chapter 2-4

Inferential Statistics

Statistics means never having to say you're certain.

Introduction

The process of sampling from a population and making inference about a phenomenon in that population is at the heart of the scientific process. In some disciplines like Physics, relationships may be highly predictable with error stemming mainly from measurement issues. In biological systems, error can also stem from measurement issues; however, a greater source of error is natural biological variation, which is often multifactorial. Statistical tests are therefore necessary to be able to address the existence of relationships or differences between variables when this biological variation is present. The heart of statistical analysis is the concept of confidence in your statement about the tested hypothesis. How likely is it that you are correct in rejecting the null hypothesis and accepting that there is indeed a relationship or difference between groups? You draw inference about the population based upon the results in the sample. By definition therefore, an **Inferential Statistical** test is one whereby inference is made about the population based upon analysis of a sample selected from that population. Regardless of the specific test, an identical process is used for all inferential statistical tests:

- **Select a sample representative of the population.** The method of sample selection is important and the size of the sample must be large enough to allow appropriate probability testing.

- **Calculate the appropriate test statistic.** The test statistic used is determined by the hypothesis being tested and the research design as a whole.

- **Test the Null hypothesis.** Compare the calculated test statistic to its critical value at the predetermined level of acceptance.

Once the results of the statistical analysis are known, it is important to report them in a fashion that is clear and consistent with the way other scientists will be reporting their results. Guidelines for the reporting of statistical results are outlined in the latter part of this chapter. These will be of particular importance to you as you write your term project paper.

Select a Sample Representative of the Population

Since we cannot measure all the individuals in a population, we measure only a portion of the population. This is referred to as our sample. Ideally the sample is representative of the whole population, this being determined by our sampling strategy and the size of the sample.

Sample Selection Method: Random sampling is the obvious choice as the best selection method in many situations. True random samples are often hard to achieve however. In random sampling each member of the population has an equal likelihood for selection into the sample. The simplest way that this can be achieved is put all the names of the members of the population into a hat and pull out the required number of members for your population. Another method is to use random number tables to select which members of the population to include in the sample.

Be wary of using the term random sample when in fact it is a selected sample. For instance, if you set up outside the university cafeteria and “randomly” asked students to participate in your study from those who walked by, you would have far from a random sample of university students. There may be a bias in the type of students who frequent that part of the university; you may have a subconscious bias in the type of person you are “randomly” selecting; there may be a bias in the type of students that would volunteer for your study. All of these factors, and more, conspire to ensure that you do not have a random sample representative of the whole university population. The only way you could select a random sample would be to obtain all students id numbers and select from them using random number tables or more convenient if the data is in a computer file, carry out the selection electronically. In EXCEL you can assign a random number to each case and then sort by this number to select your cases. In SPSS you can select a random sample of a given number of cases from your total file. Even if you do make your random selection of students you still have to convince them to partake in the study. This can introduce a bias if for some reason a certain type of student is more likely to refuse to partake in the study. In this case, it is important to try to keep track of any data that might be analyzed to detect if there are differences in those who refuse participation.

In the 1981 Canada Fitness Study, a very good attempt was made to carry out a random sampling of Canadians. Stats Canada produced lists of randomly selected residences in Canada. Two person measurement teams would then knock on the door and ask to measure all occupants. Surprisingly, there was about an 80% success rate with this method! If a household refused, there was a back up list of residences to try. There is one major deficiency of this method, in that only people who lived in residences could be selected. This is merely an example of how hard it is to set up a true random sampling procedure.

Most times a selected sample based upon certain criteria is used. It is therefore beholden upon the researcher to report the sampling method so that the reader can have insight into how representative the sample is of the population as a whole. Theoretically, random sampling is a prerequisite for inferential statistical analysis; however, inferential statistics are often used on non-randomly selected samples. A common selection process is to use a team or school or a certain workplace. These convenience samples are acceptable as long as the selection criteria are made clear.

