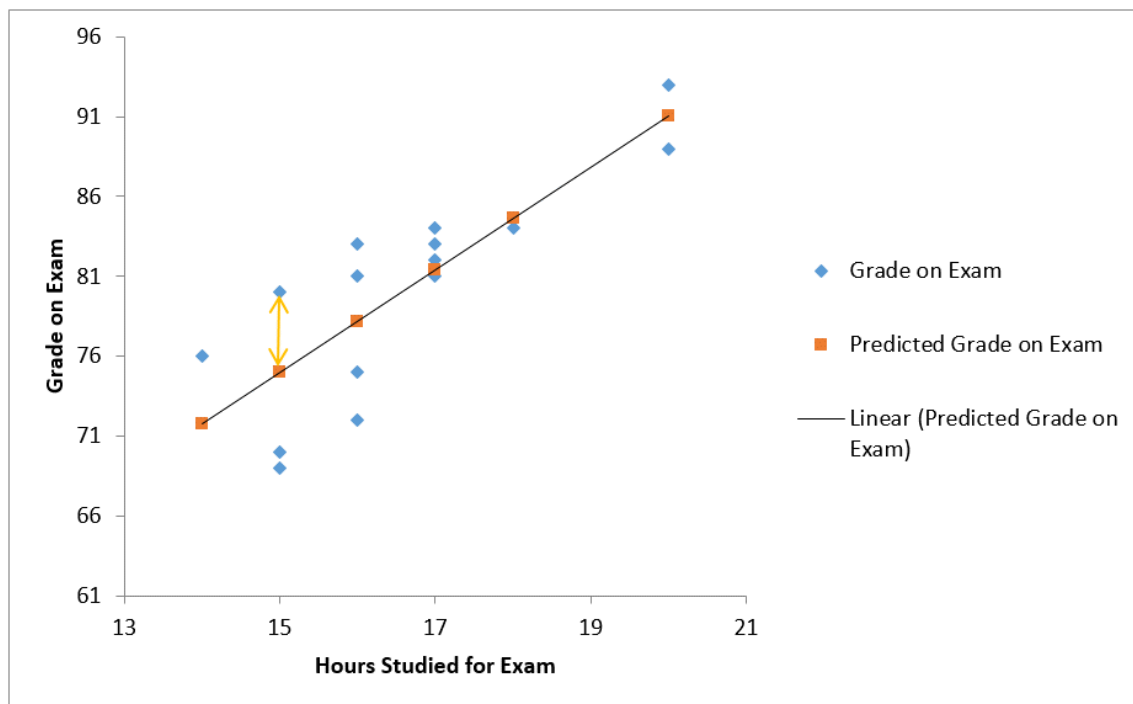


Figure 11-13

The blue diamonds represent the sample data points. The orange squares are the predicted \hat{y} for each value of x . If we connect the orange squares, we get the linear regression equation. The vertical distance between each data point and the regression equation is called the residual. The numeric value can be found by subtracting the observed y with its corresponding predicted value, $y - \hat{y}$. We use e_i to represent the i^{th} residual where $e_i = y_i - \hat{y}_i$. The residual for the point (15, 80) is drawn on the scatterplot vertically as a yellow double-sided arrow to visually show the size of the residual.

If you were to predict a student's exam grade when they studied 15 hours, you would get a predicted grade of $\hat{y} = 26.742 + 3.216346 \cdot 15 = 74.9865$. The residual for the point (15, 80) then would be $y - \hat{y} = 80 - 74.9865 = 5.0135$. This is the length of the vertical red dashed line connecting the point (15, 80) to the point (15, 74.9865).

Standard Error of Estimate

The standard deviation of the residuals is called the standard error of estimate or s . Some texts will use a subscript s_e or s_{est} to distinguish the different standard deviations from one another. When all of your data points line up in a perfectly straight-line $s = 0$, since none of your points deviate from the regression line. As your data points get more scattered away from a regression line, s gets larger. When you are analyzing a regression model you want s to be as small as possible.

Standard Error of Estimate

$$s_{\text{est}} = s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{MSE}$$

The **standard error of estimate** is the standard deviation of the residuals. The standard error of estimate measures the deviation in the vertical distance from data points to the regression equation. The units of s are the same as the units of y .

Example 11-8: Use the exam data to find the standard error of estimate.

Solution: To find the $\sum(y_i - \hat{y}_i)^2$ you would need to find the residual for every data point, square the residuals and sum them up. This is a lot of math. Recall the regression ANOVA table found earlier. The $MSE = 15.4912$.

Source	SS	df	MS	F
Regression	430.3471154	1	430.3471154	27.780
Error	201.3862	13	15.4912	
Total	631.7333	14		

The mean square error is the variance of the residuals, if we take the square root of the MSE we find the standard deviation of the residuals, which is the standard error of estimate.

$$s = \sqrt{MSE} = \sqrt{15.4912} = 3.9359$$

You can also use the technology to find s . Figure 11-14 shows the output for Excel, SPSS, and TI calculators.

Excel:

Regression Statistics	
Multiple R	0.825358
R Square	0.681216
Adjusted R Square	0.656695
Standard Error	3.935892
Observations	15

TI-84:

```
LinRegTTest
y=a+bx
b≠0 and ρ≠0
↑b=3.216346154
s=3.935892216
r²=.6812164131
r=.8253583544
```

TI-89:

```
Linear Regression T Test
1.1
1.1
1.5
1.6
1.7
1.6
1.7
1.4
1.1
1.1
```

s = standard error

SPSS:

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825 ^a	.681	.657	3.93589

a. Predictors: (Constant), Hours Studied

Figure 11-14

Residual Plots

Most software packages will plot the residuals for each x on the y -axis against either the x -variable or \hat{y} along the x -axis. This plot is called a residual plot. Residual plots help determine some of these assumptions. A good residual plot has a random scattering of points with no discernible pattern that form a horizontal band across the horizontal axis.

Example 11-9: Use technology to compute the residuals and make a residual plot for the hours studied and exam grade data.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: Plot the Residuals.

Excel: Run the regression the same as in the last section when testing to see if there is a significant correlation. Type the data into two columns in Excel. Select the Data tab, then Data Analysis, then choose Regression and select OK.

The screenshot shows an Excel spreadsheet with two columns: 'Hours Studied for Exam' (A1:A16) and 'Grade on Exam' (B1:B16). The 'Data Analysis' dialog box is open, with 'Regression' selected. The 'Regression' dialog box is also open, showing the following settings:

- Input Y Range: \$B\$1:\$B\$16
- Input X Range: \$A\$1:\$A\$16
- Labels
- Constant is Zero
- Confidence Level: 95%
- Output options: Output Range: \$D\$1
- Residuals: Residuals, Standardized Residuals, Residual Plots, Line Fit Plots
- Normal Probability: Normal Probability Plots

Be careful here, the second column is the y range, and the first column is the x range.

Only check the Labels box if you highlight the labels in the input range. The output range is one cell reference where you want the output to start.

Check the residuals, residual plots and normal probability plots, then select OK.

Figure 11-15 shows the Excel Output.

Regression Statistics	
Multiple R	0.825358
R Square	0.681216
Adjusted R Square	0.656695
Standard Error	3.935892
Observations	15

Annotations for Regression Statistics:

- Absolute value of the correlation coefficient $|r|$ (points to Multiple R)
- Coefficient of Determination R^2 (points to R Square)
- Standard Error of Estimate (points to Standard Error)
- Sample size n (points to Observations)

ANOVA	df	SS	MS	F	Significance F
Regression	1	430.3471	430.3471	27.78002	0.000151
Residual	13	201.3862	15.49125		
Total	14	631.7333			

Annotations for ANOVA:

- F-test statistic (points to F)
- F p-value (points to Significance F)

	Coefficients	Standard Error	t Stat	P-value
Intercept	26.74199	10.18074	2.626725	0.020917
Hours Studied for Exam	3.216346	0.610234	5.270675	0.000151

Annotations for Coefficients:

- y-intercept b_0 (points to Intercept)
- Slope b_1 (points to Hours Studied for Exam)
- t-test statistic (points to t Stat for Hours Studied for Exam)
- t p-value (points to P-value for Hours Studied for Exam)

Figure 11-15

Additional output from Excel gives the residuals, residual plot Figure 11-16, and normal probability plot Figure 11-17, see below.

RESIDUAL OUTPUT		
<i>Observation</i>	<i>Predicted Grade on Exam</i>	<i>Residuals</i>
1	91.0689	-2.0689
2	78.2035	-6.2035
3	91.0689	1.9311
4	84.6362	-0.6362
5	81.4199	-0.4199
6	78.2035	-3.2035
7	74.9872	-4.9872
8	81.4199	0.5801
9	74.9872	-5.9872
10	78.2035	4.7965
11	74.9872	5.0128
12	81.4199	1.5801
13	78.2035	2.7965
14	81.4199	2.5801
15	71.7708	4.2292

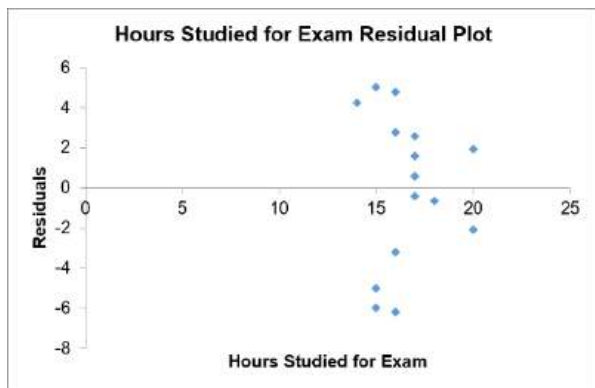


Figure 11-16

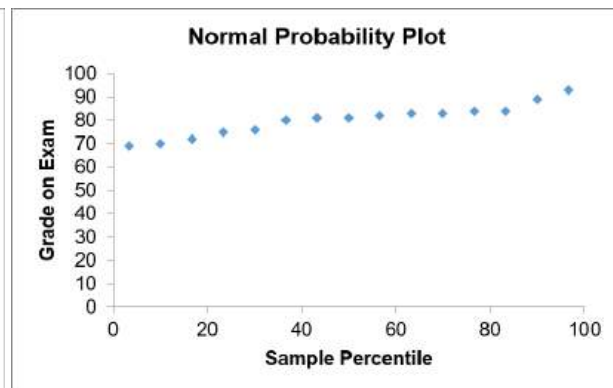


Figure 11-17

With this additional output, you can check the assumptions about the residuals. The residual plot should have no patterns and the normal probability plot forms should form an approximately straight line.

TI-84: Find the least-squares regression line as described in the previous section. Press [Y=] and clear any equations that are in the y-editor. Press [2nd] then [STAT PLOT] then press 1 or press [ENTER] to select **Plot1**. Select **On** and press [ENTER] to activate plot 1. For “Type” select the first graph that looks like a scatterplot and press [ENTER]. For “Xlist” enter whichever list where your explanatory variable data is stored. For our example, enter L₁. For “Ylist” press [2nd] [LIST] then scroll down to **RESID** and press [ENTER]. The calculator automatically computes the residuals and stores them in a list called RESID. Press [ZOOM] then press 9 or scroll down to **ZoomStat** and press [ENTER].



TI-89: Find the least-squares regression line as described in the previous section. Press [♦] then [F1] (Y=) and clear any equations that are in the y-editor. In the **Stats/List Editor** select F2 for the **Plots** menu. Use cursor keys to highlight **1:Plot Setup**. Make sure that the other graphs are turned off by pressing F4 button to remove the check marks. Under “Plot 2” press F1 for the **Define** menu. In the “Plot Type” menu select “Scatter.” In the “x” space type in the name of your list with the x variable without space, for our example “list1.” In the “y” space press [2ND] [-] for the **VAR-LINK** menu. Scroll down the list and find “resid” in the “STATVARS” menu. Press [ENTER] twice and you will be returned to the Plot Setup menu. Press F5 **ZoomData** to display the graph. Press F3 **Trace** and use the arrow keys to scroll along the different points.



When examining a residual plot, there are several characteristics to look for.

- **Constant Variance:** Homoscedasticity means that the spread of residuals should be relatively consistent across all levels of the predictor variables. If the residuals' spread increases or decreases with the predictor's values, it indicates heteroscedasticity and shows as a sideways V or fan shape. This may imply that the model's assumption of a constant variance is violated and transformations may be necessary.
- **Linearity:** Check for any systematic patterns, such as curves, U-shapes, or waves in the residual plot. If there are clear patterns in the residuals, such as a curved shape, it suggests that the linear model may not be appropriate.
- **Outliers:** Look for outliers or high-leverage points in the residual plot that deviate significantly from the random scatter. Outliers impact the regression model's fit.

A scatterplot for a person's age and how much time they spent on a video game is shown in Figure 11-18. The corresponding residual plot, Figure 11-19 exaggerates the curve seen in the scatterplot. The downward U shape of the residuals shows that a linear model is not a good fit, and we should use a non-linear regression model instead. Figure 11-20 shows the scatterplot and Figure 11-21 is the corresponding residual plot for cherry tree diameters and their corresponding heights. The residual plot for Figure 11-21 is scattered about the horizontal axis with no curve shapes or outliers, and a linear model could be a good fit. Figure 11-22 is the scatterplot and Figure 11-23 is the corresponding residual plot for the measurement depth and degrees Celsius for a geological mantel study. The residuals in Figure 11-23 fan out in a sideways V shape, which indicates heteroscedastic variance of the residuals. In this case there are more advanced techniques to fix the issue, beyond the scope of this course, such as a log transformation.

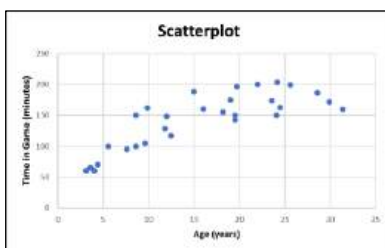


Figure 11-18

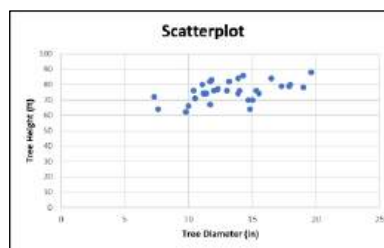


Figure 11-20

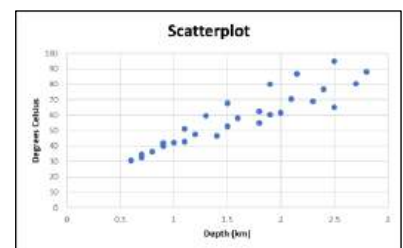


Figure 11-22

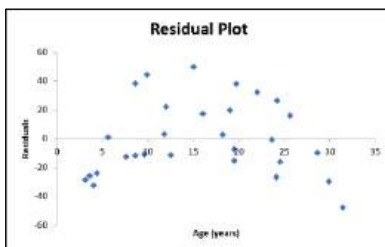


Figure 11-19

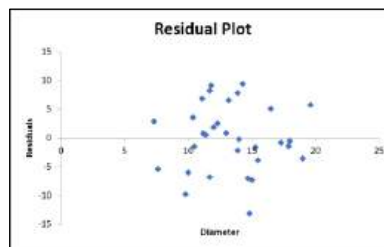


Figure 11-21

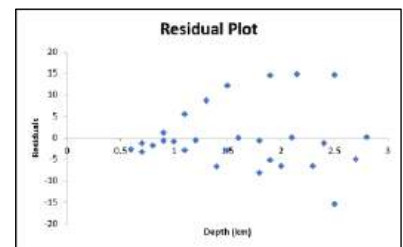


Figure 11-23

11.6 Outliers

Secondly, residual analysis can help identify influential observations or outliers that have a disproportionate impact on the regression results. **Outliers** are data points that deviate significantly from the overall pattern and can exert a strong influence on the estimated regression coefficients. By examining the residuals, we can identify these influential observations and assess their impact on the model's conclusions.

Example 11-10: Should linear regression be used with this data set?

x	1	3	8	2	1	3	2	2	3	1
y	2	3	8	2	3	1	3	1	2	1

Solution: A regression analysis for the data set was run on Excel.

Regression Statistics	
Multiple R	0.844
R Square	0.712
Adjusted R Square	0.676
Standard Error	1.176
Observations	10

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.406	0.618	0.658	0.529
x	0.844	0.19	4.446	0.002

If we test for a significant correlation: $H_0: \rho = 0$
 $H_1: \rho \neq 0$.

The correlation is $r = 0.844$ and the p-value is 0.002 which is less than $\alpha = 0.05$ so we would reject H_0 and conclude there is a significant linear relationship between x and y .

However, if look at the scatterplot in Figure 11-24, with the regression equation we can clearly see that the point (8,8) is an outlier. The outlier is pulling the slope up towards the point (8,8).

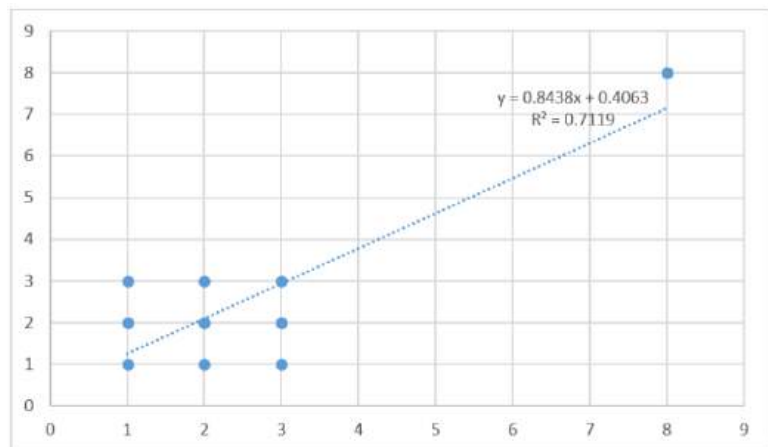


Figure 11-24

If we were to take out the outlier point (8,8) and run the regression analysis again on the modified data set we get the following Excel output.

Regression Statistics	
Multiple R	0
R Square	0
Adjusted R Square	-0.142857
Standard Error	0.92582
Observations	9

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	2	0.816	2.449	0.044
x	0	0.378	0	1

See Figure 11-25, note the correlation is now 0 and the p-value is 1 so there is no relationship at all between x and y .

