Descriptive statistics is the discipline of quantitatively describing the main features of a collection of data. Descriptive statistics are distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential statistics, are not developed on the basis of probability theory.

Statistical inference is the process of drawing conclusions from data that is subject to random variation, for example, observational errors or sampling variation.
More substantially, the terms **statistical inference, statistical induction and inferential statistics** are used to describe systems of procedures that can be used to draw conclusions from datasets arising from systems affected by random variation, such as observational errors, random sampling, or random experimentation. Initial requirements of such a system of procedures for inference and induction are that the system should produce reasonable answers when applied to well-defined situations and that it should be general enough to be applied across a range of situations.

Unit: A single entity of statistical interests

Population: A complete set of collection of unit.

Statistical population: The set of all measurements.

Sample: A subset from the statistical population

A good sample should be randomized and is representative of the population while a **bad sample** contains subjectivity and is biased.

Descriptive measures of a statistical sample (or population) computed from the data.

Given data, $x_1, x_2, \cdots, x_i, \cdots, x_n$

the most standard measures are:

Sample mean (often simply called the **mean**) is the average value

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median describes the center of the data set ***when the data is ordered by value***

- If *n* is an odd number, there is exactly one data value lying at the center of the ordered data values. It is the *k*'th data value where *k = (n+1)/2.* In this case the median value is x_k

- If *n* is an even number, there are two values "lying at the center" of the ordered data values. Those are the *k*'th and *k*+1'st values where *k* = *n/2.* In this case the median value is defined as

$$\frac{x_k + x_{k+1}}{2}$$

Variance and Standard Deviation: In addition to knowing the *average* behavior (value) of a set of data, we would like to know how much the data is spread about the average. This is computed by **deviations from the mean**

Let $\bar{x}$ be the mean value of the data $x_1, x_2, \cdots, x_i, \cdots, x_n$

The differences, $x_1 - \bar{x}, \quad x_2 - \bar{x}, \cdots, x_i - \bar{x}, \cdots, x_n - \bar{x}$ are called the deviations from the mean

The sum of the deviations $\sum_{i=1}^{n}(x_i - \bar{x})$ = 0

(i.e. the sum of the positive deviations exactly cancels the sum of the negative deviations.) This is one of the “nice” properties of the mean.

Summing the squares of the deviations gives a non-zero result which gives a useful measure of the spread of the data about the mean. The common measure for this spread is the **sample variance***, s*^{2}

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Since the deviations sum to 0, the value of 1 of them can always be calculated from the other *n* – 1. Therefore only *n* – 1 of the deviations are independent, explaining the *n* – 1 (instead of *n*) appearing in the denominator of *s*^{2}

Since the units of the sample variance are the square of the units of the data, we define the **sample standard deviation**

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

If the deviations from the mean are generally large, the standard deviation will be large. If the deviations from the mean are generally small, the standard deviation will be small.

When do we know the standard deviation is large? Is s = 1.67 large or small?

That depends on the mean value. Thus we use

the fraction **relative variation,** ${}^{s}/_{\bar{x}}$

or the percent **coefficient of variation**, $V = {}^{s}/_{\bar{x}} \cdot 100\%$

to quantify the size of the standard deviation.

e.g. a set of measurements made with micrometer's A and B give the following table of results. Which micrometer measurements are more accurate?

Micrometer	mean	std dev	coef of var
A	3.92 mm	0.0152 mm	0.39%
B	1.54 in	0.0086 in	0.56%

Alternate way to compute sample variance

$$s^2 \equiv \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad = \quad \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}$$

Operations count

“+”	n-1	n-1
“−”	n+1	2
“×”	n	n+2
“÷”	1	1
Total	3n+1	2n+4
more efficient on a	*computer*	*hand calculator*

Example: meals served over 5 weeks: 15 14 2 27 13

n = 5, mean number meals served is $\bar{x} = \frac{15+14+2+27+13}{5} = 14.2$

the ordered data is 2 13 14 15 27
median value of meals served is given by
the *k*'th data value where *k* = (*n+1*)/2 = 6/2 = 3
That is median = $x_3 = 14$

e.g. In the alloy strength data, *n* = 58.

The mean value of alloy strength is $\bar{x} =$ 70.7

To compute the median value with *n* even, we want the
k = 58/2 = 29'th and *k* + 1 = 30'th
values in the ordered data array.
From the stem-leaf display, these two values are rapidly found to be 70.5 and 70.6.
Therefore, the median value of alloy strength is (70.5 + 70.6)/2 = 70.55

Except in very special cases, the mean and median values are **NOT** the same.

Mean values are very susceptible to outlier values, median values are not.

e.g. reconsider the meals per week data: 2 13 14 15 27
The mean value is $\bar{x} = 14.2$; the median value is 14.

Suppose the last value is changed from 27 to 67, giving the data set 2 13 14 15 67
The mean value is now $\bar{x} = 22.2$; the median value is **unchanged** at 14.

The **mean** is the most common "single number" used to discribe the "**average**" behavior of a data set. This is commonly misinterpreted to also be the "central" behavior of the data set. Except in special cases to be discussed later in the course, the mean *does not* characterize the central behavior of the data set.

If you want to know the **central** behavior of the data, use the **median.**

e.g. if you are told that the average grade on an exam is 77.6, *does that mean half of the class scored above* 77.6?