Random Allocation: If you run an experiment whereby one group receives a treatment and the other does not, even if the sample is not randomly selected, you can assign subjects to the two treatment groups randomly. This is called random allocation, and is a way of ensuring that any differences between groups are due to the treatment and not some selection bias. It is this random allocation that is referred to in a “randomized controlled trial.” It is a very valuable design methodology.

Sample Size: Commonsense would say that a sample size of 1 would be insufficient, but how about 10 or 50 or 100? How big does the sample need to be? Unfortunately, the answer to this is not straightforward. It depends upon the research design, the precision of measurements, the variability of the measures in the population, and the magnitude of difference or relationship expected. **Formal sample size calculations can be carried out for each type of test to determine the required sample size. These calculations are beyond the scope of this text.** However, during later discussions of inferential tests we will discuss the impact of sample size upon findings.

Sampling Errors: Suppose you sample a given population and calculate the mean of the variable that you have measured. If you sample the population again and calculate the mean for this second sample, it will most likely be different than the first sample mean. The two means may be slightly different or very different. If we keep repeatedly sampling from the population we will eventually have very many sample means. The mean of all these sample means would be our best estimate of the population mean. Interestingly, a characteristic of these means is that they would tend to be normally distributed. The difference between any sample mean and the true population mean could be regarded as an error due to the sampling. Thus, the term Sampling Errors is used to describe the deviation of these sample means from the population mean. Sampling errors tend to be normally distributed, such that the characteristics of the normal distribution can be applied:

- 68% of the sample means are within ± 1 standard deviation of the true population mean
- 95% of the sample means are within ± 1.96 standard deviation of the true population mean

- Mean of all of the sample means is the best estimate of the true population mean
- Standard deviation of the sample means is the best estimate of the population standard deviation.

Calculate the Appropriate Test Statistic.

Since computer software for statistical analysis is so readily available, the actual process of calculating a test statistic using some complicated equation is irrelevant to most researchers. What is key however, is which test to use in any given situation. This text was written with this in mind. Therefore, the actual calculations have been down played while understanding when and how to use each statistic has been emphasized. Hopefully the chapters here will provide you with the information required to ensure in the future that you will select the appropriate test statistic for the question at hand.

Test the Null Hypothesis

If two samples are selected, and their means are different, this may be due to sampling errors or it could be because they truly are samples from different populations. Null hypothesis testing aims to resolve this issue. The Null hypothesis is a statement that the observed difference or relationship was due to chance (i.e., due to sampling errors). If the null hypothesis is rejected, the researcher is stating that the likelihood is that the observed finding was not due to chance and indeed reflects a true phenomenon. Central to these statements is the associated probability or confidence. No inferential statistic can give absolute certainty that the null hypothesis is false, or conversely that a difference or relationship truly exists in the population. Rather, an inferential statistic can simply provide a level of confidence in that statement. Indeed, in research areas which are particularly bombarded with numerous confounding factors such as in Epidemiology, scientists do not rely on one study to make confident statements about the existence of relationships to disease and mortality rates, but require multiple reported studies of similar and dissimilar design each with at least 95% confidence in results pointing in the similar direction. Only if every member of the population were measured, could you ever have 100% confidence in your findings.

Setting a Probability Level for Acceptance: It is common to select a level of probability that will be used to accept that there is a significant result. Prior to analysis the researcher must decide upon their level of acceptance. Tests of significance are conducted at pre-selected probability levels, symbolized by ***p*** or **α**. The vast majority of the time the probability level of 0.05, is used. A ***p*** of .05 means that if you reject the null hypothesis, then you expect to find a result of this magnitude by chance only 5 in 100 times. Or conversely, if you carried out the experiment 100 times you would expect to find a result of this magnitude 95 times. You therefore have 95% confidence in your result. A more stringent test would be one where the ***p*** =

0.01, which translates to 99% confidence in the result. The choice of acceptance level is up to the researcher, but once chosen it should be adhered to.