This type of outlier is called a **leverage point**. Leverage points are positioned far away from the main cluster of data points. A leverage point can significantly influence the slope of the regression line. Leverage points can pull the regression line towards them or push the regression line away, depending on their location relative to the other observations. These points have high leverage because they exert leverage on the estimated regression coefficients. However, a leverage point may not necessarily have a strong influence on the response variable or the overall regression fit.

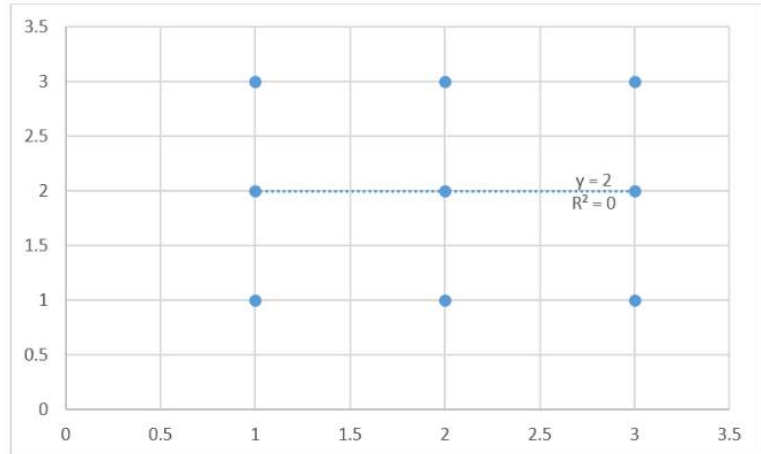


Figure 11-25

There is another type of outlier called an **influential point**. An influential point refers to an observation that has a substantial impact on the regression analysis, affecting both the estimated regression coefficients and the overall fit of the model. Influential points can arise from a combination of extreme values in both predictor and response variables, and they can significantly influence the slope, intercept, and overall shape of the regression line. These points have a high influence on the regression results, often leading to changes in the estimated coefficients, standard errors, and statistical significance of the model. There is an option in most software packages to get the “standardized” residuals. Standardized residuals are z -scores of the residuals. Any standardized residual that is not between -2 and 2 may be an outlier. If it is not between -3 and 3 then the point is an outlier. When this happens, the points are called influential points or influential observations. An outlier can be both influential and have high leverage.

Example 11-11: Use technology to compute the residuals and standardized residuals. Should linear regression be used with this data set?

Solution: A regression analysis for the following data set was run on Excel.

x	1	3	2	2	4	5	7	9	6	8
y	1	3	10	2	4	5	7	9	6	8

<i>Observation</i>	<i>Predicted y</i>	<i>Residuals</i>	<i>Standard Residuals</i>
1	2.974	-1.974	-0.831
2	4.339	-1.339	-0.564
3	3.656	6.344	2.671
4	3.656	-1.656	-0.698
5	5.022	-1.022	-0.430
6	5.705	-0.705	-0.297
7	7.070	-0.070	-0.030
8	8.436	0.564	0.237
9	6.388	-0.388	-0.163
10	7.753	0.247	0.104

The point (2, 10) shown in Figure 11-26 is pulling the left side of the line up away from the points that form a line. This influential point has high leverage and changes both the y-intercept and slope.

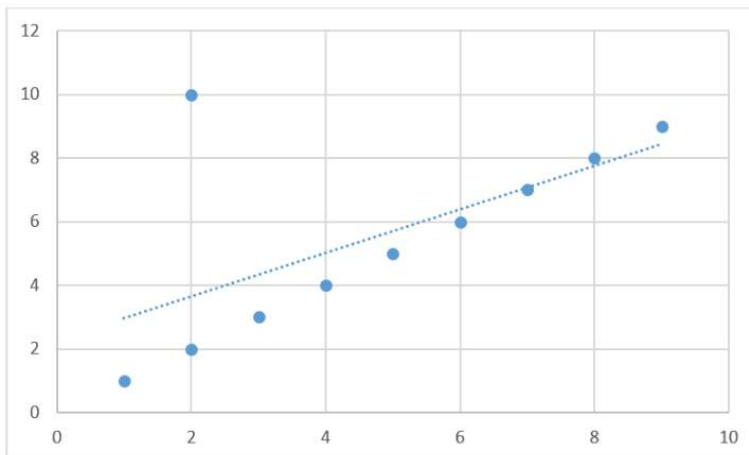


Figure 11-26

A **lurking variable** is a variable other than the independent or dependent variables that may influence the regression line. For instance, the highly correlated ice cream sales and home burglary rates probably have to do with the season. Hence, linear regression does **not** imply cause and effect.

Two variables are **confounded** when their effects on the dependent variable cannot be distinguished from each other. For instance, if we are looking at diet predicting weight, a confounding variable would be age. As a person gets older, they can gain more weight with fewer calories compared to when they were younger. Another example would be predicting someone’s midterm score from hours studied for the exam. Some confounding variables would be GPA, IQ score, and teacher’s difficulty level.

11.7 Prediction Interval

When there is a statistically significant linear relationship, we can use the least squares regression equation for prediction.

Example 11-12: Use the regression equation to predict the grade for a student who has studied for 18 hours for their exam using the previous data.

Hours Studied for Exam	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Grade on Exam	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Solution: We found the regression equation $\hat{y} = 26.742 + 3.216346x$. The x -variable is the hours studied so let $x = 18$ hours. \hat{y} is the symbol for the predicted y .

Substitute $x = 18$ into the regression equation and you get: $\hat{y} = 26.742 + 3.216346 \cdot 18 = 84.636228$.

We would estimate the student’s grade of 84.6 when they studied 18 hours. This is a point estimate for the grade on the exam for a student that studied 18 hours.

Prediction Interval

Every time we take a new sample this point estimate will change. We can find a special type of confidence interval to estimate the true value of y called the **prediction interval**. A prediction interval provides an estimate of the range within which a future observation or outcome is expected to fall, given a specific level of confidence. Unlike a confidence interval from previous chapters, which provided an interval estimate for a population parameter such as

the mean, a prediction interval gives a range of values that encompasses a new observation or outcome with a certain level of confidence. It takes into account both the variability of the data points around the regression line and the uncertainty in predicting a specific outcome.

The prediction interval is the confidence interval for the actual value of y .

$$\hat{y} \pm t_{\alpha/2} \cdot s \sqrt{\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_{xx}}\right)}$$

Where \hat{y} is the predicted value of y for the given value of x .

Example 11-13: Using the previous exam data, find and interpret the 95% prediction interval for a student who studies 18 hours.

Solution: From the question $x = 18$. From previous examples we found that $\hat{y} = 26.742 + 3.216346 \cdot 18 = 84.636228$ and $s = 3.935892$. Find the critical value from the invT using $df = n - 2 = 13$ we get $t_{\alpha/2} = 2.160369$. Make sure to go out at least 6 decimal places in between steps. Ideally you never round between steps. Use the 2-Var Stats from your calculator to find the sums and then substitute values back into the equation to get:

$$84.636228 \pm 2.1600369 \cdot 3.935892 \sqrt{\left(1 + \frac{1}{15} + \frac{(18-16.6)^2}{41.6}\right)}$$

$\Rightarrow 84.636228 \pm 8.9723$
 $\Rightarrow 75.6639 < y < 93.6085.$

We are 95% confident that the predicted exam grade for a student that studies 18 hours is between 75.6639 and 93.6085.

A confidence interval can be more accurate (narrower) when you increase the sample size. Note in the last example the predicted grade for an individual student could have been anywhere from a C to an A grade. If you wanted to predict y with more accuracy then you would want to sample more than 15 students to get a smaller margin of error. The confidence interval for a mean will have a smaller margin of error than for an individual's predicted value.

Extrapolation is the use of a regression line for prediction far outside the range of values of the independent variable x . As a general rule, one should not use linear regression to estimate values too far from the given data values. The further away you move from the center of the data set, the more variable results become. For instance, we would not want to estimate a student's grade for someone that studied way less than 14 hours or more than 20 hours.

Excel, the TI-83, and TI-84 do not have built in prediction intervals.

TI-89: Enter the x -values in list1 and the y -values in list2, select [F7] Intervals, then select option 7:LinRegTInt... Use the Var-Link button to enter in list1 and list2 for the X List and Y List. Select Response in the drop-down menu for Interval. Enter in the x -value given in the question. Change the confidence level (C-Level) to match what was in the question, the [Enter]. Scroll down to Pred Int for the prediction interval. The calculator does not round between steps so if you rounded b_0 and b_1 for instance when doing hand calculations your answer may be slightly different than the calculator results.



11.8 Linear Regression Analysis

Overall, linear regression analysis is a powerful statistical tool for exploring and modeling the relationship between variables, make predictions, and inform decision-making processes. The linear regression analysis helps us understand how changes in the independent variable affect the dependent variable.

Model Assumptions

There are assumptions that need to be met when running simple linear regression. If these assumptions are not met, then one should use more advanced regression techniques.

The assumptions for simple linear regression are:

- The data need to follow a linear pattern.
- The observations of the dependent variable y are independent of one another.
- Residuals are approximately normally distributed with a mean of zero.
- The variance of the residuals is constant (homoscedasticity).

Violations of these assumptions may affect the validity and reliability of the regression results, so it is essential to evaluate and address them. Additionally, if there is a statistically significant linear relationship, the analysis allows us to make predictions of the dependent variable based on the values of the independent variable.

Example 11-14: High levels of hydrogen sulfide (H_2S) in the ocean can be harmful to animal life. It is expensive to run tests to detect these levels. A scientist would like to see if there is a relationship between Sulfate (SO_4) and H_2S levels since SO_4 is much easier and less expensive to test in ocean water. A sample of SO_4 and H_2S were recorded together at different depths in the ocean. The sample is reported below in millimolar (mM). Run a regression analysis using a 5% level of significance, to see if there is a significant relationship between H_2S and SO_4 . If the model is significant, compute the 95% prediction interval to predict the H_2S level in the ocean when the SO_4 level is 25 mM.

Sulfate	22.5	27.5	24.6	27.3	23.1	24	24.5	28.4	25.1	24.4
Sulfide	0.6	0.3	0.6	0.4	0.7	0.5	0.7	0.2	0.3	0.7

Solution: Start with a scatterplot to see if a linear relation exists.

The scatterplot in Figure 11-27 shows a negative linear relationship. Test to see if the linear relationship is statistically significant. Use $\alpha = 0.05$. You could use an F- or a t-test. I would recommend the t-test if you are using a TI calculator and an F-test if you are using a computer program like Excel or SPSS. We will do the F-test for the following example.

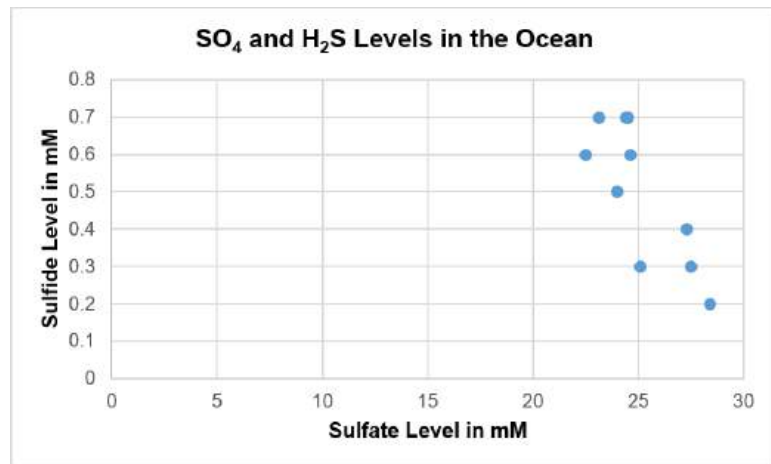


Figure 11-27

The hypotheses are: $H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$.

Compute the sum of squares.

$$SS_{xx} = (n - 1)s_x^2 = (10 - 1)1.959138^2 = 34.544$$

$$SS_{yy} = (n - 1)s_y^2 = (10 - 1)0.188561^2 = 0.32$$

$$SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y} = 123.04 - 10 \cdot 25.14 \cdot 0.5 = -2.66$$

Next, compute the test statistic and add the values to the ANOVA table.

$$SSR = \frac{(SS_{xy})^2}{SS_{xx}} = \frac{(-2.66)^2}{34.544} = 0.2048286$$

$$SST = SS_{yy} = 0.32$$

$$SSE = SST - SSR = 0.32 - 0.2048286 = 0.1151714$$

$$df_T = n - 1 = 9$$

$$df_E = n - p - 1 = 10 - 1 - 1 = 8$$

$$MSR = \frac{SSR}{p} = \frac{0.2048286}{1} = 0.2048286$$

$$MSE = \frac{SSE}{n-p-1} = \frac{0.1151714}{8} = 0.014396$$

$$F = \frac{MSR}{MSE} = \frac{0.2048286}{0.014396} = 14.228$$

Source	SS	df	MS	F
Regression	0.2048	1	0.2048	14.228
Error	0.1152	8	0.0144	
Total	0.32	9		

Compute the p-value. This is a right-tailed F-test with $df = 1, 8$, which gives a p-value of $=F.DIST.RT(14.2277, 1, 8) = 0.00545$.

We could also use Excel to generate the p-value. Excel uses “Significance F” for a the label for a p-value.

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	0.204829	0.204829	14.22775	0.00545
Residual	8	0.115171	0.014396		
Total	9	0.32			

The p-value = 0.00545 < $\alpha = 0.05$ therefore reject H_0 . There is a statistically significant linear relationship between hydrogen sulfide and sulfate levels in the ocean.

From the linear regression check the assumptions and make sure there are no outliers.

The standardized residuals are between -2 and 2 and the scatterplot do not indicate any outliers.

Standard Residuals
-0.91306
-0.16153
0.516413
0.586326
0.379351
-0.776
1.332336
-0.43289
-1.79521
1.264266

The Normal Probability Plot in Figure 11-28 forms an approximately straight line. This indicates that the residuals are approximately normally distributed.

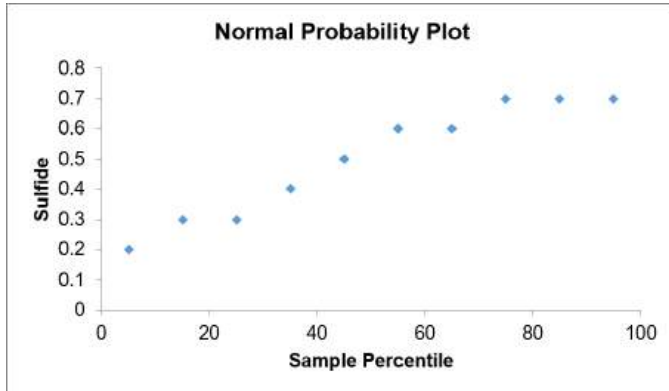


Figure 11-28

The residual plot in Figure 11-29 has no unusual pattern. This indicates that a linear model would work well for this data.

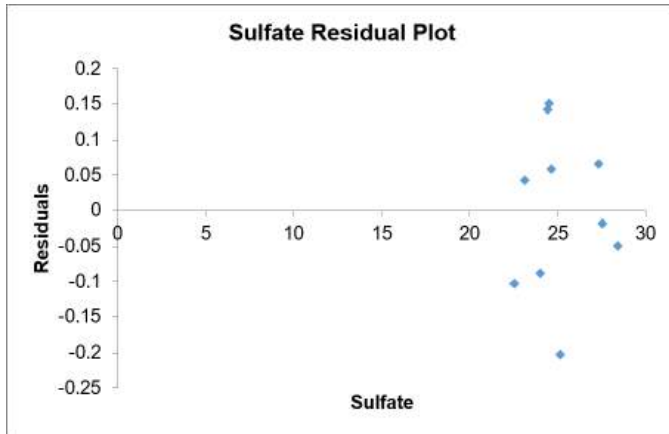


Figure 11-29

Now find and use the regression equation to calculate the 95% prediction interval to predict the sulfide level in the ocean when the sulfate level is 25 mM.

Find the regression equation. Calculate the slope: $b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-2.66}{34.544} = -0.077$.

Then calculate the y-intercept: $b_0 = \bar{y} - b_1 \cdot \bar{x} = 0.5 - (-0.077) \cdot 25.14 = 2.43586$.

Put the numbers back into the regression equation and write your answer as:
 $\hat{y} = 2.4359 + (-0.077)x$ or as $\hat{y} = 2.4359 - 0.077x$.

We can use technology to get the regression equation. Coefficients are found in the first column in the computer output.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	2.43586	0.5146	4.7333	0.0015	1.2491	3.6226
Sulfate	-0.0770032	0.0204	-3.7720	0.0055	-0.1241	-0.0299

We would expect variation in our predicted value every time a new sample is used. Find the 95% prediction interval to estimate the sulfide level when the sulfate level is 25 mM.

Use the prediction interval equation $\hat{y} \pm t_{\alpha/2} \cdot s \sqrt{\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}\right)}$.

Substitute $x = 25$ into the equation to get $\hat{y} = 2.43586 - 0.0770032 \cdot 25 = 0.51078$.

To find $t_{\alpha/2}$ use your invT with $df_E = n - 2 = 8$ and left-tail area $\alpha/2 = 0.05/2 = 0.025$, gives $t_{0.025} = \pm 2.306004$

The standard error of estimate $s = \sqrt{MSE} = \sqrt{0.014396} = 0.11998$ which can also be found using technology.

Regression Statistics	
Multiple R	0.800056
R Square	0.640089
Adjusted R Square	0.595101
Standard Error	0.119985
Observations	10

From the earlier descriptive statistics, we have $n = 10$, $\bar{x} = 25.14$, $SS_{xx} = 34.544$. Substitute each of these values into the prediction interval to get the following:

$$0.51078 \pm 2.306004 \cdot 0.119985 \sqrt{\left(1 + \frac{1}{10} + \frac{(25-25.14)^2}{34.544}\right)}$$

$$0.51078 \pm 0.290265$$

$$0.2205 < y < 0.8010$$



We can be 95% confident that the true sulfide level in the ocean will be between 0.2205 and 0.801 mM when the sulfate level is 25 mM.