Computing mean and sample variance from data with repetition

Consider the data set: 7.6, 7.6, 7.7, 7.7, 7.7, 7.7, 7.8, 7.8, 7.8, 7.9
The mean value is

$$\bar{x} = \frac{7.6+7.6+7.7+7.7+7.7+7.7+7.8+7.8+7.8+7.9}{10}$$

$$= \frac{7.6 \cdot 2+7.7 \cdot 4+7.8 \cdot 3+7.9 \cdot 1}{10}$$

Therefore, if a data set consists of *n* values, but only *k* < *n* of the values are different, where the value x_i occurs g_i times then

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot g_i}{n}$$

Similarly, for variance:

$$s^2 = \frac{\sum_{i=1}^{k} g_i \, (x_i - \bar{x})^2}{n-1}$$

Consider Sample Data:

compressive strength measurements of 58 samples of an aluminum alloy

66.4, 67.7, 68.0, 68.0, 68.3, 68.4, 68.6, 68.8, 68.9, 69.0, 69.1,
69.2, 69.3, 69.3, 69.5, 69.5, 69.6, 69.7, 69.8, 69.8, 69.9, 70.0,
70.0, 70.1, 70.2, 70.3, 70.3, 70.4, 70.5, 70.6, 70.6, 70.8, 70.9,
71.0, 71.1, 71.2, 71.3, 71.3, 71.5, 71.6, 71.6, 71.7, 71.8, 71.8,
71.9, 72.1, 72.2, 72.3, 72.4, 72.6, 72.7, 72.9, 73.1, 73.3, 73.5,
74.2, 74.5, 75.3

Estimating mean and sample variance from a frequency distribution

Suppose the original data is lost and all you have is the frequency distribution

You can approximate the computation of the sample mean and variance using

$$\bar{x} \approx \frac{\sum_{i=1}^{k} x_i \, f_i}{n}$$

$$s^2 \approx \frac{\sum_{i=1}^{k} f_i \, (x_i - \bar{x})^2}{n-1}$$

Alloy Strength

Bin	Bin Mark x_i	Frequency f_i
(66.3 67.3]	66.8	1
(67.3 68.3]	67.8	4
(68.3 69.3]	68.8	9
(69.3 70.3]	69.8	13
(70.3 71.3]	70.8	11
(71.3 72.3]	71.8	10
(72.3 73.3]	72.8	6
(73.3 74.3]	73.8	2
(74.3 75.3]	74.8	2