You will often see a table of results in a published paper where significance is indicated at two or three different levels. For instance they may have carried out many correlation coefficient analyses and then reported that some were non-significant, some were significant at p<0.05, some at p<0.01 and the rest significant at p<0.001. Clearly these researchers did not select a single level of acceptance. Some refer to this approach as using a rubber yard stick. Snedecor and Cochran (1967) were unambiguous on this issue. Either the researcher should pre-select one level of acceptance and stick to it, or do away with a set level of acceptance all together and simply report the exact probability of each test statistic. If for instance, you had calculated a t statistic and it had an associated probability of p = 0.032, you could either say the probability is lower than the pre-set acceptance level of 0.05 therefore a significant difference at the 95% level of confidence or simply talk about 0.032 as a percentage confidence (96.8%)

Type I and II Errors: Based upon the calculated test statistic for your sample, you choose to accept or reject the null hypothesis. You may however be incorrect in your decision. If you reject a true null hypothesis (you conclude that there is a significant relationship when there is not) it is referred to as a **type I error**. If you accept a false null hypothesis (you conclude there is no significant relationship when there is) it is referred to as a **type II error**. You can not eliminate the possibility of these errors. If you selected p=0.01 instead of p=0.05 as your acceptance level you would reduce the likelihood of committing a type I error, but unfortunately increase the chance of committing a type II error. This is one of the reasons why you should predetermine the acceptance level and not carry out significance testing and wait to see how significant the findings are. If you are more concerned about committing a type I error than a type II error then set a stringent acceptance level such as p = 0.01 or even p = 0.001. This might be done because the consequences of a type I error are more profound, such as medical research where life and death may be at stake. Most commonly an acceptance level of p = 0.05 is selected, which seems to represent the best balance between committing the two types of errors.

One and Two-Tailed tests: Most tests of significance are two-tailed, meaning that the null hypothesis can be rejected irrespective of the direction of the effect. For example, when two sample means are compared, the null hypothesis is rejected if the mean of sample A is sufficiently larger than the mean of sample B, or if the mean of sample A is sufficiently smaller than the mean of sample B. A one- tailed test of significance is used when the researcher is sure that differences can occur only in one direction. When this is the case, for a particular probability level, a smaller magnitude of difference will result in rejection of the null hypothesis.

Critical Values of the Test Statistic: Most statistical tests result in the calculation of a test statistic which is evaluated in the following manner:

- The test statistic is calculated
- The critical value of the test statistic is determined based upon sample size and probability acceptance level (found in a table at the back of a stats book or part of the EXCEL or SPSS output)
- The calculated test statistic must be greater than the critical value of the test statistic to reject the null hypothesis and accept a significant difference or relationship.

The actual value of the test statistic is relatively unimportant. What is important is the probability associated with the value of the test statistic in relation to the associated degrees of freedom (sample size). In the SPSS output, the calculated value of the test statistic will be reported along with the actually probability level associated with it. If the value is less than 0.05 you can reject the null hypothesis at $p<0.05$ or the 95% confidence level. You do not actually need to know the critical value of the test statistic. The critical value becomes more important when you are relying on a stats book to look up the critical value to compare to your calculated value of the test statistic.