Summary

A simple linear regression should only be performed if you observe visually that there is a linear pattern in the scatterplot and that there is a statistically significant correlation between the independent and dependent variables. Use technology to find the numeric values for the y-intercept = $a = b_0$ and slope = $b = b_1$, then make sure to use the correct notation when substituting your numbers back in the regression equation $\hat{y} = b_0 + b_1x$. Another measure of how well the line fits the data is called the coefficient of determination R^2 . When R^2 is close to 1 (or 100%) then the line fits the data very closely. The advantage over using R^2 over r is that we can use R^2 for nonlinear regression, whereas r is only for linear regression. Make sure that the residual plots have a completely random horizontal band around zero. There should be no patterns in the residual plots such as a sideways V that may indicate a non-constant variance. A pattern like a slanted line, a U, or an upside-down U shape would suggest a non-linear model. Check that the residuals are normally distributed; this is not the same as the population being normally distributed. Check to make sure that there are no outliers. Be careful with lurking and confounding variables.

Chapter 11 Formulas

$SS_{xx} = (n - 1)s_x^2$ $SS_{yy} = (n - 1)s_y^2$ $SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y}$	Correlation Coefficient $r = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}}$
Correlation t-test $H_0: \rho = 0$ $H_1: \rho \neq 0$ $t = r \sqrt{\frac{(n-2)}{1-r^2}}$ $df = n - 2$	Regression Equation (Line of Best Fit) $\hat{y} = b_0 + b_1x$ Slope $b_1 = \frac{SS_{xy}}{SS_{xx}}$ y-Intercept $b_0 = \bar{y} - b_1\bar{x}$
Slope t-test $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ $t = \frac{b_1}{\sqrt{\frac{MSE}{SS_{xx}}}}$ $df = n - p - 1 = n - 2$	Slope/Model F-test $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ See ANOVA table below
Standard Error of Estimate $s_{est} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{MSE}$	Residual $e_i = y_i - \hat{y}_i$
Prediction Interval $\hat{y} \pm t_{\frac{\alpha}{2}} \cdot s_{est} \sqrt{\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_{xx}}\right)}$	Coefficient of Determination $R^2 = (r)^2 = \frac{SSR}{SST}$

Source	SS = Sum of Squares	df	MS = Mean Square	F
Regression	$SSR = \frac{(SS_{xy})^2}{SS_{xx}}$	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error	$SSE = SS_{yy} - SSR$	$n - p - 1$	$MSE = \frac{SSE}{n - p - 1}$	
Total	$SST = SS_{yy}$	$n - 1$		

Chapter 11 Exercises

- To test the significance of the correlation coefficient, we use the t-distribution with how many degrees of freedom?
 - $n - 1$
 - n
 - $n + 1$
 - $n - 2$
 - $n_1 + n_2 - 2$
- What are the hypotheses for testing to see if a correlation is statistically significant?
 - $H_0: r = 0$ $H_1: r \neq 0$
 - $H_0: \rho = 0$ $H_1: \rho \neq 0$
 - $H_0: \rho = \pm 1$ $H_1: \rho \neq \pm 1$
 - $H_0: r = \pm 1$ $H_1: r \neq \pm 1$
 - $H_0: \rho = 0$ $H_1: \rho = 1$
- The coefficient of determination is a number between _____.
 - 1 and 1
 - 10 and 10
 - 0 and 10
 - 0 and ∞
 - 0 and 1
 - $-\infty$ and ∞
- The correlation coefficient is a number between _____.
 - 1 and 1
 - 10 and 10
 - 0 and 10
 - 0 and ∞
 - 0 and 1
 - $-\infty$ and ∞
- The standard error of estimate is a number between _____.
 - 1 and 1
 - 10 and 10
 - 0 and 10
 - 0 and ∞
 - 0 and 1
 - $-\infty$ and ∞
- The sum of the residuals should equal _____.
 - Slope
 - Y-intercept
 - 0
 - 1
 - 10
 - r
- True or False: Correlation measures the strength and direction of the linear relationship between two variables.

8. True or False: A correlation coefficient of -0.95 indicates a weak negative relationship between two variables.
9. True or False: Correlation implies causation, meaning that if two variables are strongly correlated, one variable causes the other.
10. True or False: In simple linear regression, the slope of the regression line represents the change in the dependent variable associated with a one-unit increase in the independent variable.
11. True or False: A p-value less than the significance level indicates that there is a significant linear relationship between the independent and dependent variables.
12. True or False: A correlation of 0 indicates that there is no relationship between the two variables.
13. True or False: In simple linear regression, the y-intercept represents the predicted value of the dependent variable when the independent variable is zero.
14. True or False: The coefficient of determination represents the proportion of the dependent variable's variance explained by the independent variable.
15. Which of the following is not the correct notation for a linear regression equation?
 - a) $\hat{y} = -5 + \frac{2}{9}x$
 - b) $\hat{y} = 3x + 2$
 - c) $\hat{y} = \frac{2}{9} - 5x$
 - d) $y = 5 + 0.4x$
16. What is the purpose of residual analysis in simple linear regression?
 - a) To check the assumptions of linearity.
 - b) To detect any outliers.
 - c) To check the assumptions of constant variance of errors.
 - d) All of the above.
17. It has long been thought that the length of one's femur is positively correlated to the length of one's tibia. The following are data for a classroom of students who measured each (approximately) in inches. Use $\alpha = 0.10$ to test to see if there is a significant correlation between the two variables.

Femur Length	Tibia Length
18.7	14.2
20.5	15.9
16.2	13.1
15.0	12.4
19.0	16.2
21.3	15.8
21.0	16.2
14.3	12.1
15.8	13.0
18.8	14.3
18.7	13.8

18. A new fad diet called Trim-to-the-MAX is running some tests that they can use in advertisements. They take a random sample of their users and record the number of days each has been on the diet along with how much

weight they have lost in pounds. The data is below. Use $\alpha = 0.05$ to test to see if there is a significant correlation between the two variables.

Days on Diet	7	12	16	19	25	34	39	43	44	49
Weight Lost	5	7	12	15	20	25	24	29	33	35

19. An elementary school uses the same system to test math skills at their school throughout the course of the 5 grades at their school. The age and score (out of 100) of several students is displayed below. Use $\alpha = 0.10$ to test to see if there is a significant correlation between the two variables.

Student Age	6	6	7	8	8	9	10	11	11
Math Score	54	42	50	61	67	65	71	72	79

20. The following data represent the age of a car and the average monthly cost for repairs. Use $\alpha = 0.05$ to test to see if there is a significant correlation between the two variables.

Age of Car (yrs.)	1	2	3	4	5	6	7	8	9	10
Monthly Cost (\$)	25	34	42	45	55	71	82	88	87	90

21. Body frame size is determined by a person's wrist circumference in relation to height. A researcher measures the wrist circumference and height of a random sample of individuals. The data is displayed below. Use the Excel output and $\alpha = 0.05$ to answer the following questions.

<i>Regression Statistics</i>	
Multiple R	0.7938
R Square	0.6301
Adjusted R Square	0.6182
Standard Error	3.8648
Observations	33

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	31.6304	5.2538	6.0205	1.16E-06
Circumference	5.4496	0.7499	7.2673	3.55E-08

- a) What is the value of the test statistic to see if the correlation is statistically significant?
- 6.0205
 - 1.16E-06
 - 3.55E-08
 - 5.2538
 - 7.2673
 - 0.7938
 - 0.7499
- b) What is the correct p-value and conclusion for testing if there is a significant correlation?
- 1.16E-06; There is a significant correlation.
 - 3.55E-08, There is a significant correlation.
 - 1.16E-06; There is not a significant correlation.
 - 3.55E-08, There is not a significant correlation.
 - 0.7938, There is a significant correlation.
 - 0.7938, There is not a significant correlation.

- c) Which number is the standard error of estimate?
- d) Which number is the coefficient of determination?
- e) Compute the correlation coefficient.
- f) What is the correct test statistic for testing if the slope is significant $H_1: \beta_1 \neq 0$?
- g) What is the correct p-value for testing if the slope is significant $H_1: \beta_1 \neq 0$?
- h) At the 5% level of significance, is there a significant linear relationship between wrist circumference and height?

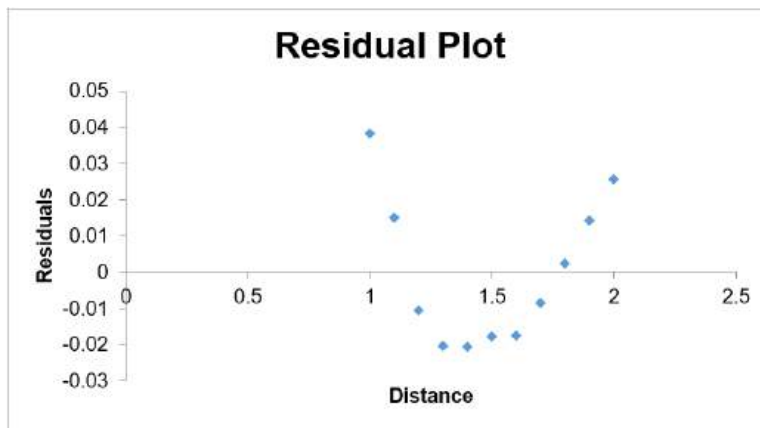
22. A survey asked participants at a local gym their age (x) and the number of hours they exercised (y) each week. The following regression equation was found $\hat{y} = 35.7 - 1.25x$. Choose the correct interpretation of the slope coefficient.

- a) The average age at the gym was 35.7 years.
- b) For every hour of exercise, participants age decreased by 1.25 years.
- c) Every year older a person gets; their predicted hours of exercise decrease by 1.25 hours per week.
- d) Every year older a person gets; their predicted hours of exercise increase by 1.25 hours per week.
- e) For every 35.7 hours exercised each week a person was 1.25 years younger.

23. The intensity (in candelas) of a 100-watt light bulb was measured by a sensing device at various distances (in meters) from the light source. A linear regression was run and the following residual plot was found.

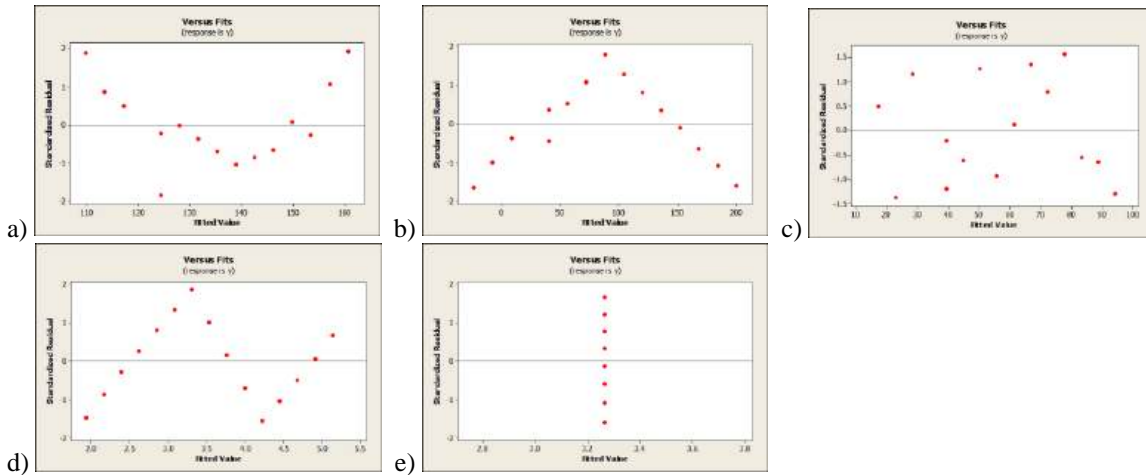
<i>Regression Statistics</i>	
Multiple R	0.95936
R Square	0.920371
Adjusted R Square	0.911523
Standard Error	0.021636
Observations	11

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	0.468618	0.031624	14.81856	1.25E-07
Distance	-0.2104	0.020629	-10.1992	3.04E-06

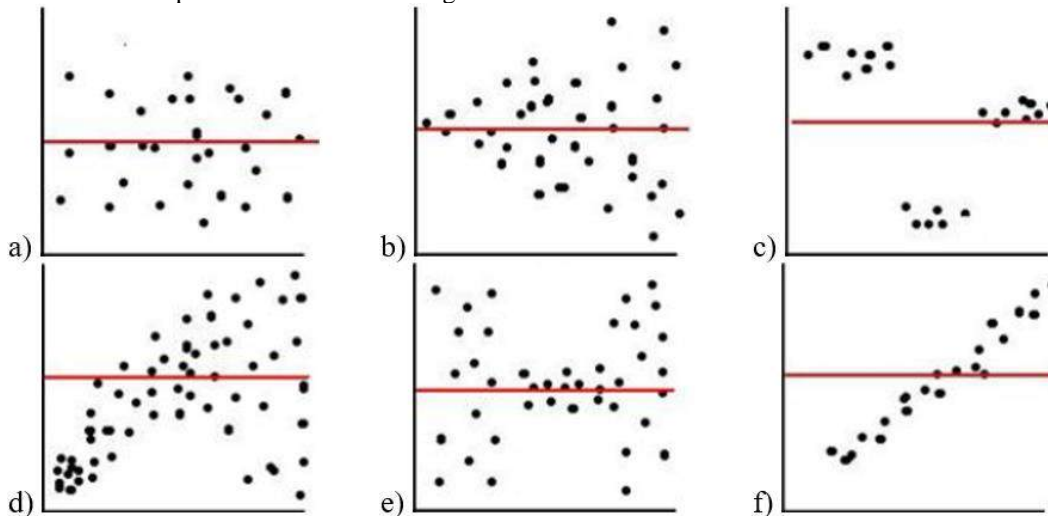


- a) Is linear regression a good model to use?
- b) Write a sentence explaining your answer.

24. Which residual plot has the best linear regression model?



25. Which residual plot has the best linear regression model?



26. An object is thrown from the top of a building. The following data measure the height of the object from the ground for a five-second period.

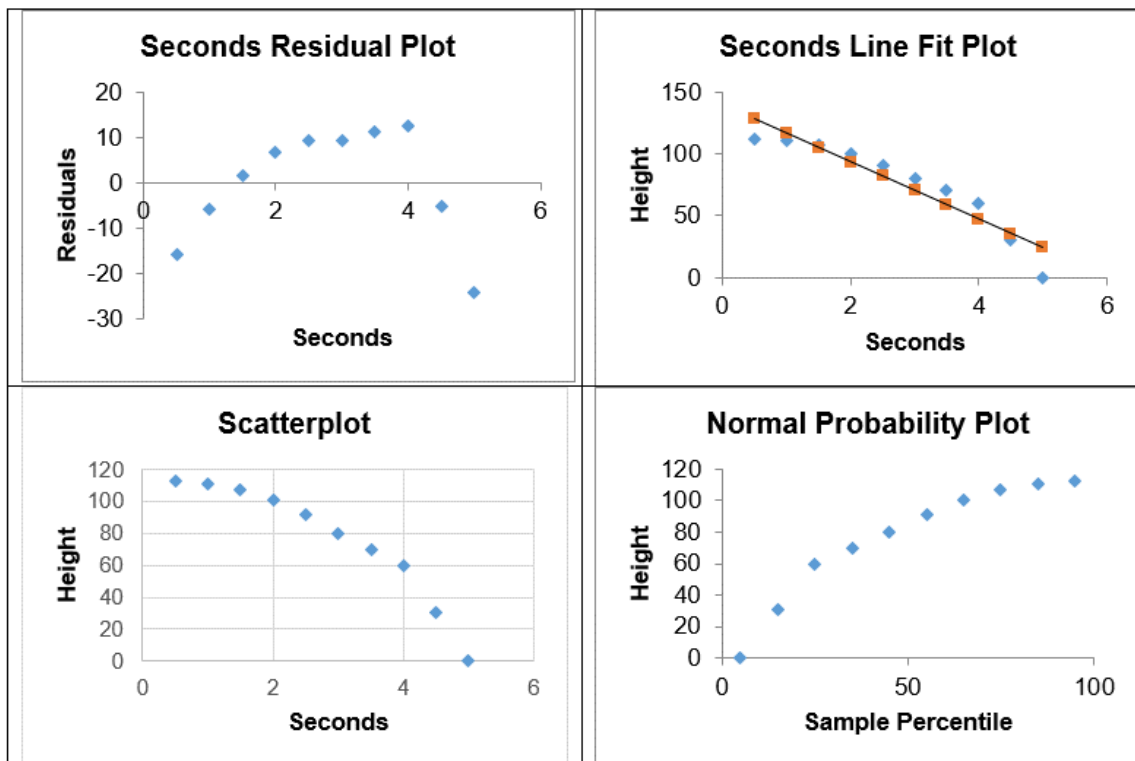
Seconds	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Height	112.5	110.875	106.8	100.275	91.3	79.875	70.083	59.83	30.65	0

- a) State the hypotheses to test for a significant correlation.
- b) Compute the correlation coefficient.
- c) Compute the p-value to see if there is a significant correlation.
- d) State the correct decision.
- e) Is there a significant correlation?
- f) Compute the regression equation.

27. An object is thrown from the top of a building. The following data measure the height of the object from the ground for a five-second period.

Seconds	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Height	112.5	110.875	106.8	100.275	91.3	79.875	70.083	59.83	30.65	0

The following four plots were part of the regression analysis.

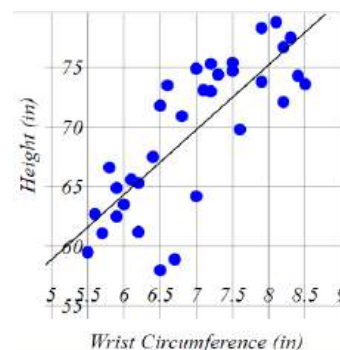


There is a statistically significant correlation between time and height, $r = -0.942$, $p\text{-value} = 0.0000454$. Should linear regression be used for this data? Why or why not? Choose the correct answer.

- Yes, the p -value indicates that there is a significant correlation so we can use linear regression.
- Yes, the normal probability plot has a nice curve to it.
- Yes, there is a nice straight line in the line fit plot.
- No, there is a curve in the residual plot, normal plot and the scatterplot.

28. Body frame size is correlated with a person's wrist circumference in relation to height. A researcher measures the wrist circumference and height of a random sample of individuals. The Excel output and scatterplot are displayed below. Find the regression equation and predict the height (in inches) for a person with a wrist circumference of 7 inches. Then, compute the residual for the point (7, 75).