where

x_i are the bin marks

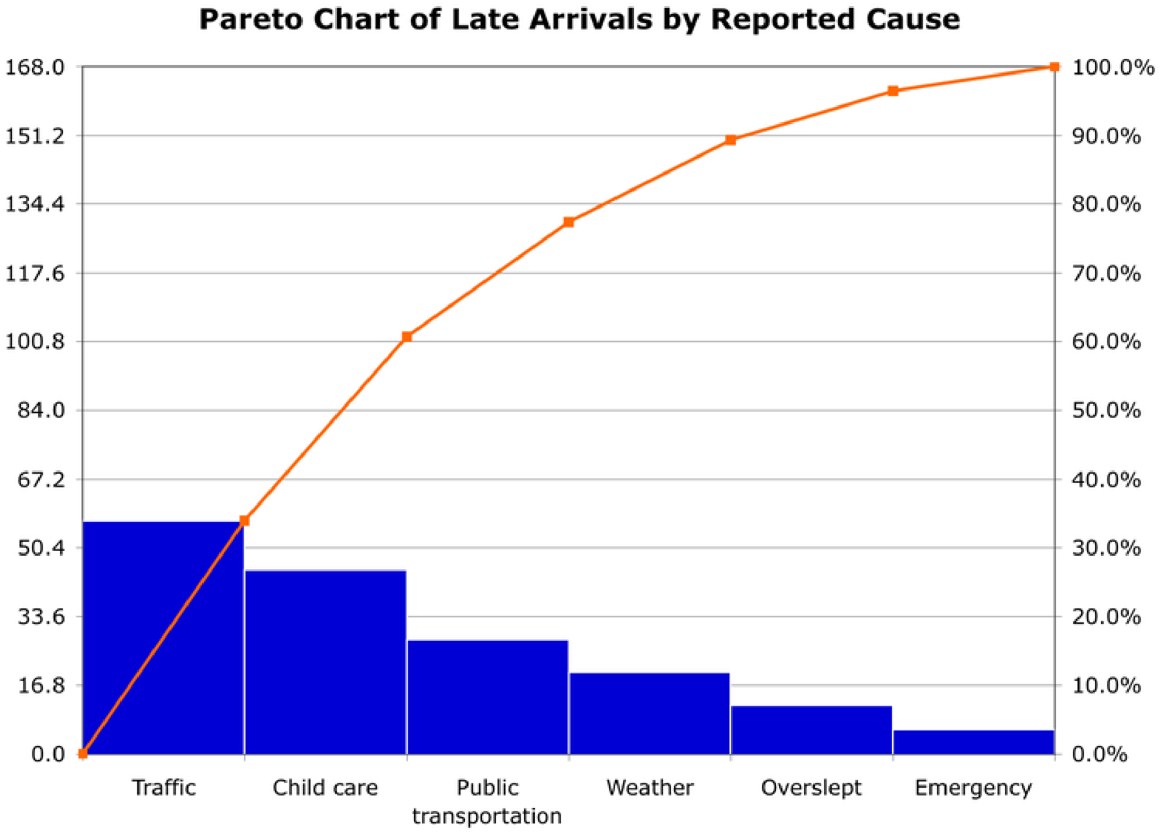
f_i are the bin frequencies

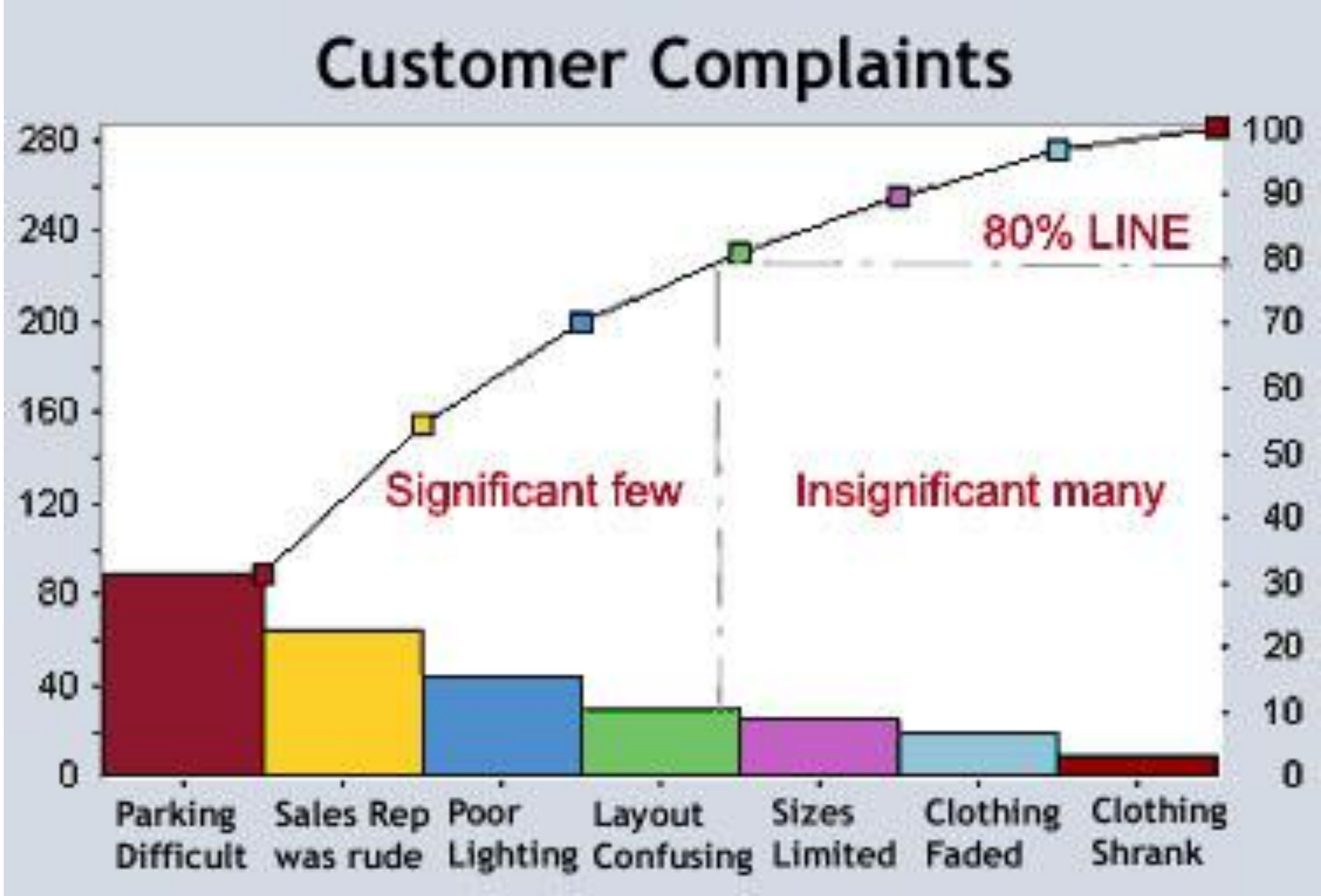
k is the number of bins (9 in the example table)

and $n = \sum_{i=1}^{k} f_i$ is the number of data points (58 in the example table)