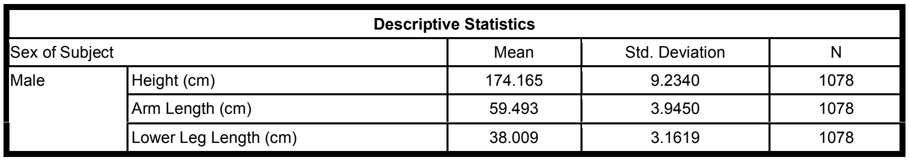
Statistical Significance Is Not the Same as Practical Significance: The fact that you have determined that there is a statistically significant result does not mean that this finding will be of practical significance. Over the years, measures of human morphology (eg Somatotype) have been shown to be statistically significantly related to measures of temperament. However, the very large sample sizes, in the thousands, cause low correlations (around $r=0.15$), to be statistically significant. The inference here is that we are at least 95% confident that the very weak relationship indicated by the correlation coefficient exists in the population from which we have sampled. Unfortunately, with such a weak relationship there would be no practical significance to these findings. For example, there would be no predictive ability. Another example might be a finding that one method was significantly better than another in training workers. If the statistically better method was a lot more time consuming and costly it might not be practical to use the better methodology unless the increase in training quality far outreached the expense.

Reporting Statistics

When reporting statistical results in your paper it is important that you are clear, concise, and consistent. In the methods section of your paper make sure that you identify your statistical methods, and cite them using textbooks or review papers. Also cite the commercial software you used for your analysis.

When you run your statistical analysis in SPSS you will be faced with copious output with lots of superfluous information. Your task will be to select the pertinent information and report it in well constructed tables or simply as part of the text. The tendency is to want to simply copy and paste the SPSS out put in your paper. However, it is vital that you take the effort to select only the appropriate information. Figure 2-4.1 shows a typical Descriptive Statistics output from SPSS. You might think that it is ready to go straight into you paper. However, Table 1 below the SPSS output is much more appropriate for inclusion in a paper.

SPSS Output:



Descriptive Statistics

Sex of Subject		Mean	Std. Deviation	N
Male	Height (cm)	174.165	9.2340	1078
	Arm Length (cm)	59.493	3.9450	1078
	Lower Leg Length (cm)	38.009	3.1619	1078

Table produced from relevant SPSS output information:

Table 1: Mean (SD) of Height, Arm Length and Lower Leg Length of Men (N=1078)

Height, Free Standing (cm)	174.2 (9.2)
Arm Length, Right (cm)	59.5 (3.9)
Lower Leg Length, Right (cm)	38.0 (3.2)

Figure 2-4.1: Example of a table layout based upon information from an SPSS statistical output

American Physiological Society Guidelines for Reporting Statistics: In 2004 the American Physiological Society published Guidelines for Reporting Statistics in their journals (Curran-Everett and Benos, 2004). The full article describing the guidelines is posted on the course website, and the guidelines are listed below. Guidelines cannot substitute for understanding statistical concepts and procedures, but they do improve the quality of reporting of statistical information in published journal articles. These guidelines should be helpful for your term paper project.

Guideline 1. *If in doubt, consult a statistician when you plan your study.*

Guideline 2. *Define and justify a critical significance level* $\propto$ *appropriate to the goals of your study.*

Guideline 3. *Identify your statistical methods, and cite them using textbooks or review papers. Cite separately commercial software you used to do your statistical analysis.*

Guideline 4. *Control for multiple comparisons.*

Guideline 5. *Report variability using a standard deviation.*

Guideline 6. *Report uncertainty about scientific importance using a confidence interval.*

Guideline 7. *Report a precise P value.*

Guideline 8. *Report a quantity so the number of digits is commensurate with scientific relevance.*

Guideline 9. *In the Abstract, report a confidence interval and a precise P value for each main result.*

Guideline 10. *Interpret each main result by assessing the numerical bounds of the confidence interval and by considering the precise P value.*

Further Reading

Snedecor, G.W. and W.G. Cochran: Statistical Methods. Iowa State University Press, Ames, Iowa, U.S.A. 1967.

Council of Science Editors, Style Manual Subcommittee. Scientific Style and Format: the CSE Manual for Authors, Editors, and Publishers (7th ed.). Reston, VA: Rockefeller Univ. Press, 2006.

Curran-Everett D, Benos DJ, American Physiological Society. Guidelines for reporting statistics in journals published by the American Physiological Society. Am J Physiol Endocrinol Metab. 2004 Aug;287(2):E189-91.