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	31.6304	5.2538	6.0205	1.16E-06
Circumference	5.4496	0.7499	7.2673	3.55E-08



29. The data below show the predicted average high temperature ($^{\circ}\text{F}$) per month by the Farmer's Almanac in Portland, Oregon alongside the actual high temperature per month that occurred.

Farmer's Almanac	45	50	57	62	69	72	81	90	78	64	51	48
Actual High	46	52	60	61	72	78	82	95	85	68	52	49

- Compute the regression equation.
- Test to see if the slope is significantly different from zero, use $\alpha = 0.01$.
- Predict the high temperature in the coming year, given that the Farmer's Almanac is predicting the high to be 58 °F.
- Compute the 99% prediction interval for the actual high temperature in the coming year, given that the Farmer's Almanac is predicting the high to be 58 °F.

30. The following data represent the leaching rates (percent of lead extracted vs. time in minutes) for lead in solutions of magnesium chloride (MgCl_2). Use $\alpha = 0.05$.

Time (x)	4	8	16	30	60	120
Percent Extracted (y)	1.2	1.6	2.3	2.8	3.6	4.4

- State the hypotheses to test for a significant linear relationship.
- Compute the correlation coefficient.
- Compute the p-value to see if there is a significant linear relationship.
- State the correct decision.
- Is there a significant linear relationship?
- Compute the coefficient of determination.
- Compute the regression equation.
- Compute the 95% prediction interval for 100 minutes.
- Write a sentence interpreting this interval using units and context.

31. Bone mineral density and cola consumption have been recorded for a sample of patients. Let x represent the number of colas consumed per week and y the bone mineral density in grams per cubic centimeter. Assume the data is normally distributed.

x	y
1	0.883
2	0.8734
3	0.8898
4	0.8852
5	0.8816
6	0.863
7	0.8634
8	0.8648
9	0.8552
10	0.8546
11	0.862

- State the hypotheses to test for a significant linear relationship.
- Compute the correlation coefficient.
- Compute the test statistic and p-value to see if there is a significant linear relationship.
- State the correct decision, use $\alpha = 0.05$. Is there a significant linear relationship?
- Compute the coefficient of determination.
- Compute the regression equation.
- Interpret the slope coefficient.
- Compute the predicted bone mineral density for a person that consumes 7 colas per week.
- Compute the residual for the point (7, 0.8634).

32. A new fad diet called Trim-to-the-MAX is running some tests that they can use in advertisements. They sample 25 of their users and record the number of days each has been on the diet along with how much weight they have lost in pounds. The data is below. A significant linear correlation was found between the two variables. Compute the 95% prediction interval for the weight lost when a person has been on the diet for 60 days.

Days on Diet	7	12	16	19	25	34	39	43	44	49
Weight Lost	5	7	12	15	20	25	24	29	33	35

33. An elementary school uses the same system to test math skills at their school throughout the course of the 5 grades at their school. The age and score (out of 100) of several students is displayed below. A significant linear relationship is found between the student's age and their math score. Compute a 90% prediction interval for the score a student would earn given that they are 5 years old.

Student Age	6	6	7	8	8	9	10	11	11
Math Score	54	42	50	61	67	65	71	72	79

34. A teacher believes that the third homework assignment is a key predictor of how well students will do on the midterm. Let x represent the third homework score and y the midterm exam score. A random sample of last term's students were selected and their grades are shown below. Assume scores are normally distributed. Use $\alpha = 0.05$.

HW3	Midterm
13.1	59
21.9	87
8.8	53
24.3	95
5.4	39
13.2	66
20.9	89
18.5	78

HW3	Midterm
6.4	43
20.2	79
21.8	84
23.1	92
22	87
11.4	54
14.9	71
18.4	76

HW3	Midterm
20	86
15.4	73
25	93
9.7	52
15.1	70
15	65
16.8	77
20.1	78

- State the hypotheses to test for a significant correlation.
 - Compute the correlation coefficient.
 - Compute the p-value to see if there is a significant correlation.
 - State the correct decision.
 - Is there a significant correlation?
 - Compute the coefficient of determination.
 - Write a sentence interpreting R^2 .
 - Does doing poorly on homework 3 cause a student to do poorly on the midterm exam? Explain.
 - Compute the standard error of estimate.
 - Compute the regression equation.
 - Compute the predicted midterm score when the homework 3 score is 15.
 - Compute the residual for the point (15, 65).
 - Compute the 95% prediction interval for the midterm score when the homework 3 score is 15.
35. A study was conducted to determine if there was a linear relationship between a person's age and their peak heart rate. Use $\alpha = 0.05$.

Age (x)	16	26	32	37	42	53	48	21
Peak Heart Rate (y)	220	194	193	178	172	160	174	214

- What is the estimated regression equation that relates number of hours worked and test scores for high school students?
- Interpret the slope coefficient for this problem.
- Compute and interpret the coefficient of determination.
- Compute the coefficient of nondetermination.
- Compute the standard error of estimate.
- Compute the correlation coefficient.
- Compute the 95% Prediction Interval for peak heart rate for someone who is 25 years old.

36. The following data represent the weight of a person riding a bike and the rolling distance achieved after going down a hill without pedaling.

Weight (lbs.)	59	84	97	56	103	87	88	92	53	66	71	100
Rolling Distance (m.)	26	43	48	20	59	44	48	46	28	32	39	49

- a) Can it be concluded at a 0.05 level of significance that there is a linear correlation between the two variables?
- b) Using the regression line for this problem, find the predicted bike rolling distance for a person that weighs 110 lbs.
- c) Find the 99% prediction interval for bike rolling distance for a person that weighs 110 lbs.
37. It has long been thought that the length of one's femur is positively correlated to the length of one's tibia. The following are data for a classroom of students who measured each (approximately) in inches. A significant linear correlation was found between the two variables. Find the 90% prediction interval for the length of someone's tibia when it is known that their femur is 23 inches long.

Femur Length	Tibia Length
18.7	14.2
20.5	15.9
16.2	13.1
15.0	12.4
19.0	16.2
21.3	15.8
21.0	16.2
14.3	12.1
15.8	13.0
18.8	14.3
18.7	13.8

38. The following data represent the age of a car and the average monthly cost for repairs. A significant linear correlation is found between the two variables. Use the data to find a 95% prediction interval for the monthly cost of repairs for a vehicle that is 15 years old.

Age of Car (yrs.)	1	2	3	4	5	6	7	8	9	10
Monthly Cost (\$)	25	34	42	45	55	71	82	88	87	90

39. The following data represent the enrollment at a small college during its first ten years of existence. A significant linear relationship is found between the two variables. Compute a 90% prediction interval for the enrollment after the college has been open for 14 years.

Years	1	2	3	4	5	6	7	8	9	10
Enrollment	856	842	923	956	940	981	1025	996	1057	1088

40. A nutritionist feels that what mothers eat during the months they are nursing their babies is important for healthy weight gain of their babies. She samples several of her clients and records their average daily caloric intake for the first three months of their babies' lives and also records the amount of weight the babies gained in those three months. The data are below.

Daily Calories	1523	1649	1677	1780	1852	2065	2096	2145	2378
Baby's Weight Gain (lbs.)	4.62	4.77	4.62	5.12	5.81	5.34	5.89	5.96	6.05

- a) Compute the regression equation.

- b) Test to see if the slope is significantly different from zero, use $\alpha = 0.05$.
- c) Predict the weight gain of a baby whose mother gets 2,500 calories per day.
- d) Compute the 95% prediction interval for the weight gain of a baby whose mother gets 2,500 calories per day.

41. The data below represent the driving speed (mph) of a vehicle and the corresponding gas mileage (mpg) for several recorded instances.

Driving Speed	Gas Mileage
57	21.8
66	20.9
42	25.0
34	26.2
44	24.3
44	26.3
25	26.1
20	27.2
24	23.5
42	22.6
52	19.4
54	23.9
60	24.8
62	21.5
66	20.5
67	23.0
52	24.2
49	25.3
48	24.3
41	28.4
38	29.6
26	32.5
24	30.8
21	28.8
19	33.5
24	25.1

- a) Do a hypothesis test to see if there is a significant correlation. Use $\alpha = 0.10$.
- b) Compute the standard error of estimate.
- c) Compute the regression equation and use it to find the predicted gas mileage when a vehicle is driving at 77 mph.
- d) Compute the 90% prediction interval for gas mileage when a vehicle is driving at 77 mph.

42. In a sample of 20 football players for a college team, their weight and 40-yard-dash time in minutes were recorded.

Weight (lbs.)	40-Yard-Dash
285	5.95
185	4.99
165	4.92
188	4.77
160	4.52
156	4.67
256	5.22
169	4.95
210	5.06
165	4.83

Weight (lbs.)	40-Yard-Dash
195	4.85
254	5.12
140	4.87
212	5.05
158	4.75
188	4.87
134	4.53
205	4.92
178	4.88
159	4.79

- Do a hypothesis test to see if there is a significant correlation. Use $\alpha = 0.01$.
- Compute the standard error of estimate.
- Compute the regression equation and use it to find the predicted 40-yard-dash time for a football player that is 200 lbs.
- Compute the 99% prediction interval for a football player that is 200 lbs.
- Write a sentence interpreting the prediction interval.

“Well I was in fact, I was moving backwards in time. Hmmm. Well I think we've sorted all that out now. If you'd like to know, I can tell you that in your universe you move freely in three dimensions that you call space. You move in a straight line in a fourth, which you call time, and stay rooted to one place in a fifth, which is the first fundamental of probability. After that it gets a bit complicated, and there's all sorts of stuff going on in dimensions 13 to 22 that you really wouldn't want to know about. All you really need to know for the moment is that the universe is a lot more complicated than you might think...”
 (Adams, 2002)

Answers to Odd Numbered Exercises

Chapter 1

- 1) 42
- 3) c
- 5) b, c, e
- 7) a) Population Parameter
b) Observation
c) Sample Statistic
d) Variable
- 9) a) Population Parameter
b) Sample Statistic
c) Individual
d) Variable
- 11) Stratified
- 13) Convenience
- 15) Systematic
- 17) a) Experimental
b) Observational
c) Experimental
d) Experimental
- 19) a, b, d, f, g
- 21) a) Ratio
b) Nominal
c) Ordinal
d) Ratio
e) Ratio
f) Ratio
g) Nominal
- 23) a) Quantitative
b) Qualitative
c) Qualitative
d) Qualitative
e) Qualitative
- 25) a) Continuous
b) Discrete
c) Continuous
d) Continuous
e) Continuous
- 27) a) Systematic
b) Convenience
c) Random
d) Cluster
- 29) a) Random
b) Systematic
c) Stratified
d) Cluster
- 31) a) Cluster
b) Systematic
c) Convenience
d) Random
- 33) b
- 35) b
- 37) b
- 39) a

41) b

Chapter 2

- 1) a, c, d
- 3) True
- 5) a)

```

6 | 3 5 6 9
7 | 0 1 3 5 6 9
8 | 0 1 2 2 3 3 3 3 5 5 6 7
9 | 4 6 9
    
```

b)

Class	Freq.	Cum. Freq.	Rel. Freq.	Cum. Rel. Freq.
60-69	4	4	$4/25 = 0.16$	$4/25 = 0.16$
70-79	6	$4+6 = 10$	$6/25 = 0.24$	$10/25 = 0.4$
80-89	12	$10+12 = 22$	$12/25 = 0.48$	$22/25 = 0.88$
90-99	3	$22+3 = 25$	$3/25 = 0.12$	$25/25 = 1$
Total	25		1	

- c) 1
- d) The sample size n .
- e) 6
- f) 0.24
- g) 80-89
- h) 90-99
- i) 0.88
- j) 70-79
- 7) a) 40
b) 50%

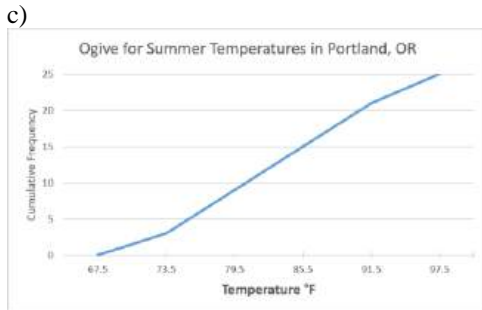
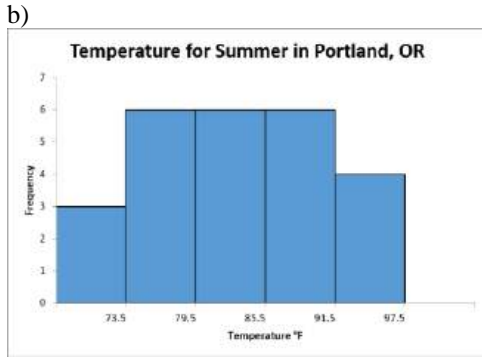
9)

```

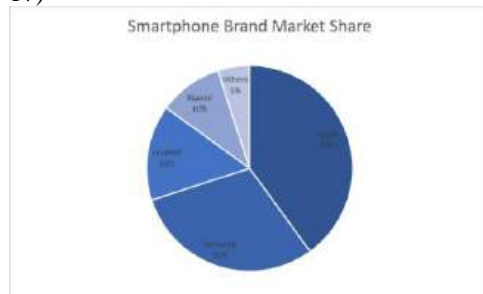
15 | 4
16 | 2 5 7 9
17 | 0 1 2 3 5 5 6 7 8 8 8
18 | 2 4 5 5 5 6 8
19 |
20 | 1
    
```

11) a)

Class	Frequency	Cumulative Frequency
68 - 73	3	3
74 - 79	6	9
80 - 85	6	15
86 - 91	6	21
92 - 97	4	25

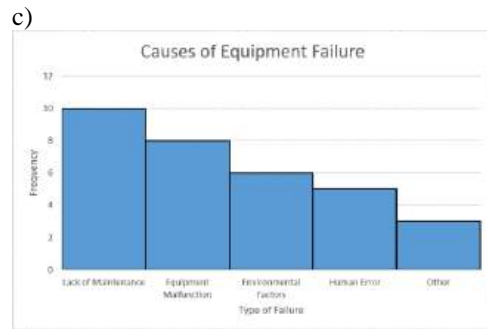
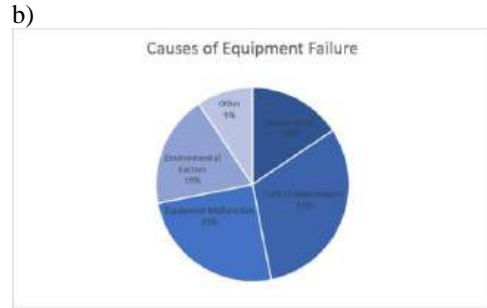


- 13) a) 50
 b) 78
 c) 16.55
 15) 20%
 17)



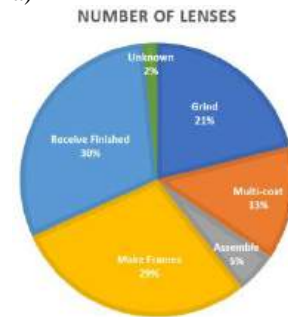
19) a)

Factor	Number of Failures	Relative Frequency
Human Error	5	0.15625
Lack of Maintenance	10	0.3125
Equipment Malfunction	8	0.25
Environmental Factors	6	0.1875
Other	3	0.09375
Total	32	1

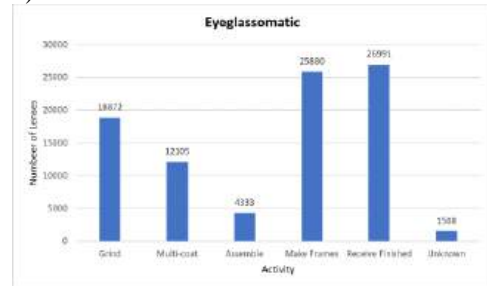


d) Lack of Maintenance

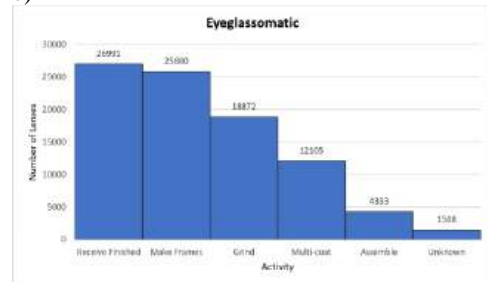
21) a)



b)



c)



23) a)



b)

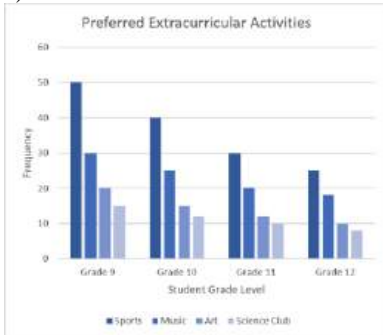


c)

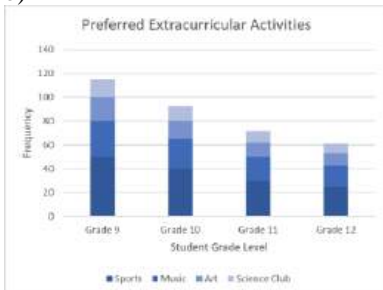


25) Chevy & Toyota

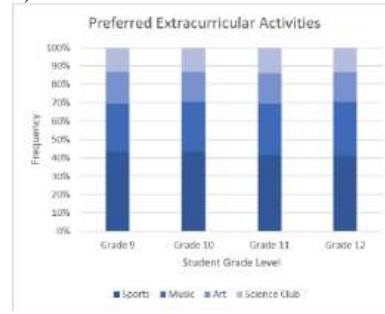
27) a)



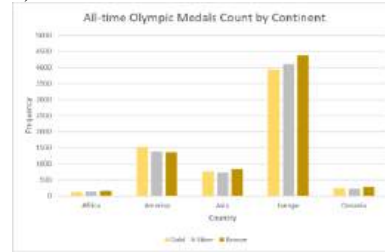
b)



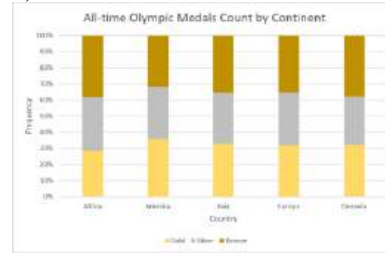
c)



29) a)



b)

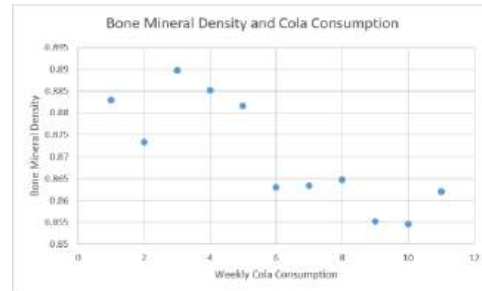


31) 1.5

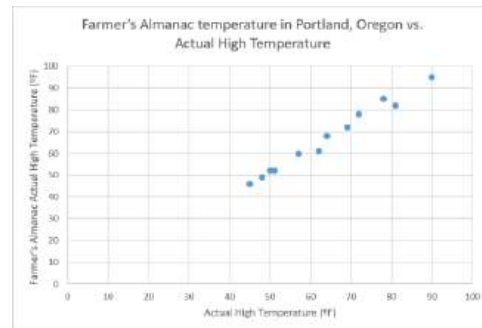
33) a) Poland

b) Greece

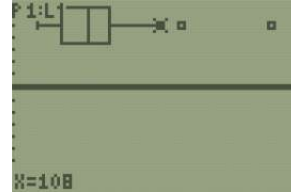
35)



37)

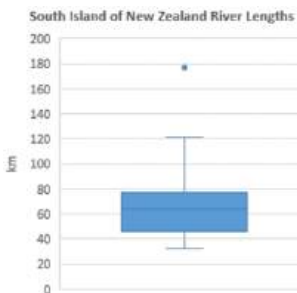


- 39) There are no labels for both axis or categories.
 41) The vertical axis is reversed, making the graph appear to increase when it is actually decreasing. There are no labels for both axes.