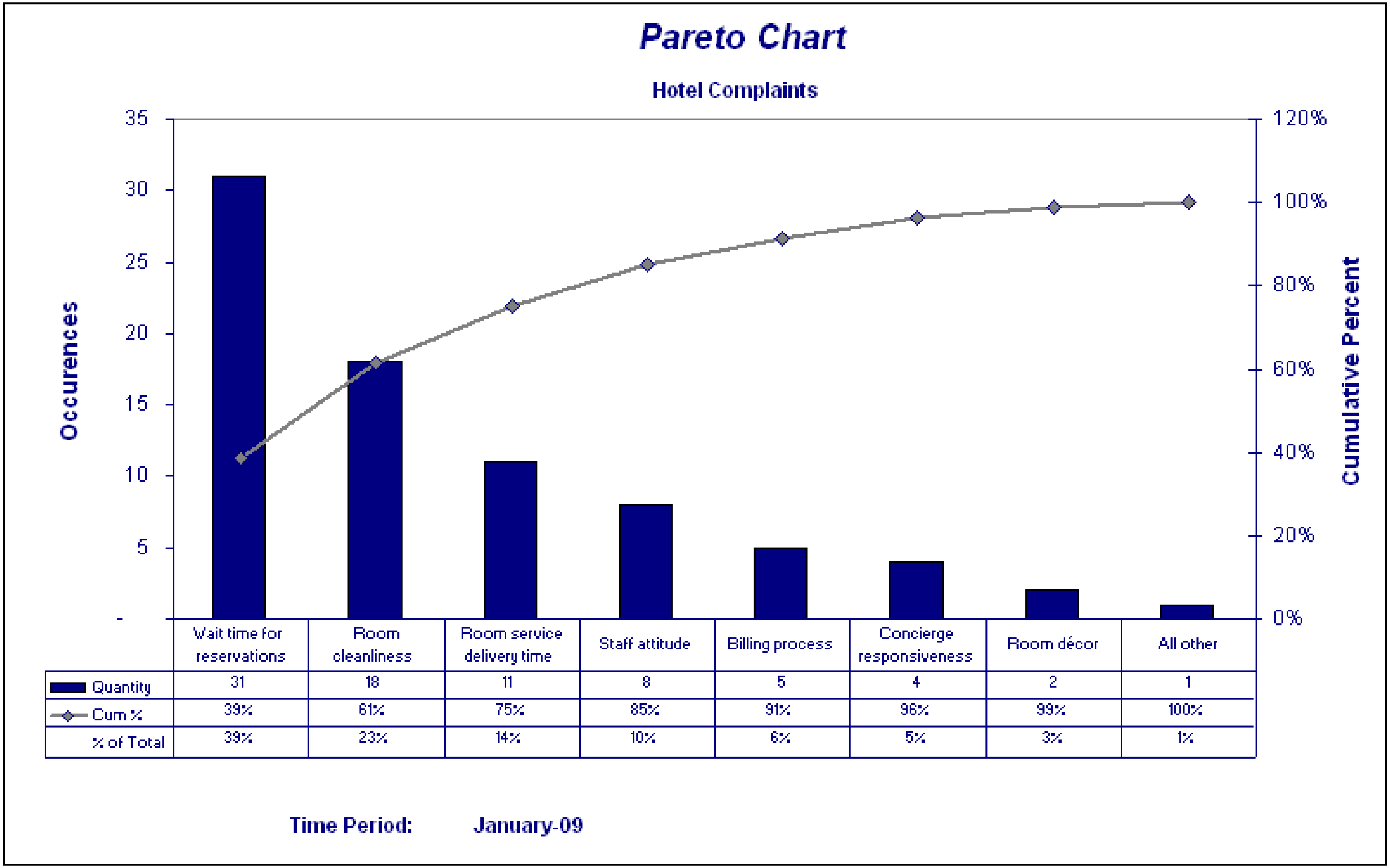
A Pareto chart, named after Vilfredo Pareto (1848 – 1923, was an Italian engineer, sociologist, economist, political scientist and philosopher), is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line.

Pareto's empirical law: any assortment contains a few major components and many minor components.



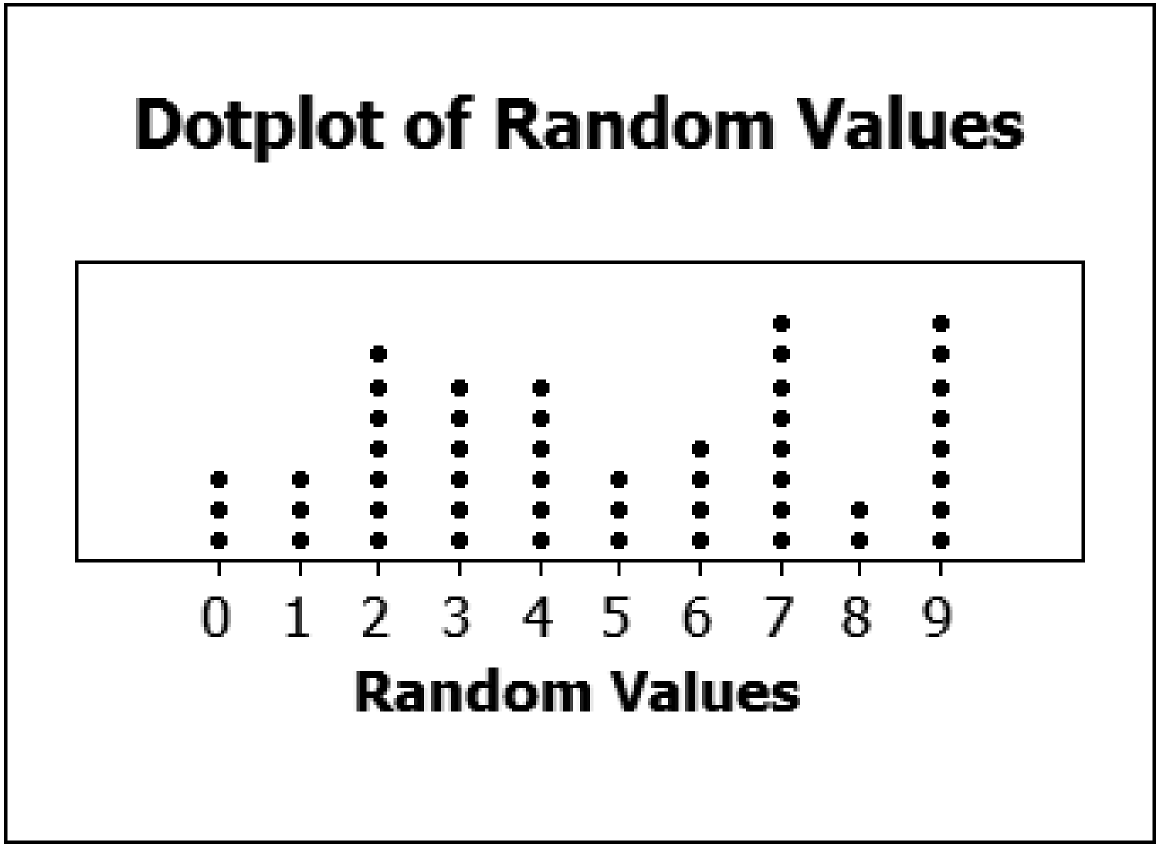

Customer Complaints
280
240
200
160
120
80
40
0
100
90
80
70
60
50
40
30
20
10
0
80% LINE
Significant few
Insignificant many
Parking Difficult
Sales Rep was rude
Poor Lighting
Layout Confusing
Sizes Limited
Clothing Faded
Clothing Shrank

Pareto diagrams: Pareto diagrams are bar charts that show the percentage of the total response variance attributable to each factor and interaction.

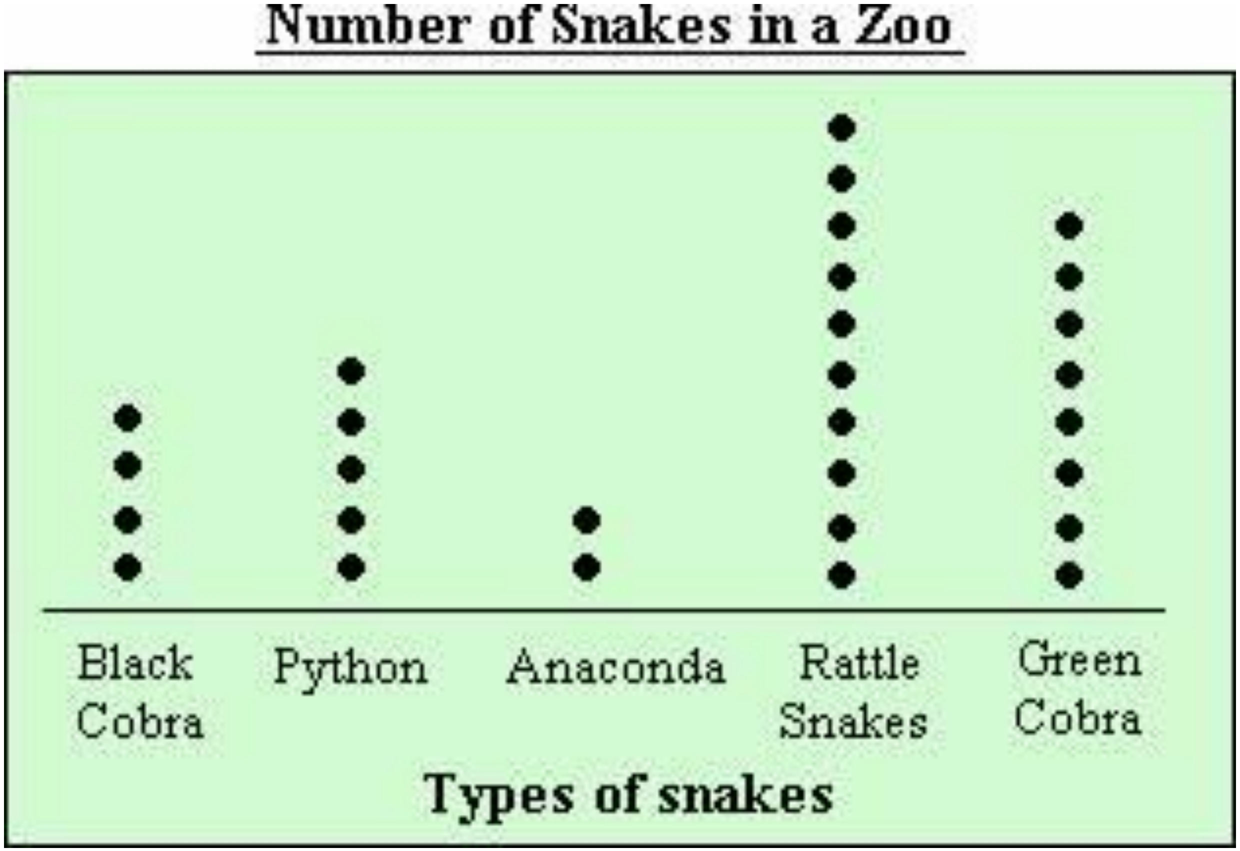


	Wait time for reservations	Room cleanliness	Room service delivery time	Staff attitude	Billing process	Concierge responsiveness	Room décor	All other
▬ Quantity	31	18	11	8	5	4	2	1
—◆— Cum %	39%	61%	75%	85%	91%	96%	99%	100%
% of Total	39%	23%	14%	10%	6%	5%	3%	1%

Wilkinson dot plots: as a representation of a distribution consists of group of data points plotted on a simple scale.