Chapter 3

- 1) a) mode = 4
 b) median = 3.75
 c) $\bar{x} = 3.725$
 3) a) 56 & 64
 b) 64
 c) 67.6818
 5) 2.833
 7) 72.25
 9) a) Range = 0.9
 b) $s^2 = 0.0870$
 c) $s = 0.2949$
 11) a) Negatively skewed
 b) The median is higher.
 13) $s_1 = 9.9348$, $s_2 = 3.3466$; Training 1 is more variable
 15) Before range = 42, after range = 42, both groups have the same range.
 17) a) 30.35
 b) 136.7
 c) 32.7
 d) 38.52%
 e) 28.4
 19) a) $CV_{\text{height}} = 3.04\%$; $CV_{\text{weight}} = 9.84\%$
 b) Weight, because it has a higher coefficient of variation.
 21) -10
 23) a) -2.8507
 b) -3.1808
 c) Tennessee Lane
 25) Min = 3.2, $Q_1 = 3.45$ (TI: 3.55), $Q_2 = 3.7$, $Q_3 = 3.95$ (TI: 3.85), Max = 4.1
 27) a) Min = 32, $Q_1 = 46$ (TI: 48), $Q_2 = 64$, $Q_3 = 77$ (TI: 76), Max = 177
 b) lower limit = -0.5 (TI: 6), upper limit = 123.5 (TI: 118), outliers = 177 (TI: 121 & 177)
 c)



- 29) a) 4
 b) Yes, the treatment was effective.
 31) b, a, c
 33) 1. c. ii; 2. a. iii; 3. b. i
 33) 99.7%
 35) a
 37) 0.8241
 39) a) 0.9815
 b) $\hat{y} = 25.6472 + 2.8212x$
 c) 67.96566
 41) a) 0.9403
 b) $\hat{y} = 1.6307 + 0.0257x$
 c) 4.2003

Chapter 4

- 1) 0.1420
 3) 0.375
 5) 0.001
 7) 0.05
 9) 0.1944
 11) 0.2
 13) 0.513
 15) 0.19
 17) a) 0.25
 b) 0.0192
 c) 0.3077
 d) 0.1538
 e) 0
 f) Yes, $P(7 \cap \text{Ace}) = 0$
 19) a) 0.5627
 b) 0.1756
 c) 0.1868
 d) 0.3579
 e) 2034
 21) a) 0.8696
 b) 0.2754
 c) 0.5938
 d) 0.7826
 e) 0.4638
 f) 0.5909
 23) a)

	Inoculated	Not Inoculated	Total
Lived	0.0382	0.8252	0.8634
Died	0.0010	0.1356	0.1366
Total	0.0392	0.9608	1

- b) 0.0392
 c) 0.8643
 d) 0.1748

- e) 0.026
- f) 0.141
- 25) 0.2222
- 27) 0.42
- 29) a) 0.3271
- b) 0.0062
- c) 0.3427
- d) 0.6729
- e) 0.2170
- f) 0.7451
- g) 0.6243
- 31) 8,000,000
- 33) 175,760,000
- 35) 24,360
- 37) 60,949,324,800
- 39) 2,821,109,907,456
- 41) 125,000

Chapter 5

- 1) a) Yes
- b) No
- c) No
- 3) a) -\$0.425
- b) \$2.9592
- 5) \$69.283
- 7) a)

x	-112.1	887.9
$P(X = x)$	0.76	0.24
- b) \$127.90
- c) For many of these extended warranties bought by customers, they can expect to gain 127.9 dollars per warranty on average.
- 9) a) Yes, $\sum P(x) = 1$ and $0 \leq P(x) \leq 1$
- b) 0.8
- c) 0.9055
- d) 0.82
- 11) a) 0.0016
- b) 0.0021
- c) 0.0005
- d) 0.9979
- e) 0.9995
- f) 21.25
- g) 1.7854
- h) 3.1875
- 13) a) 0.1222
- b) 0.9464
- c) 0.1759
- d) 0.9873
- 15) a) 0.0428
- b) 0.9444
- c) 0.9873
- d) 0.0556
- e) 0.0127
- f) 1.5
- g) 1.1619
- h) 1.35

- 17) 0.995
- 19) 0.0002
- 21) 0.3438
- 23) a) 0.2193
- b) 0.0191
- c) 0.9801
- d) 0.0199
- 25) a) 0.6826
- b) 0.087
- 27) 99.7%
- 29) 2.35%
- 31) a) 0.9803
- b) 0.5809
- c) 0.9429
- d) 0.0455
- e) 0.0024
- f) 0.0162
- g) 0.0227
- h) 0.305
- 33) a) 0.0107
- b) 0.0571
- c) 0.9
- d) 0.9772
- e) 0.1227
- f) 0.9418
- 35) a) -0.5534
- b) 0.6745
- c) ± 2.5758
- 37) a) 0.1333
- b) 0.6585
- c) 80.4686 in.
- 39) a) 0.0548
- b) 0.0541
- c) 11.1569 years
- 41) a) 0.2179
- b) 0.3034
- c) 0.5853
- d) 62.7477 in.
- 43) 0.2215
- 45) a) 1023
- b) 19.3
- c) 26.1
- 47) 0.0146
- 49) 0.1333
- 51) e
- 53) a) ii.
- b) iii.
- c) i.
- 55) a) iii.
- b) i.
- c) ii.
- 57) a) 0.0036
- b) 0.0718
- c) 0.9213
- d) 0.9696
- 59) a) 0.9922

- b) 0.0399
 - c) 0.0026
 - d) 0.9088
- 61) ± 2.7633

Chapter 6

- 1) a
- 3) ± 1.6449
- 5) d
- 7) $0.6101 < p < 0.7399$
- 9) $0.4710 < p < 0.749$
- 11) $0.3226 < p < 0.3574$
- 13) $0.1658 < p < 0.1742$
- 15) b
- 17) $t = \pm 1.6991$
- 19) $t = \pm 2.023$
- 21) $88.872 < \mu < 97.128$
- 23) $12.6989 < \mu < 14.7011$
- 25) $3.3073 < \mu < 3.7527$; We can be 99% confident that the population mean GPA for all graduate business students admitted into the top schools is between 3.3073 and 3.7527.
- 27) $35.13195 < \mu < 50.86805$
- 29) $12.3471 < \mu < 19.2529$; We can be 99% confident that the population mean germination time for the new bean is between 12.3 and 19.3 days.
- 25) $t = \pm 2.807$
- 27) $17795.7087 < \mu < 18684.2913$; We can be 95% confident that the population mean amount of money in all money market accounts is between \$17,795.71 and \$18,684.29.
- 29) $12.3471 < \mu < 19.2529$; We can be 99% confident that the population mean germination time for the new bean is between 12.3 and 19.3 days.
- 31) a) 2.55
b) ii.
c) $2.0923 < \mu < 3.0077$
d) i.
e) It is possible, but unlikely since 2 is not contained within the confidence interval boundaries
- 33) $26.7589 < \mu < 33.6411$
- 35) d
- 37) 205
- 39) 82
- 41) 554

Chapter 7

- 1) b
- 3) a
- 5) ± 2.5758
- 7) ± 1.96

- 9) a) ii.
b) viii.
c) iv.
d) vi.
e) vii.
f) i.
g) v.
h) iii.
- 11) $100(1 - \alpha)\%$ = Confidence Level
 $1 - \beta$ = Power
 β = P(Type II Error)
 μ = Parameter
 α = Significance Level
- 13) a) $H_0: \mu = 1000; H_1: \mu \neq 1000$
b) The test writers conclude that the average test score is not 1000 when it really was. They would need to change the exam when they really did not need to.
c) The test writers conclude that average test score is 1000 points when it really was not. Students taking the exam would have a higher or lower mean than the previous years. This could go either way, the students could have a harder test, score lower and hence not get into their choice of colleges. Or, on the flip side, students could have a higher mean score than previous years and get an unfair edge into colleges when they are not necessarily prepared.
- 15) d
- 17) b
- 19) a
- 21) The implication of a Type I error from the clinical trial is that the vaccination will be approved when it indeed does not reduce the risk of contracting the virus. The implication of a Type II error from the clinical trial is that the vaccination will not be approved when it indeed does reduce the risk of contracting the virus.
- 23) The local government decides that the data do not provide convincing evidence of an average commute time higher than 30 minutes, when the true average commute time is in fact higher than 30 minutes.
- 25) 0.0123
- 27) $H_0: p = 0.093; H_1: p > 0.093; z = 2.7116; p\text{-value} = 0.0027$; Reject H_0 . There is enough evidence to support the claim the population proportion of American adults that suffer from depression or a depressive illness is more than 9.3%.
- 29) $H_0: p = 0.31; H_1: p < 0.31; z = -1.5831; p\text{-value} = 0.0567$; Do not reject H_0 . There is not enough evidence to support the claim the population proportion of men over the age

- of 50 who regularly have their prostate examined is significantly less than 0.31.
- 31) $H_0: p = 0.64; H_1: p > 0.64; z = 1.4907; p\text{-value} = 0.0680; \text{Reject } H_0$. There is enough evidence to support the claim the population proportion of adults that have their teeth cleaned by a dentist yearly is higher than 64%.
- 33) a) iii.
b) i.
- 35) a
- 37) $H_0: \mu = 36; H_1: \mu < 36; t = -1.9758; p\text{-value} = 0.0438; \text{Reject } H_0$. There is enough evidence to support the claim the average weekly household garbage weight is less than the company's weekly 36 lb. maximum.
- 39) $H_0: \mu = 842; H_1: \mu > 842; t = 0.8218; p\text{-value} = 0.2152; \text{Do not reject } H_0$. We do not have evidence to support the claim the average calories from a fast food meal is higher than reported.
- 41) $H_0: \mu = 98.6; H_1: \mu > 98.6; z = 2.1639; p\text{-value} = 0.018; \text{Do not reject } H_0$. There is not enough evidence to support the claim that the mean body temperature for all athletes is more than 98.6°F.
- Chapter 8
- 1) $H_0: p_1 = p_2; H_1: p_1 > p_2; z = 1.1104; p\text{-value} = 0.1334; \text{Do not reject } H_0$. There is not enough evidence to support the claim that the proportion of all freshman that purchase most of their textbooks from the college's bookstore is greater than the proportion of all seniors.
- 3) $H_0: p_1 = p_2; H_1: p_1 > p_2; z = 3.7177; p\text{-value} = 0.0001; \text{Reject } H_0$. Yes, there is evidence that the proportion of those who caught pertussis is higher for those who were not up to date on their booster.
- 5) $H_0: p_1 = p_2; H_1: p_1 < p_2; z = -1.8064; p\text{-value} = 0.0354; \text{Do not reject } H_0$. There is not enough evidence that the new facial recognition tool is performing better.
- 7) $0.04126 < p_1 - p_2 < 0.25874$
- 9) $H_0: \mu_D = 0; H_1: \mu_D < 0; t = -0.7514; p\text{-value} = 0.2311; \text{Do not reject } H_0$. There is not enough evidence to support the claim on average the new online learning module increased placement scores.
- 11) a) $-11.9129 < \mu_D < -8.4871$
b) Yes, since $\mu_D = 0$ is not captured in the interval $(-11.9129, -8.4871)$.
- 13) $H_0: \mu_D = 0; H_1: \mu_D > 0; t = 4.7951; p\text{-value} = 0.0001; \text{Reject } H_0$. There is enough evidence to support the claim that the mean reaction time is significantly faster for a person's dominant hand.
- 15) $H_0: \mu_1 = \mu_2; H_1: \mu_1 < \mu_2; t = -3.0908; p\text{-value} = 0.0011; \text{Reject } H_0$. There is enough evidence to support the claim that the online homework system for introductory math courses improved student's average test scores.
- 17) $H_0: \mu_1 = \mu_2; H_1: \mu_1 > \mu_2; t = 0.4653; p\text{-value} = 0.3222; \text{Do not reject } H_0$. There is not enough evidence to support the claim that the American League team would score on average more runs for games in which the designated hitter was used.
- 19) $H_0: \mu_1 = \mu_2; H_1: \mu_1 > \mu_2; t = 2.6612; p\text{-value} = 0.0056; \text{Reject } H_0$. There is enough evidence to support the claim that the mean number of frozen pizzas sold during the winter months is more than during the summer months.
- 21) $H_0: \mu_1 = \mu_2; H_1: \mu_1 < \mu_2; t = -1.2639; p\text{-value} = 0.1098; \text{Do not reject } H_0$. There is not enough evidence to support the claim that the mean wrinkle recovery angle for Hylite is better than Permafresh.
- 23) $H_0: \mu_1 = \mu_2; H_1: \mu_1 > \mu_2; t = 2.9106; p\text{-value} = 0.0027; \text{Reject } H_0$. There is enough evidence to support the claim that new medicine is effective.
- 25) $H_0: \mu_1 = \mu_2; H_1: \mu_1 > \mu_2; t = 22.9197; p\text{-value} = 0.000; \text{Reject } H_0$. There is enough evidence to support the claim that women with preeclampsia have a higher mean blood pressure in the late stages of labor.
- 27) $H_0: \mu_1 = \mu_2; H_1: \mu_1 > \mu_2; t = 3.7017; p\text{-value} = 0.0001; \text{Reject } H_0$. There is enough evidence to support the claim that community college transfer students take longer to earn a bachelor's degree.
- 29) $H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2; -18.6812 < \mu_1 - \mu_2 < -1.4521; \text{Reject } H_0$. There is enough evidence to support the claim that there is a difference in the average gym usage of sophomores and senior college students.
- 31) $H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2; -4632.49 < \mu_1 - \mu_2 < 554.49; \text{Fail to reject } H_0$. There is not enough evidence to support the claim that the mean starting salary differs based on level of experience.
- 33) $H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2; t = -0.8052; p\text{-value} = 0.4275; \text{fail to reject } H_0$. There is not enough evidence to support the claim that there is a statistically significant difference

- in the mean comprehension score between text and visual illustrations.
- 35) $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$; $t = -2.0919$; p-value = 0.0402; Do not reject H_0 . There is not enough evidence to support the claim that there is a statistically significant difference in the mean productivity level between the two locations.
- 37) $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$; $t = 2.0435$; p-value = 0.0437; Reject H_0 . There is enough evidence to support the claim that the mean cost for a pair of shoes in the Midwest and the South are different.
- 39) $-19.3226 < \mu_1 - \mu_2 < -12.4974$
- 41) $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$; $1.6542 < \mu_1 - \mu_2 < 7.1458$; Reject H_0 . There is enough evidence to support the claim that there is a statistically significant difference in the mean scores on the problem-solving test for mathematics and computer science majors.

Chapter 9

- 1) d
- 3) b & c
- 5) 15.0863
- 7) 12.017
- 9) 6
- 11) 10
- 13) $H_0: p_1 = 0.6, p_2 = 0.25, p_3 = 0.15$; H_1 : At least one proportion is different. $\chi^2 = 86.5529$; p-value = $1.604E-19 = 0$; Reject H_0 . There is enough evidence to support the claim that the distribution is different than expected. There were more students than expected that would read the text online.
- 15) $H_0: p_1 = 0.301, p_2 = 0.176, p_3 = 0.125, p_4 = 0.097, p_5 = 0.079, p_6 = 0.067, p_7 = 0.058, p_8 = 0.051, p_9 = 0.046$; H_1 : At least one proportion is different. $\chi^2 = 11.8466$; CV = 15.5073; Do not reject H_0 . There is no evidence of tax fraud so law enforcement officials should not pursue the case.
- 17) $H_0: p_1 = 0.25, p_2 = 0.25, p_3 = 0.25, p_4 = 0.25$; H_1 : At least one proportion is different. $\chi^2 = 3.16$; p-value = 0.3676; Do not reject H_0 . There is not enough evidence to support the claim that preference among the four display designs.
- 19) $H_0: p_1 = 0.25, p_2 = 0.25, p_3 = 0.25, p_4 = 0.25$; H_1 : At least one proportion is different. $\chi^2 = 2.64$; p-value = 0.4505; Do not reject H_0 . There is not enough evidence to support the claim that car accidents are not equally distributed throughout the year.
- 21) $H_0: p_1 = 0.27, p_2 = 0.32, p_3 = 0.41$; H_1 : At least one proportion is different. $\chi^2 = 8.352$;

- p-value = 0.0154; Reject H_0 . There is enough evidence to support the claim that there a significant difference in where young adults lived in 2000 versus 2008. There are fewer young adults living at home than expected.
- 23) $H_0: p_1 = 0.727, p_2 = 0.197, p_3 = 0.048, p_4 = 0.028$; H_1 : At least one proportion is different. $\chi^2 = 20.0291$; p-value = 0.0002; Reject H_0 . There is enough evidence to support the claim that the distribution of Oregon residents is different now compared to 1990. There were more Oregonians in central Oregon than expected.
- 25) $H_0: p_1 = 0.1, p_2 = 0.15, p_3 = 0.34, p_4 = 0.4$; H_1 : At least one proportion is different. $\chi^2 = 1.9196$; p-value = 0.5893; Do not reject H_0 . There is not enough evidence to support the claim that the Mr. Mendoza's course evaluation scores are different compared to the rest of the History Department's evaluations.
- 27) $H_0: p_1 = 0.42, p_2 = 0.31, p_3 = 0.16, p_4 = 0.11$; H_1 : At least one proportion is different. $\chi^2 = 7.6986$; p-value = 0.05267; Do not reject H_0 . There is not enough evidence to support the claim that the patients who took the medicine healed at a different rate than these percentages.
- 29) c
- 31) d
- 33) H_0 : The proportion of customers who will go out to eat on New Year's Eve is independent of location. H_1 : The proportion of customers who will go out to eat on New Year's Eve is dependent on location. $\chi^2 = 2.2772$; p-value = 0.3203; Do not reject H_0 . There is not enough evidence to support the claim that the proportion of customers who will go out to eat on New Year's Eve is dependent on location.
- 35) H_0 : The math placement exam and where students are placed are independent. H_1 : The math placement exam and where students are placed are dependent. $\chi^2 = 0.1642$; p-value = 0.9212; Do not reject H_0 . There is not enough evidence to support the claim that the math placement exam and where students are placed are dependent.
- 37) H_0 : Department and LMS preference are independent. H_1 : Department and LMS preference are dependent. $\chi^2 = 24.7778$; p-value = 0.000056; Reject H_0 . There is enough evidence to support the claim that department and LMS preference are dependent.