Cleveland dot plots: refer to plots of points that each belong to one of several categories. They are an alternative to bar charts or pie charts, and look somewhat like a horizontal bar chart where the bars are replaced by a dots at the values associated with each category.



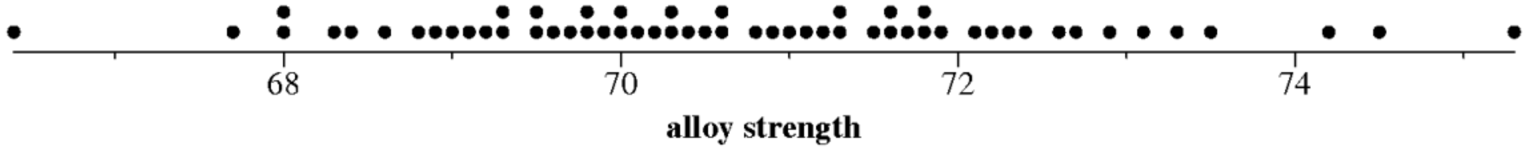
Sample Data:

compressive strength measurements of 58 samples of an aluminum alloy

66.4, 67.7, 68.0, 68.0, 68.3, 68.4, 68.6, 68.8, 68.9, 69.0, 69.1,
69.2, 69.3, 69.3, 69.5, 69.5, 69.6, 69.7, 69.8, 69.8, 69.9, 70.0,
70.0, 70.1, 70.2, 70.3, 70.3, 70.4, 70.5, 70.6, 70.6, 70.8, 70.9,
71.0, 71.1, 71.2, 71.3, 71.3, 71.5, 71.6, 71.6, 71.7, 71.8, 71.8,
71.9, 72.1, 72.2, 72.3, 72.4, 72.6, 72.7, 72.9, 73.1, 73.3, 73.5,
74.2, 74.5, 75.3

Dot diagram: each data value represented as a point, with care taken so that points with the same value do not overlap

Youtube teaching page: http://www.youtube.com/watch?v=GK5NgLzLmpc
Or search “dot plot statstics

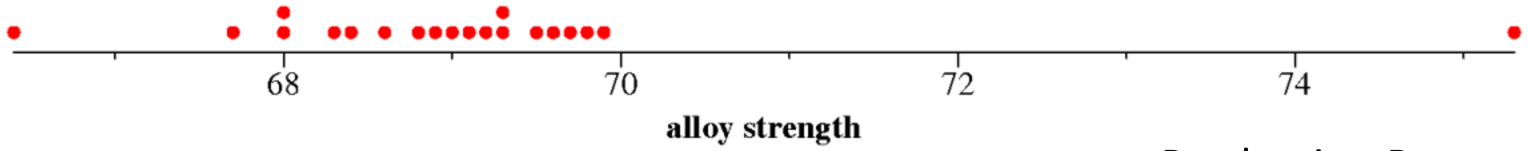


A **dot chart or dot plot** is a statistical chart consisting of data points plotted on a simple scale, typically using filled in circles. There are two common, yet very different, versions of the dot chart. The first is described by Wilkinson as a graph that has been used in hand-drawn (pre-computer era) graphs to depict distributions. The other version is described by Cleveland as an alternative to the bar chart, in which dots are used to depict the quantitative values (e.g. counts) associated with categorical variables.

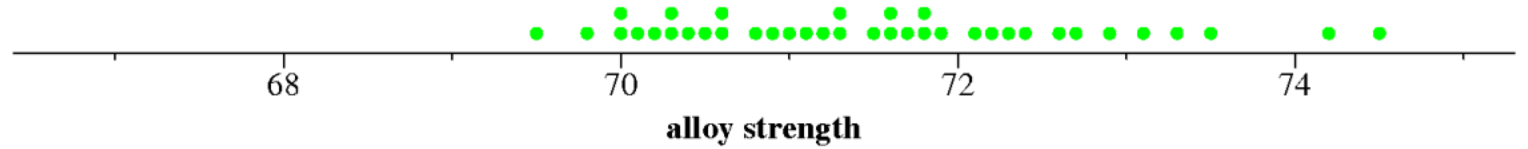
Data assessment through visualization

- outliers
- differences between sets of samples

Production A



Production B



Frequency Distributions (Tables)

Look at frequency of distribution of data over subintervals (bins, classes) spanning the entire range of data

Alloy Strength

Subinterval	Frequency	Relative Frequency	Percent Frequency	Cumulative Frequency
(66.3 67.3]	1	0.0172	1.72	1
(67.3 68.3]	4	0.0690	6.90	5
(68.3 69.3]	9	0.1552	15.52	13
(69.3 70.3]	13	0.2241	22.41	26
(70.3 71.3]	11	0.1897	18.97	37
(71.3 72.3]	10	0.1724	17.24	47
(72.3 73.3]	6	0.1034	10.34	53
(73.3 74.3]	2	0.0345	3.45	56
(74.3 75.3]	2	0.0345	3.45	58
Total	58	1.0000	100.00	