- 39) H_0 : Type of electronic device and store branch are dependent. H_1 : Type of electronic device and store branch are dependent. $\chi^2 = 7.4224$; p-value = 0.8285; Do not reject H_0 . There is not enough evidence to support the claim that type of electronic device and store branch are dependent.
- 41) H_0 : The number of defective parts is independent on the machine that produced it. H_1 : The number of defective parts is dependent on the machine that produced it. $\chi^2 = 2.3536$; p-value = 0.5023; Do not reject H_0 . There is not enough evidence to support the claim that the number of defective parts is dependent on the machine that produced it.

Chapter 10

- 1) b
 3) a
 5) b
 7) d
 9) a
 11) 4.1591
 13) 0.3018
 15) $H_0: \mu_1 = \mu_2 = \mu_3$; H_1 : At least one mean is different. $F = 0.5902$; p-value = 0.5605. Do not reject H_0 . There is not enough evidence to support the claim that there is a difference in the mean per-pupil costs for private school tuition for three counties in the Portland, Oregon, metro area.
- 17) c
 19) False
 21) b
 23) a) $H_0: \mu_A = \mu_B = \mu_C$; H_1 : At least one mean is different.
 b) $F = 10.64046$
 c) $F_\alpha = 3.0781$
 d) Reject H_0 . There is enough evidence to support the claim that there is a difference in the mean price of the three types of plastic.
 e) $H_0: \mu_A = \mu_B$; $H_1: \mu_A \neq \mu_B$; p-value = 0; Reject H_0 . There is significant difference in price between plastics A and B. $H_0: \mu_A = \mu_C$; $H_1: \mu_A \neq \mu_C$; p-value = 1; Do not reject H_0 . There is not a significant difference in price between plastics A and C. $H_0: \mu_B = \mu_C$; $H_1: \mu_B \neq \mu_C$; p-value = 0.003; Reject H_0 . There is significant difference in price between plastics B and C.
- 25) $H_0: \mu_1 = \mu_2 = \mu_3$; H_1 : At least one mean is different. $F = 2.7121$; p-value = 0.0896; Reject H_0 . There is sufficient evidence to support the claim that course delivery type is a factor for the mean final exam scores.

- 27) $H_0: \mu_1 = \mu_2 = \mu_3$; H_1 : At least one mean is different. $F = 2.5459$; p-value = 0.0904; Fail to reject H_0 . There is not enough evidence to support the claim that there is a difference in the mean movie ticket prices by geographical regions.

- 29) $H_0: \mu_A = \mu_B = \mu_C$; H_1 : At least one mean is different. $F = 2.895$; p-value = 0.06; Reject H_0 . There is sufficient evidence to support the claim that there is a difference in the mean cost between three different types of fabric. $H_0: \mu_A = \mu_B$; $H_1: \mu_A \neq \mu_B$; p-value = 1; Do not reject H_0 . There is not a significant difference in the mean cost of fabrics A and B. $H_0: \mu_A = \mu_C$; $H_1: \mu_A \neq \mu_C$; p-value = 0.222; Do not reject H_0 . There is not a significant difference in the mean cost of fabrics A and C. $H_0: \mu_B = \mu_C$; $H_1: \mu_B \neq \mu_C$; p-value = 0.07; Reject H_0 . There is significant difference in the mean cost of fabrics B and C.

- 31) $H_0: \mu_1 = \mu_2 = \mu_3$; H_1 : At least one mean is different. $F = 49.8126$; p-value = 0.000; Reject H_0 . There is sufficient evidence to support the claim that there is a difference in the average per-pupil costs for private school tuition for three counties in the Portland, Oregon, metro area. $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$; $t = 9.2122$; p-value = 0; Reject H_0 . There is significant difference in the mean per-pupil costs for private school tuition for Clackamas and Multnomah counties. $H_0: \mu_1 = \mu_3$; $H_1: \mu_1 \neq \mu_3$; $t = 0.4224$; p-value = 1, Fail to reject H_0 . There is not a significant difference in the mean per-pupil costs for private school tuition for Clackamas and Washington counties. $H_0: \mu_2 = \mu_3$; $H_1: \mu_2 \neq \mu_3$; $t = -7.7312358$; p-value = 0.000, Reject H_0 . There is a significant difference in the mean per-pupil costs for private school tuition for Multnomah and Washington counties.

Chapter 11

- 1) d
 3) e
 5) d
 7) True
 9) False
 11) True
 13) True
 15) d
 17) $H_0: \rho = 0$; $H_1: \rho \neq 0$; $t = 7.6221$; p-value = 0.0000325; Reject H_0 . There is a significant correlation between a person's femur and tibia length.

- 19) $H_0: \rho = 0; H_1: \rho \neq 0; t = 6.2131; p\text{-value} = 0.0004; \text{Reject } H_0. \text{ There is a significant correlation between a student's age and their math score.}$
- 21) a) v
 b) ii
 c) 3.8648
 d) 0.6301
 e) 0.7938
 f) $t = 7.2673$
 g) 3.55E-08
 h) Reject H_0 . Yes, there is a significant relationship between wrist circumference and height.
- 23) a) No
 b) The p-value = 0.00000304 suggests that there is a significant linear relationship between intensity (in candelas) of a 100-watt light bulb was measured by a sensing device at various distances (in meters) from the light source. However, the residual plot clearly shows a nonlinear relationship. Even though we can fit a straight line through the points, we would get a better fit with a curve.
- 25) a
 27) d
- 29) a) $\hat{y} = -3.2852 + 1.0944x$
 b) $H_0: \beta_1 = 0; H_1: \beta_1 \neq 0; t = 25.5639, p\text{-value} = 0; \text{Reject } H_0. \text{ There is a significant linear relationship between the predicted high temperature in the Farmer's Almanac and the actual high temperatures.}$
 c) 60.1913
- d) $53.394 < y < 66.989$
- 31) a) $H_0: \beta_1 = 0; H_1: \beta_1 \neq 0$
 b) $r = 0.8241$
 c) $p\text{-value} = 0.0018$
 d) Reject $H_0; \text{ Yes}$
 e) $R^2 = 0.6792$
 f) $\hat{y} = 0.8893 - 0.0031x$
 g) For every additional average weekly soda consumption, a person's bone density decreases by 0.0031 grams per cubic centimeter.
- h) $\hat{y} = 0.8893 - 0.0031 \cdot 7 = 0.8674$
 i) $y - \hat{y} = 0.8643 - 0.8674 = -0.004$
- 33) $31.636 < y < 54.585$
- 35) a) $\hat{y} = 241.8127 - 1.5618x$
 b) Every year a person ages, their peak heart rate decreases by an average of 1.5618.
 c) $R^2 = 0.93722$
 d) $1 - R^2 = 0.06278$
 e) $s = 5.6924$
 f) $r = -0.9681$
 g) $187.5161 < y < 218.0181$
- 37) $15.817 < y < 18.488$
- 39) $1122.72 < y < 1243.425$
- 41) a) $H_0: \rho = 0; H_1: \rho \neq 0; p\text{-value} = 0.000022; \text{Reject } H_0; \text{ There is a significant correlation between driving speed of a vehicle and the corresponding gas mileage.}$
 b) $s = 2.49433$
 c) $\hat{y} = 32.40313 - 0.1662x$
 d) $14.8697 < y < 24.3424$

Data Sources and References

- Adams, D. (2002). *The Ultimate Hitchhiker's Guide to the Galaxy*. New York: Del Rey/Ballantine Books.
- Kozak, K. (2015). *Statistics Using Technology* (2nd ed.). United States.
- Diez, D. M., Barr, C. D., & Çetinkaya-Rundel Mine. (2016). *OpenIntro Statistics* (3rd ed.). Retrieved from <https://www.openintro.org/book/os/>
- Sagan, C. (1997). The Fine Art of Baloney Detection. *The Demon-Haunted World: Science as a Candle in the Dark*. New York: Ballantine Books.
- B1 assets of financial institutions*. (2013, June 27). Retrieved from www.rba.gov.au/statistics/tables/xls/b01hist.xls
- Benen, S. (2011, September 02). Retrieved from http://www.washingtonmonthly.com/political-animal/2011_09/gop_leaders_stop_taking_credit031960.php
- Capital and rental values of Auckland properties*. (2013, September 26). Retrieved from <http://www.statsci.org/data/oz/rentcap.html>
- Contraceptive use*. (2013, October 9). Retrieved from <http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=35>
- Deaths from firearms*. (2013, September 26). Retrieved from <http://www.statsci.org/data/oz/firearms.html>
- DeNavas-Walt, C., Proctor, B., & Smith, J. U.S. Department of Commerce, U.S. Census Bureau. (2012). *Income, poverty, and health insurance coverage in the United States: 2011* (P60-243). Retrieved from <https://www.census.gov/prod/2012pubs/p60-243.pdf>
- Department of Health and Human Services, ASPE. (2013). *Health insurance marketplace premiums for 2014*. Retrieved from http://aspe.hhs.gov/health/reports/2013/marketplacepremiums/ib_premiumslandscape.pdf
- Electricity usage*. (2013, October 9). Retrieved from <http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=162>
- Fertility rate*. (2013, October 14). Retrieved from <http://data.worldbank.org/indicator/SP.DYN.TFRT.IN>
- Fuel oil usage*. (2013, October 9). Retrieved from <http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=164>
- Gas usage*. (2013, October 9). Retrieved from <http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=165>
- Health expenditure*. (2013, October 14). Retrieved from <http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS>
- Hinatov, M. U.S. Consumer Product Safety Commission, Directorate of Epidemiology. (2012). *Incidents, deaths, and in-depth investigations associated with non-fire carbon monoxide from engine-driven generators and other engine-driven tools, 1999-2011*. Retrieved from <http://www.cpsc.gov/PageFiles/129857/cogenerators.pdf>.
- Life expectancy at birth*. (2013, October 14). Retrieved from <http://data.worldbank.org/indicator/SP.DYN.LE00.IN>
- Median income of males*. (2013, October 9). Retrieved from <http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=137>
- Median income of males*. (2013, October 9). Retrieved from <http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=136>
- Prediction of height from metacarpal bone length*. (2013, September 26). Retrieved from <http://www.statsci.org/data/general/stature.html>

Pregnant woman receiving prenatal care. (2013, October 14). Retrieved from <http://data.worldbank.org/indicator/SH.STA.ANVC.ZS>

United States unemployment. (2013, October 14). Retrieved from <http://www.tradingeconomics.com/united-states/unemployment-rate>

Weissmann, J. (2013, March 20). A truly devastating graph on state higher education spending. *The Atlantic*. Retrieved from <http://www.theatlantic.com/business/archive/2013/03/a-truly-devastating-graph-on-state-higher-education-spending/274199/>

Population Reference Bureau. (2018, March 19). Data. Retrieved from <http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=35>

Nico, G. (2020). Breadcrumb Statistics - Multiple Linear Regression. Retrieved from https://gerardnico.com/data_mining/multiple_regression

O'Connor, J. J., Robertson, E. F., Banach, S., & University of St Andrews, Scotland. (2020). MacTutor History of Mathematics Archive. Retrieved from <https://mathshistory.st-andrews.ac.uk/>

Ryan, B. F., Joiner, B. L., & Ryan, Jr, T. A. (1985). *Cholesterol levels after heart attack*. Retrieved from <http://www.statsci.org/data/general/cholest.html>

Olson, K., & Hanson, J. (1997). Using reiki to manage pain: a preliminary report. *Cancer Prev Control*, 1(2), 108-13. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9765732>

Pulse rates before and after exercise. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/ms212.html>

Ovegard, M., Berndt, K., & Lunneryd, S. (2012). Condition indices of Atlantic cod (gadus morhua) biased by capturing method. *ICES Journal of Marine Science*, doi: 10.1093/icesjms/fss145

Bhat, R., & Kushtagi, P. (2006). A re-look at the duration of human pregnancy. *Singapore Med J.*, 47(12), 1044-8. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17139400>

Life expectancy in southeast Asia. (2013, September 23). Retrieved from <http://apps.who.int/gho/data/node.main.688>

Federal Trade Commission, (2008). *Consumer fraud and identity theft complaint data: January-December 2007*. Retrieved from website: <http://www.ftc.gov/opa/2008/02/fraud.pdf>

Schultz, S. T., Klonoff-Cohen, H. S., Wingard, D. L., Askhoomoff, N. A., Macera, C. A., Ji, M., & Bacher, C. (2006). Breastfeeding, infant formula supplementation, and autistic disorder: the results of a parent survey. *International Breastfeeding Journal*, 1(16), doi: 10.1186/1746-4358-1-16

Glossary

Alternative Hypothesis	A statistical hypothesis that states that there is a difference between a parameter and a specific value, or that there is a difference between two parameters.
Blind Study	The individual does not know which treatment they are getting or if they are getting the treatment or a placebo.
Class Boundary	The halfway points that separate the class limits.
Class Limit	The smallest and largest value that could go into a class.
Cluster Sample	The populations are split into groups called clusters, then one or more clusters are randomly selected and all individuals in the chosen clusters are sampled.
Coefficient of Determination	The coefficient of determination is the fraction (or percent) of the variation in the values of y that is explained by the least-squares regression of y on x .
Coefficient of Variation	The standard deviation divided by the mean, which allows you to compare variability among data sets when the units or scale is different
Combination Rule	The number of ways to select r out of n objects where order does not make a difference.
Completely Randomized Experiment	In this experiment, individuals are randomly placed into the two or more groups. One group gets either nothing or a placebo; this group is called the control group. The groups getting the treatment are called the treatment groups.
Confidence Interval	A range of potential values for the unknown population parameter.
Continuous Random Variable	A variable that has an infinite number of possible values in an interval of numbers.
Continuous Variable	Can take on any value. Continuous variables are usually things you measure.
Convenience Sample	Picking a sample that is conveniently at hand. For example, asking other students in your statistics course or using social media. Most convenience sampling will give biased views and is not encouraged.
Correlation Coefficient	A measure of the strength and direction of the linear relationship between two variables.
Critical Value	The number of standard errors added and subtracted in order to achieve a desired confidence level.
Cross-sectional Study	Data observed, measured, or collected at one point in time.
Descriptive statistics	The process where you collect, organize and describe data.
Discrete Random Variable	A variable that has a countable number of possible values.
Discrete Variable	Can only take on particular values like integers. Discrete variables are usually things you count.
Double-blind Study	Neither the individual nor the researcher knows who is getting which treatment or who is getting the treatment and who is getting the placebo.
Event	A set of certain outcomes of an experiment that you want to have happen.

Experiment	An activity or process that has specific results that can be repeated indefinitely which has a set of well-defined outcomes.
Extrapolation	The use of a regression line for prediction far outside the range of values of the independent variable x .
Factorial Design	This design has two or more independent categorical variables called factors. Each factor has two or more different treatment levels.
Factorial Rule	The number of ways to arrange n objects is $n!$
Five-number Summary	Minimum, first quartile, second quartile, third quartile and maximum value in a data set.
Fundamental Counting Rule	The number of ways to do task 1, 2, ..., n together would be to multiply the number of ways for each task $m_1 \cdot m_2 \cdot \dots \cdot m_n$.
Hypothesis Testing	The scientific method used to evaluate claims about population parameters.
Independent Events	Two events that are not related and the probability of one event does not affect the probability of the other event.
Individual	A person or object that you are interested in finding out information about.
Inferential statistics	The process of using sample data to make inferences about the population.
Influential point	A point in a scatter plot that is positioned far away from the main cluster of data points on the y -axis.
Interquartile Range (IQR)	The distance between the first and third quartile.
Intersection	Where two events overlap and happen at the same time.
Interval data	Numeric data where there is a known difference between values, but zero does not mean "nothing."
Law of Large Numbers	As n increases, the relative frequency tends toward the theoretical probability.
Leverage Point	A point in a scatter plot that is positioned far away from the main cluster of data points on the x -axis.
Lurking Variable	A variable other than the independent or dependent variables that may influence the regression equation.
Margin of Error	Half the width of the confidence interval and is a statistic that represents the amount of random sampling error.
Matched Pairs Design	This is a subset of the randomized block design where the treatments are given to two groups that can be matched up with each other in some way.
Mean	The mean is the arithmetic average of the numbers. This is the center that most people call the average.
Median	The median is the data value in the middle of the data that has 50% of the data below that point and 50% of the data above that point.
Midrange	The average of the smallest and largest data values.
Modal Class	The modal class is the class with the highest frequency.