Histogram:

Frequency Distribution: number n_i of data values in bin *i*

Relative Frequency Distribution: relative number n_i / n of data values in bin *i*

(*n* is the total number of data values)

Percent Frequency Distribution: percent ((n_i / n)*100) of data values in bin *i*

Cumulative Distribution: Total number of data values $\sum_{j=1}^{i} n_j$

up to and including bin *i*

Note:

- **Bins must not overlap**
- **Bins must accommodate all data**
- **Bins should be of the same width**

Frequency Distribution Decisions

- number of bins & bin width

  Goals - enough (≥ 30 – 100) data values in most bins

  - enough bins (≥ 10) to see shape of distribution

  If the data set is small (< 100), the number of bins and amount of data per bin will be reduced from the optimum goals. Aim for ≥ 5 bins.

  strength data: 58 values from 66.4 to 75.3.

  9 bins over the range 66.3 to 75.3 gives a bin width of (75.3 – 66.3)/9 = 1.0 with an average frequency of 58/9 = 6.4 data values per bin.

- bin ranges must respect data accuracy

  e.g. (66.3, 67.3], (67.3, 68.3] if data is given to "tenths" values (1 decimal place)

  (65, 67], (67, 69] if data is given only to "ones" value (no decimal places)

- numerical assignment to bin

  each data value must be uniquely assigned to a single bin

  ( ] convention
  (66.3, 67.3] ⟵ data value 67.3 assigned to bin 1
  (67.3, 68.3]

  [ ) convention
  [66.3, 67.3)
  [67.3, 68.3) ⟵ data value 67.3 assigned to bin 2

The textbook adopts the following notation for bins (classes)

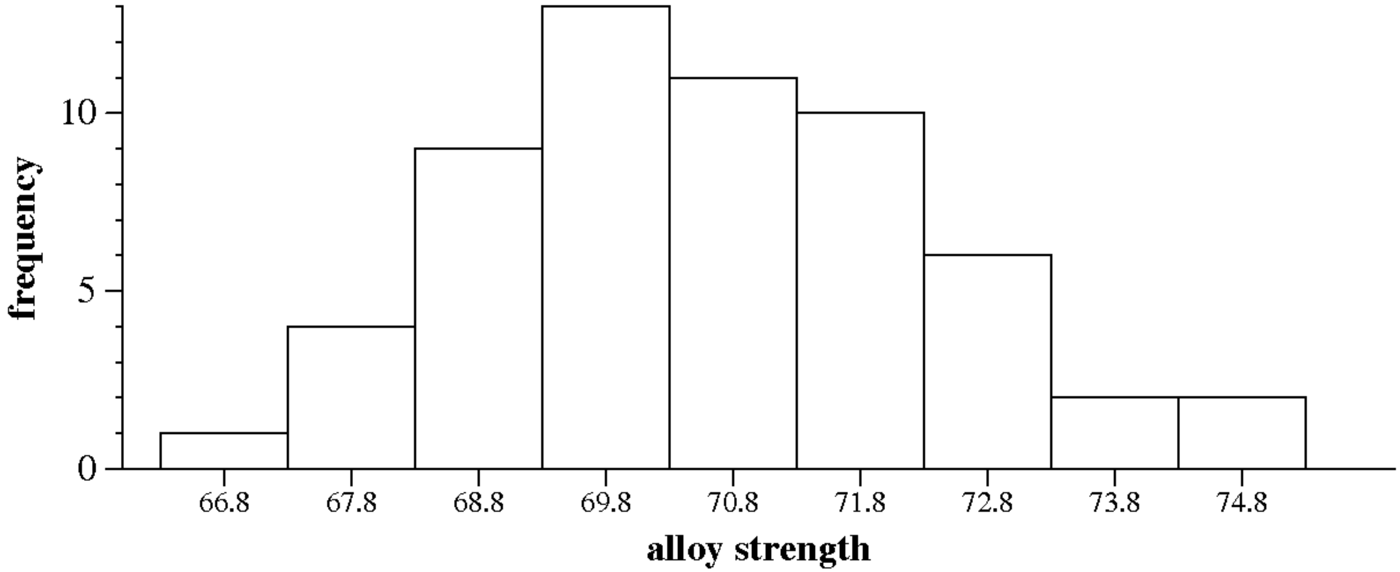
Class limits (or class boundaries) – the endpoints of each bin interval