Mode	The mode is the data value that occurs the most frequently in the data.
Mutually Exclusive Events	Disjoint events that cannot occur at the same time.
Nominal data	Categorical data that has no order or rank. For example, the color of your car, ethnicity, race, or gender.
Null Hypothesis	A statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters.
Observational Study	An observational study is when the investigator collects data by observing, measuring, counting, watching or asking questions. The investigator does not change anything.
Ordinal data	Categorical data that has a natural order to it.
Outcomes	The results of an experiment.
Outlier	An outlier is a data value that is very different from the rest of the data and is very far from the center.
Parameter	Any characteristic or measure from a population. This number is a fixed, unknown number that you want to estimate.
Pareto Chart	A bar graph that is ordered from the most frequent down to the least frequent category.
Permutation Rule	The number of ways to arrange r out of n objects where order does make a difference.
Point Estimate	A sample statistic used to estimate the population parameter.
Population	The total set of all the observations that are the subject of a study.
Probability Distribution	An assignment of probabilities to all the possible values of the random variable.
Prediction Interval	The confidence interval for the predicted value of y .
Prospective (or Longitudinal) Study	Data collected in the future from groups sharing common factors.
Qualitative or Categorical Variable	A word or name that describes a quality of the individual
Quantitative or Numerical Variable	A number (quantity), something that can be counted or measured from the individual.
Quartile	Numbers that divide a data set into fourths. One fourth (or a quarter) of the data falls between consecutive quartiles.
Random Variable	If you have a variable, and can find a probability associated with that variable, it is called a random variable.
Range	The difference between the largest and smallest data values.
Randomized Block Design	A block is a group of subjects that are similar or the same subject measured multiple times, but the blocks differ from each other. Then randomly assign treatments to subjects inside each block.

Ratio Data	Numeric data that has a true zero, meaning when the variable is zero nothing is there. Most measurement data are ratio data.
Replication	Repetition of an experiment on more than one subject so you can make sure that the sample is large enough to distinguish true effects from random effects. It is also the ability for someone else to duplicate the results of the experiment.
Residual	The vertical distance between the actual value of y and the predicted value of y .
Retrospective Study	Data collected from the past using records, interviews, and other similar artifacts.
Sample	A subset from the population.
Sample Space	Collection of all possible outcomes of the experiment.
Scatterplot	A scatterplot shows the relationship between two quantitative variables measured on the same individuals.
Simple Random Sample (SRS)	Selecting a sample size of n objects from the population so that every sample of the same size n has equal probability of being selected as every other possible sample of the same size from that population.
Standard Deviation	The average distance of the data points from the mean
Standard Error of Estimate	The standard deviation of the residuals.
Statistic	Any characteristic or measure from a sample.
Statistical Ethics	The ethical principles and considerations that guide the responsible and ethical conduct of statistical analysis, including ensuring integrity, accuracy, transparency, and fairness in data collection, analysis, interpretation, and reporting. It involves addressing issues such as privacy, confidentiality, bias, and the potential impact of statistical results on individuals and society.
Statistical Hypothesis	An educated conjecture about a population parameter.
Statistics	Statistics is the study of how to collect, organize, analyze, and interpret data collected from a group.
Stratified Sample	The population is split into groups called strata, then a random sample is taken from each stratum.
Systematic Sample	We list the entire population, then randomly pick a starting point n and then take every n^{th} value until the sample size is reached.
Type I Error	Rejecting the null hypothesis when it was actually true.
Type II Error	Failing to reject the null hypothesis when it is actually false.
Union	The junction of two events including their intersection.
Variable (also known as a random variable)	The measurement or observation of the individual.
Variance	The average squared distance from the data points to the mean.
Z-Score	The number of standard deviations a data point is from the mean.

Symbols

n	Sample Size
N	Population Size
Σ	Sum
\bar{x}	Sample Mean
μ	Population Mean
s	Sample Standard Deviation
σ	Population Standard Deviation
s^2	Sample Variance
σ^2	Population Variance
CV	Coefficient of Variation
z	Z-Score
IQR	Interquartile Range
P_i	Percentile
Q_i	Quartile
$P(x)$	Probability of $X = x$
A^C	Complement Rule
$A \cap B$	Intersection
$A \cup B$	Union
$A B$	Conditional Probability
$n!$	Factorial Rule
${}_n C_r$	Combination Rule
${}_n P_r$	Permutation Rule
p	Probability of a Success
q	Probability of a Failure
p	Population Proportion
\hat{p}	Sample Proportion
p	P-Value
α	Significance Level = P(Type I Error) = False Positive Rate
$1 - \alpha$	P(True Negative)
β	P(Type II Error) = False Negative Rate
$1 - \beta$	Power = P(True Positive)
H_0	Null Hypothesis

$H_1 = H_a$	Alternative Hypothesis
$z_{\alpha/2}$	Critical Value for Standard Normal Distribution
z	z Test Statistic
df	Degrees of Freedom
$t_{\alpha/2}$	Critical Value for t-Distribution
t	t Test Statistic
D	Difference
\bar{D}	Mean of the Differences
s_D	Standard Deviation of the Differences
F	F Test Statistic
F_α	F Critical Value
χ^2	Chi-Square Test Statistic
χ_α^2	Chi-Square Critical Value
E	Expected Value
O	Observed Value
k	Number of Groups
SS	Sum of Squares
MS	Mean Square
r	Sample Correlation Coefficient
ρ	Population Correlation Coefficient
R	Multiple Correlation Coefficient
R^2	Coefficient of Determination
$1 - R^2$	Coefficient of Nondetermination
\hat{y}	Predicted Value of the Regression Equation
b_0	Sample Y-Intercept Coefficient
b_1	Sample Slope Coefficient
β_1	Population Slope Coefficient
e_i	Residual
$s_{est} = s$	Standard Error of Estimate = Standard Deviation of the Residuals
p	Number of predictors in regression
ε	Error term in regression

Greek Letter Pronunciations

Upper Case	Lower Case	Pronunciation
A	α	<i>alpha</i> – æl-fə
B	β	<i>beta</i> – bei-tə
Γ	γ	<i>gamma</i> – gæ-mə
Δ	δ	<i>delta</i> – del-tə
E	ε	<i>epsilon</i> – eps-ill-aan
Z	ζ	<i>zeta</i> – zei-tə
H	η	<i>eta</i> – ei-tə
Θ	θ	<i>theta</i> – thei-tə
I	ι	<i>iota</i> – I-oh-tə (<i>'I' pronounced like 'eye'</i>)
K	κ	<i>kappa</i> – kæ-pə
Λ	λ	<i>lambda</i> – læm-də
M	μ	<i>mu</i> – myoo
N	ν	<i>nu</i> – nyoo
Ξ	ξ	<i>xi</i> – ksaai (<i>as in sick sigh</i>)
O	ο	<i>omicron</i> – oh-mə-kraan
Π	π	<i>pi</i> – paai (<i>the same as pie</i>)
P	ρ	<i>rho</i> – roh (<i>rhymes with go</i>)
Σ	σ	<i>sigma</i> – sig-mə
T	τ	<i>tau</i> – taa'u (<i>rhyming with cow</i>)
Υ	υ	<i>upsilon</i> – 'ups' as oops, 'ilon' as ill-on
Φ	φ	<i>phi</i> – faai (<i>as in identify</i>)
X	χ	<i>chi</i> – kaai (<i>as in kite</i>)
Ψ	ψ	<i>psi</i> – saai (<i>as in side</i>)
Ω	ω	<i>omega</i> – oh-mey-gə

Mostly Harmless Elementary Statistics Formula Packet

Chapter 3 Formulas

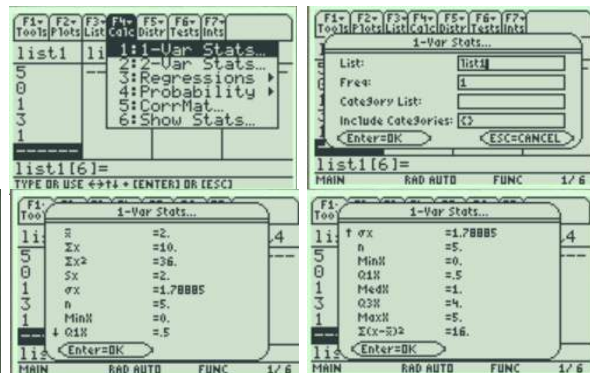
Sample Mean: $\bar{x} = \frac{\sum x}{n}$	Population Mean: $\mu = \frac{\sum x}{N}$
Weighted Mean: $\bar{x} = \frac{\sum(xw)}{\sum w}$	Range = Max – Min
Sample Standard Deviation: $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	Population Standard Deviation = σ
Sample Variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$	Population Variance = σ^2
Coefficient of Variation: $CVar = \left(\frac{s}{\bar{x}} \cdot 100\right) \%$	Z-Score: $z = \frac{x-\bar{x}}{s}$
Percentile Index: $i = \frac{(n+1)p}{100}$	Interquartile Range: $IQR = Q_3 - Q_1$
Outlier Lower Limit: $Q_1 - (1.5 \cdot IQR)$	Outlier Upper Limit: $Q_3 + (1.5 \cdot IQR)$
Correlation Coefficient $SS_{xx} = (n - 1)s_x^2$ $SS_{yy} = (n - 1)s_y^2$ $SS_{xy} = \sum(xy) - (n \cdot \bar{x} \cdot \bar{y})$ $r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$	Slope $b_1 = \frac{SS_{xy}}{SS_{xx}}$ y-Intercept $b_0 = \bar{y} - b_1\bar{x}$ Regression Equation $\hat{y} = b_0 + b_1x$

TI-84: Enter the data in a list and then press [STAT]. Use cursor keys to highlight CALC. Press 1 or [ENTER] to select **1:1-Var Stats**. Press [2nd], then press the number key corresponding to your data list. Press [Enter] to calculate the statistics. Note: the calculator always defaults to L₁ if you do not specify a data list.

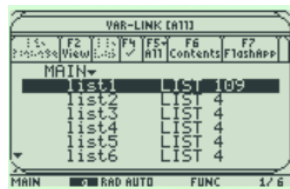


s_x is the sample standard deviation. You can arrow down and find more statistics. Use the min and max to calculate the range by hand. To find the variance simply square the standard deviation.

TI-89: Press [APPS], select **FlashApps** then press [ENTER]. Highlight **Stats/List Editor** then press [ENTER]. Press [ENTER] again to select the main folder. To clear a previously stored list of data values, arrow up to the list name you want to clear, press [CLEAR], then enter.



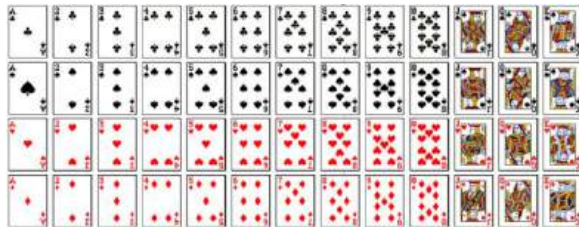
Press [F4], select 1: 1-Var Stats. To get the list name to the List box, press [2nd] [Var-Link], arrow down to list1 and press [Enter]. This will bring list1 to the List box. Press [Enter] to enter the list name



and then enter again to calculate. Use the down arrow key to see all the statistics. S_x is the sample standard deviation. You can arrow down and find more statistics. Use the min and max to calculate the range by hand. To find the variance simply square the standard deviation or take the last sum of squares divided by $n - 1$.

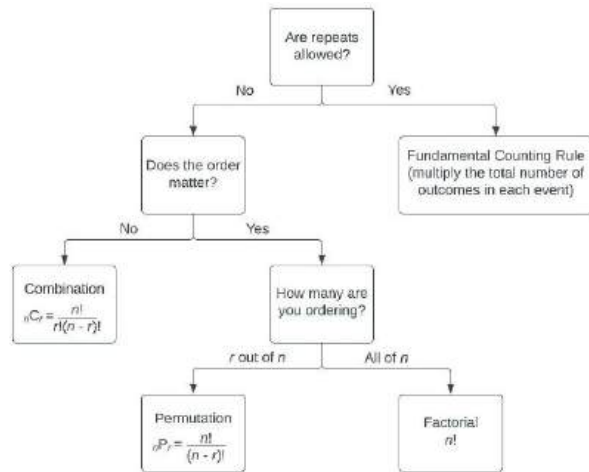
Chapter 4 Formulas

Complement Rules: $P(A) + P(A^c) = 1$ $P(A) = 1 - P(A^c)$ $P(A^c) = 1 - P(A)$	Mutually Exclusive Events: $P(A \cap B) = 0$
Union Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Independent Events: $P(A \cap B) = P(A) \cdot P(B)$
Intersection Rule: $P(A \cap B) = P(A) \cdot P(B A)$	Conditional Probability Rule: $P(A B) = \frac{P(A \cap B)}{P(B)}$
Fundamental Counting Rule: $m_1 \cdot m_2 \cdot \dots \cdot m_n$	Factorial Rule: $n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1$
Combination Rule: ${}_n C_r = \frac{n!}{r!(n-r)!}$	Permutation Rule: ${}_n P_r = \frac{n!}{(n-r)!}$



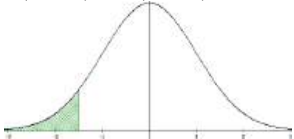
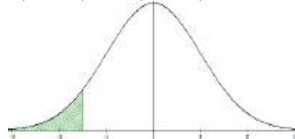
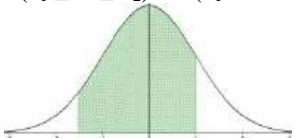
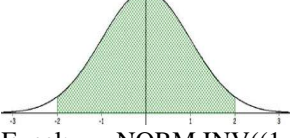
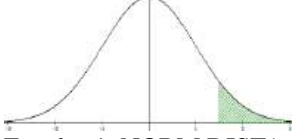
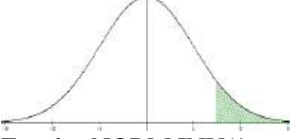
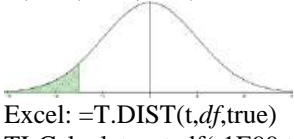
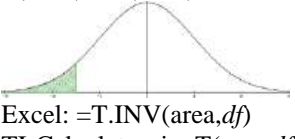
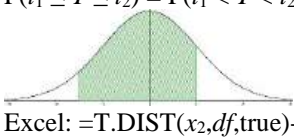
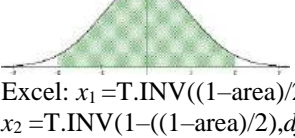
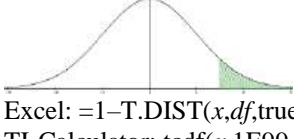
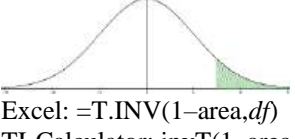
clubs = ♣, spades = ♠, hearts = ♥, diamonds = ♦

Sum of 2 Dice		Second Die					
		1	2	3	4	5	6
First Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

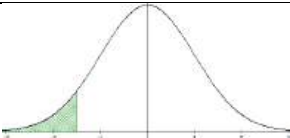
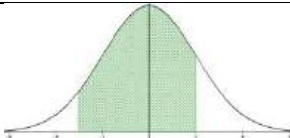
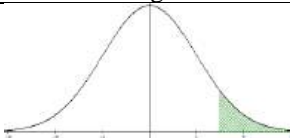


Chapter 5 Formulas

$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
Is	Is less than or equal to	Is greater than or equal to
Is equal to	Is at most	Is at least
Is exactly the same as	Is not greater than	Is not less than
Has not changed from	Within	Is more than or equal to
Is the same as		
Excel: =binom.dist(x,n,p,0) TI Calculator: binompdf(n,p,x)	Excel: =binom.dist(x,n,p,1) TI Calculator: binomcdf(n,p,x)	Excel: =1-binom.dist(x-1,n,p,1) TI Calculator: 1-binomcdf(n,p,x-1)
	$P(X > x)$	$P(X < x)$
Binomial Distribution: $P(X = x) = {}_n C_x \cdot p^x \cdot q^{(n-x)}$, $x = 0, 1, 2, \dots, n$ Mean: $\mu = n \cdot p$ Variance: $\sigma^2 = n \cdot p \cdot q$ Standard Deviation: $\sigma = \sqrt{n \cdot p \cdot q}$	More than	Less than, Fewer than
	Greater than	Below
	Above	Lower than
	Higher than	Shorter than
	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced
	Excel: =1-binom.dist(x,n,p,1) TI Calculator: 1-binomcdf(n,p,x)	Excel: =binom.dist(x-1,n,p,1) TI Calculator: binomcdf(n,p,x-1)

<p>Discrete Distribution Table: $0 \leq P(x_i) \leq 1$ $\sum P(x_i) = 1$ Mean: $\mu = \sum(x_i \cdot P(x_i))$ Variance: $\sigma^2 = \sum(x_i^2 \cdot P(x_i)) - \mu^2$ Standard Deviation: $\sigma = \sqrt{\sigma^2}$</p>	<p>Binomial Distribution: $P(X = x) = {}_n C_x \cdot p^x \cdot q^{(n-x)}$, $x = 0, 1, 2, \dots, n$ Mean: $\mu = n \cdot p$ Variance: $\sigma^2 = n \cdot p \cdot q$ Standard Deviation: $\sigma = \sqrt{n \cdot p \cdot q}$</p>
<p>Normal Distribution Probabilities: $P(X \leq x) = P(X < x)$</p>  <p>Excel: =NORM.DIST(x,μ,σ,true) TI-Calculator: normalcdf(-1E99,x,μ,σ)</p>	<p>Percentiles for Normal Distribution: $P(X \leq x) = P(X < x)$</p>  <p>Excel: =NORM.INV(area,μ,σ) TI-Calculator: invNorm(area,μ,σ)</p>
<p>$P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2) =$</p>  <p>Excel: =NORM.DIST(x2,μ,σ,true)-NORM.DIST(x1,μ,σ,true) TI-Calculator: normalcdf(x1,x2,μ,σ)</p>	<p>$P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2) =$</p>  <p>Excel: $x_1 = \text{NORM.INV}((1-\text{area})/2, \mu, \sigma)$ $x_2 = \text{NORM.INV}(1-((1-\text{area})/2), \mu, \sigma)$ TI-Calculator: $x_1 = \text{invNorm}((1-\text{area})/2, \mu, \sigma)$ $x_2 = \text{invNorm}(1-((1-\text{area})/2), \mu, \sigma)$</p>
<p>$P(X \geq x) = P(X > x)$</p>  <p>Excel: =1-NORM.DIST(x,μ,σ,true) TI-Calculator: normalcdf(x,1E99,μ,σ)</p>	<p>$P(X \geq x) = P(X > x)$</p>  <p>Excel: =NORM.INV(1-area,μ,σ) TI-Calculator: invNorm(1-area,μ,σ)</p>
<p>T-Distribution Probabilities: $P(T \leq t) = P(T < t)$</p>  <p>Excel: =T.DIST(t,df,true) TI-Calculator: tcdf(-1E99,t,df)</p>	<p>Percentiles for t-Distribution: $P(T \leq t) = P(T < t)$</p>  <p>Excel: =T.INV(area,df) TI-Calculator: invT(area,df)</p>
<p>$P(t_1 \leq T \leq t_2) = P(t_1 < T < t_2) =$</p>  <p>Excel: =T.DIST(x2,df,true)-T.DIST(x1,df,true) TI-Calculator: tcdf(t1,t2,df)</p>	<p>$P(t_1 \leq T \leq t_2) = P(t_1 < T < t_2) =$</p>  <p>Excel: $x_1 = \text{T.INV}((1-\text{area})/2, df)$, $x_2 = \text{T.INV}(1-((1-\text{area})/2), df)$ TI-Calculator: $x_1 = \text{invT}((1-\text{area})/2, df)$, $x_2 = \text{invT}(1-((1-\text{area})/2), df)$</p>
<p>$P(T \geq t) = P(T > t)$</p>  <p>Excel: =1-T.DIST(x,df,true) TI-Calculator: tcdf(x,1E99,df)</p>	<p>$P(T \geq t) = P(T > t)$</p>  <p>Excel: =T.INV(1-area,df) TI-Calculator: invT(1-area,df)</p>

In the table below, note that when $\mu = 0$ and $\sigma = 1$ use the NORM.S. DIST or NORM.S.INV function in Excel for a standard normal distribution.

P(X ≤ x) or P(X < x)	P(x₁ < X < x₂) or P(x₁ ≤ X ≤ x₂)	P(X ≥ x) or P(X > x)
Is less than or equal to	Between	Is greater than or equal to
Is at most		Is at least
Is not greater than		Is not less than
Within		More than
Less than		Greater than
Below		Above
Lower than		Higher than
Shorter than		Longer than
Smaller than		Bigger than
Decreased		Increased
Reduced		Larger
		
Excel Finding a Probability =NORM.DIST(x,μ,σ,true) Finding a Percentile =NORM.INV(area,μ,σ)	Excel Finding a Probability =NORM.DIST(x ₂ ,μ,σ,true) – NORM.DIST(x ₁ ,μ,σ,true) Finding a Percentile x ₁ =NORM.INV((1–area)/2,μ,σ) x ₂ =NORM.INV(1–((1–area)/2),μ,σ)	Excel Finding a Probability =1–NORM.DIST(x,μ,σ,true) Finding a Percentile =NORM.INV(1–area,μ,σ)
TI Calculator Finding a Probability =normalcdf(-1E99,x,μ,σ) Finding a Percentile =invNorm(area,μ,σ)	TI Calculator Finding a Probability =normalcdf(x ₁ ,x ₂ ,μ,σ) Finding a Percentile x ₁ =invNorm((1–area)/2,μ,σ) x ₂ =invNorm(1–((1–area)/2),μ,σ)	TI Calculator Finding a Probability =normalcdf(x,1E99,μ,σ) Finding a Percentile =invNorm(1–area,μ,σ)

Chapter 6 Formulas

Confidence Interval for One Proportion $\hat{p} \pm z_{\alpha/2} \sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$ $\hat{p} = \frac{x}{n}$ $\hat{q} = 1 - \hat{p}$ TI-Calculator: 1-PropZInt	t-Confidence Interval $\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}}\right)$ $df = n - 1$ TI-Calculator: TInterval
Z-Critical Values $\alpha = 1 - \text{confidence level}$ Excel: $z_{\alpha/2} = \text{NORM.INV}(1 - (\alpha/2), 0, 1)$ TI-Calculator: $z_{\alpha/2} = \text{invNorm}(1 - (\alpha/2), 0, 1)$	t-Critical Values $\alpha = 1 - \text{confidence level}$ Excel: $t_{\alpha/2} = \text{T.INV}(1 - (\alpha/2), df)$ TI-Calculator: $t_{\alpha/2} = \text{invT}(1 - (\alpha/2), df)$
Sample Size for Proportion $n = p^* \cdot q^* \left(\frac{z_{\alpha/2}}{E}\right)^2$ Always round up to whole number. If p is not given use $p^* = 0.5$. E = Margin of Error	Sample Size for Mean $n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$ Always round up to whole number. E = Margin of Error

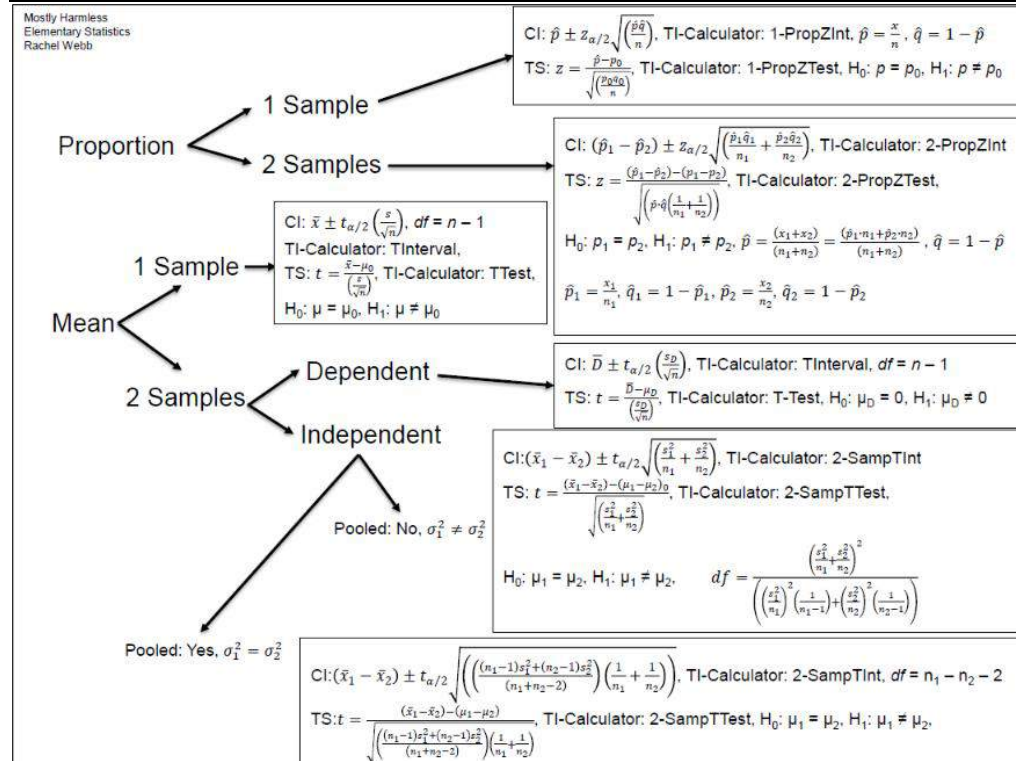
Chapter 7 Formulas

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \mu = \mu_0$ or $H_0: p = p_0$ $H_1: \mu \neq \mu_0$ $H_1: p \neq p_0$	$H_0: \mu = \mu_0$ or $H_0: p = p_0$ $H_1: \mu > \mu_0$ $H_1: p > p_0$	$H_0: \mu = \mu_0$ or $H_0: p = p_0$ $H_1: \mu < \mu_0$ $H_1: p < p_0$
Claim is in the Null Hypothesis		
=	≤	≥
Is equal to	Is less than or equal to	Is greater than or equal to
Is exactly the same as	Is at most	Is at least
Has not changed from	Is not more than	Is not less than
Is the same as	Within	Is more than or equal to
Claim is in the Alternative Hypothesis		
≠	>	<
Is not	More than	Less than
Is not equal to	Greater than	Below
Is different from	Above	Lower than
Has changed from	Higher than	Shorter than
Is not the same as	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

<p>Type I Error- Reject H_0 when H_0 is true.</p> <p>Type II Error- Fail to reject H_0 when H_0 is false.</p>	<p>Rejection Rules: P-value method: reject H_0 when the p-value $\leq \alpha$. Critical value method: reject H_0 when the test statistic is in the critical region (shaded tails).</p>
<p>Hypothesis Test for One Proportion $H_0: p = p_0$ $H_1: p \neq p_0$ $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$ TI-Calculator: 1-PropZTest</p>	<p>Hypothesis Test for One Mean $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ $t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$ TI-Calculator: T-Test</p>
<p>z-Critical Values Excel: Two-tail: $z_{\alpha/2} = \text{NORM.INV}(1-(\alpha/2), 0, 1)$ Right-tail: $z_{1-\alpha} = \text{NORM.INV}(1-\alpha, 0, 1)$ Left-tail: $z_{\alpha} = \text{NORM.INV}(\alpha, 0, 1)$</p> <p>TI-Calculator: Two-tail: $z_{\alpha/2} = \text{invNorm}(1-(\alpha/2), 0, 1)$ Right-tail: $z_{1-\alpha} = \text{invNorm}(1-\alpha, 0, 1)$ Left-tail: $z_{\alpha} = \text{invNorm}(\alpha, 0, 1)$</p>	<p>t-Critical Values Excel: Two-tail: $t_{\alpha/2} = \text{T.INV}(1-(\alpha/2), df)$ Right-tail: $t_{1-\alpha} = \text{T.INV}(1-\alpha, df)$ Left-tail: $t_{\alpha} = \text{T.INV}(\alpha, df)$</p> <p>TI-Calculator: Two-tail: $t_{\alpha/2} = \text{invT}(1-(\alpha/2), df)$ Right-tail: $t_{1-\alpha} = \text{invT}(1-\alpha, df)$ Left-tail: $t_{\alpha} = \text{invT}(\alpha, df)$</p>

Chapter 8 Formulas

<p>Hypothesis Test for Two Proportions $H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$ $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ TI-Calculator: 2-PropZTest $\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)} = \frac{(\hat{p}_1 n_1 + \hat{p}_2 n_2)}{(n_1 + n_2)}$ $\hat{q} = 1 - \hat{p} \quad \hat{p}_1 = \frac{x_1}{n_1} \quad \hat{p}_2 = \frac{x_2}{n_2}$</p>	<p>Confidence Interval for Two Proportions $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\left(\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}\right)}$ $\hat{p}_1 = \frac{x_1}{n_1} \quad \hat{p}_2 = \frac{x_2}{n_2}$ $\hat{q}_1 = 1 - \hat{p}_1 \quad \hat{q}_2 = 1 - \hat{p}_2$ TI-Calculator: 2-PropZInt For z Critical Values refer back to Chapter 7</p>
<p>Hypothesis Test for Two Dependent Means $H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$ $t = \frac{\bar{D} - \mu_D}{\left(\frac{s_D}{\sqrt{n}}\right)}$ TI-Calculator: T-Test</p>	<p>Confidence Interval for Two Dependent Means $\bar{D} \pm t_{\alpha/2} \left(\frac{s_D}{\sqrt{n}}\right)$ TI-Calculator: TInterval For t-Critical Values refer back to Chapter 7</p>
<p>Hypothesis Test for Two Independent Means $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ TI-Calculator: 2-SampTTest T-Test: Assume Variances are Unequal $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$ $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)\right)}$ T-Test: Assume Variances are Equal $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ $df = n_1 + n_2 - 2$</p>	<p>Confidence Interval for Two Independent Means T-Interval: Assume Variances are Unequal TI-Calculator: 2-SampTInt $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$ $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)\right)}$ T-Interval: Assume Variances are Equal $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ $df = n_1 + n_2 - 2$ For t-Critical Values refer back to Chapter 7</p>



Chapter 9 Formulas

Goodness of Fit Test $H_0: p_1 = p_0, p_2 = p_0, \dots, p_k = p_0.$ $H_1: \text{At least one proportion is different.}$ $\chi^2 = \sum \frac{(O-E)^2}{E}$ $df = k - 1, p_0 = 1/k \text{ or given \%}$ TI-84: χ^2 GOF-Test	Test for Independence $H_0: \text{Variable 1 and Variable 2 are independent.}$ $H_1: \text{Variable 1 and Variable 2 are dependent.}$ $\chi^2 = \sum \frac{(O-E)^2}{E}$ $df = (R - 1)(C - 1)$ TI-84: χ^2 -Test
---	--

Chapter 10 Formulas

One-Way ANOVA: $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ $k = \text{number of groups}$ $H_1: \text{At least one mean is different.}$																				
<table border="1"> <thead> <tr> <th>Source</th> <th>SS = Sum of Squares</th> <th>df</th> <th>MS = Mean Square</th> <th>F</th> </tr> </thead> <tbody> <tr> <td>Between (Factor)</td> <td>$\sum n_i(\bar{x}_i - \bar{x}_{GM})^2$</td> <td>$k - 1$</td> <td>$MSB = \frac{SSB}{k-1}$</td> <td>$F = \frac{MSB}{MSW}$</td> </tr> <tr> <td>Within (Error)</td> <td>$\sum (n_i - 1)s_i^2$</td> <td>$N - k$</td> <td>$MSW = \frac{SSW}{N-k}$</td> <td></td> </tr> <tr> <td>Total</td> <td>SST</td> <td>$N - 1$</td> <td></td> <td></td> </tr> </tbody> </table>	Source	SS = Sum of Squares	df	MS = Mean Square	F	Between (Factor)	$\sum n_i(\bar{x}_i - \bar{x}_{GM})^2$	$k - 1$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$	Within (Error)	$\sum (n_i - 1)s_i^2$	$N - k$	$MSW = \frac{SSW}{N-k}$		Total	SST	$N - 1$		
Source	SS = Sum of Squares	df	MS = Mean Square	F																
Between (Factor)	$\sum n_i(\bar{x}_i - \bar{x}_{GM})^2$	$k - 1$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$																
Within (Error)	$\sum (n_i - 1)s_i^2$	$N - k$	$MSW = \frac{SSW}{N-k}$																	
Total	SST	$N - 1$																		
$\bar{x}_i = \text{sample mean from the } i^{\text{th}} \text{ group}$ $s_i^2 = \text{sample variance from the } i^{\text{th}} \text{ group}$ $n_i = \text{sample size of the } i^{\text{th}} \text{ group}$ $N = n_1 + n_2 + \dots + n_k$ $\bar{x}_{GM} = \frac{\sum x_i}{N}$																				
F-Critical Values Excel: Two-tail: $F_{\alpha/2} = \text{F.INV}(1-\alpha/2, 0, 1)$ Right-tail: $F_{1-\alpha} = \text{F.INV}(1-\alpha, 0, 1)$ Left-tail: $F_{\alpha} = \text{F.INV}(\alpha, 0, 1)$ TI-Calculator: invF program can be downloaded at http://www.MostlyHarmlessStatistics.com .																				
Bonferroni test statistic: $t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\left(MSW \left(\frac{1}{n_i} + \frac{1}{n_j} \right) \right)}}$ $H_0: \mu_i = \mu_j$ $H_1: \mu_i \neq \mu_j$ Multiply p-value by $m = {}_kC_2$, divide area for critical value by $m = {}_kC_2$.																				

Chapter 12 Formulas

$SS_{xx} = (n - 1)s_x^2$ $SS_{yy} = (n - 1)s_y^2$ $SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y}$	Correlation Coefficient $r = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}}$	Coefficient of Determination $R^2 = (r)^2 = \frac{SSR}{SST}$																				
Slope $= b_1 = \frac{SS_{xy}}{SS_{xx}}$ y-intercept $= b_0 = \bar{y} - b_1\bar{x}$	Correlation t-test $H_0: \rho = 0; H_1: \rho \neq 0$ $t = r \sqrt{\frac{n-2}{1-r^2}}$ $df = n - 2$																					
Regression Equation (Line of Best Fit): $\hat{y} = b_0 + b_1x$	Slope t-test $H_0: \beta_1 = 0; H_1: \beta_1 \neq 0$ $t = \frac{b_1}{\sqrt{\left(\frac{MSE}{SS_{xx}} \right)}}$ $df = n - p - 1 = n - 2$																					
Residual $e_i = y_i - \hat{y}_i$ (Good residual plots should have no patterns.)	Slope/Model F-test $H_0: \beta_1 = 0; H_1: \beta_1 \neq 0$																					
Standard Error of Estimate $s_{est} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{MSE}$	<table border="1"> <thead> <tr> <th>Source</th> <th>SS = Sum of Squares</th> <th>df</th> <th>MS = Mean Square</th> <th>F</th> </tr> </thead> <tbody> <tr> <td>Regression</td> <td>$SSR = \frac{(SS_{xy})^2}{SS_{xx}}$</td> <td>$p$</td> <td>$MSR = \frac{SSR}{p}$</td> <td>$F = \frac{MSR}{MSE}$</td> </tr> <tr> <td>Error</td> <td>$SSE = SS_{yy} - SSR$</td> <td>$n - p - 1$</td> <td>$MSE = \frac{SSE}{n-p-1}$</td> <td></td> </tr> <tr> <td>Total</td> <td>$SST = SS_{yy}$</td> <td>$n - 1$</td> <td></td> <td></td> </tr> </tbody> </table>	Source	SS = Sum of Squares	df	MS = Mean Square	F	Regression	$SSR = \frac{(SS_{xy})^2}{SS_{xx}}$	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$	Error	$SSE = SS_{yy} - SSR$	$n - p - 1$	$MSE = \frac{SSE}{n-p-1}$		Total	$SST = SS_{yy}$	$n - 1$			
Source	SS = Sum of Squares	df	MS = Mean Square	F																		
Regression	$SSR = \frac{(SS_{xy})^2}{SS_{xx}}$	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$																		
Error	$SSE = SS_{yy} - SSR$	$n - p - 1$	$MSE = \frac{SSE}{n-p-1}$																			
Total	$SST = SS_{yy}$	$n - 1$																				
Prediction Interval $\hat{y} \pm t_{\alpha/2} \cdot s_{est} \sqrt{\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}} \right)}$																						