Class mark – the midpoint of each bin

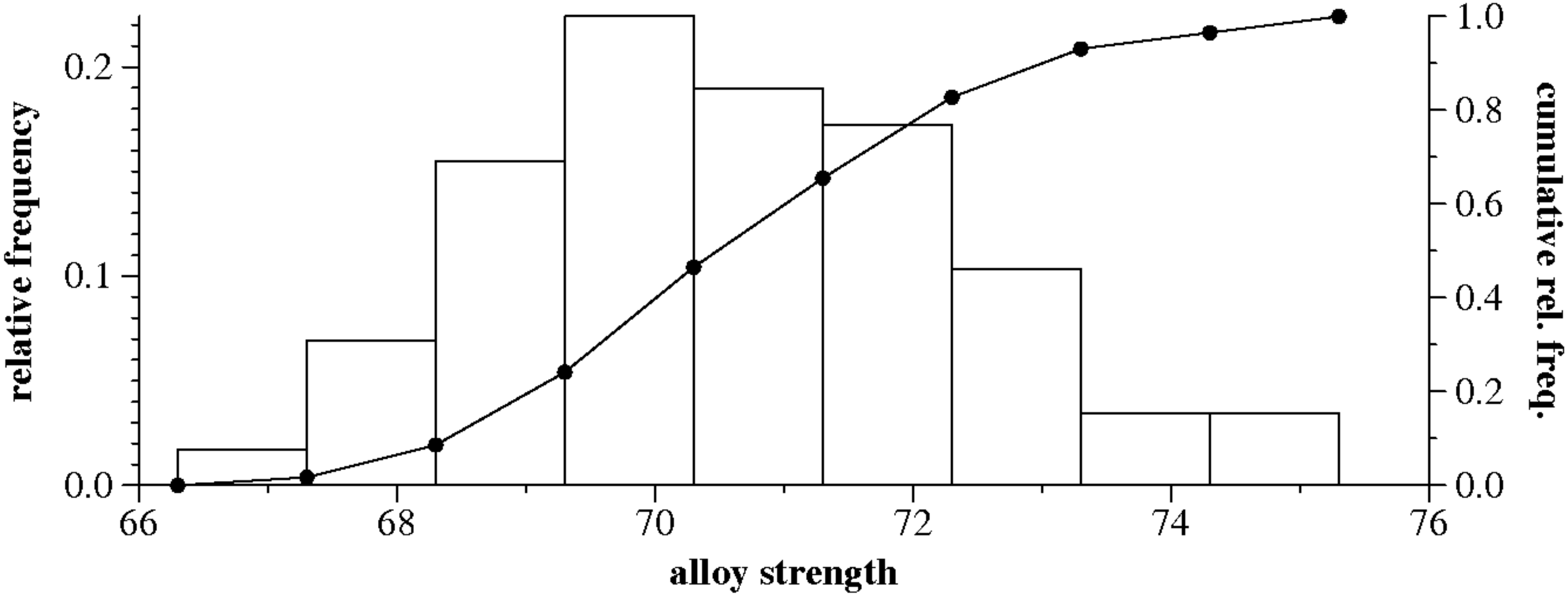
Class interval – the common width of each bin
= the distance between class marks

This notation is not necessarily standard.

Frequency Histograms (plotting frequency distributions)



Frequency plot using “class marks”



Relative frequency plot with cumulative distribution

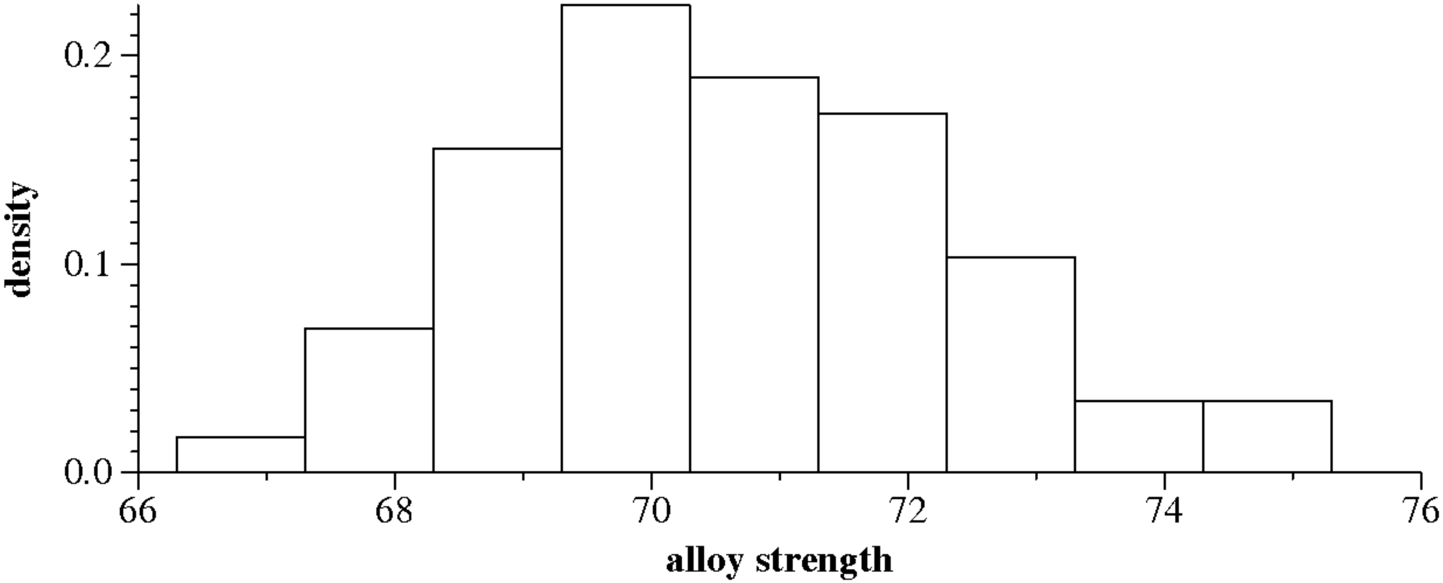
Density distributions

A density distribution is a frequency distribution for which

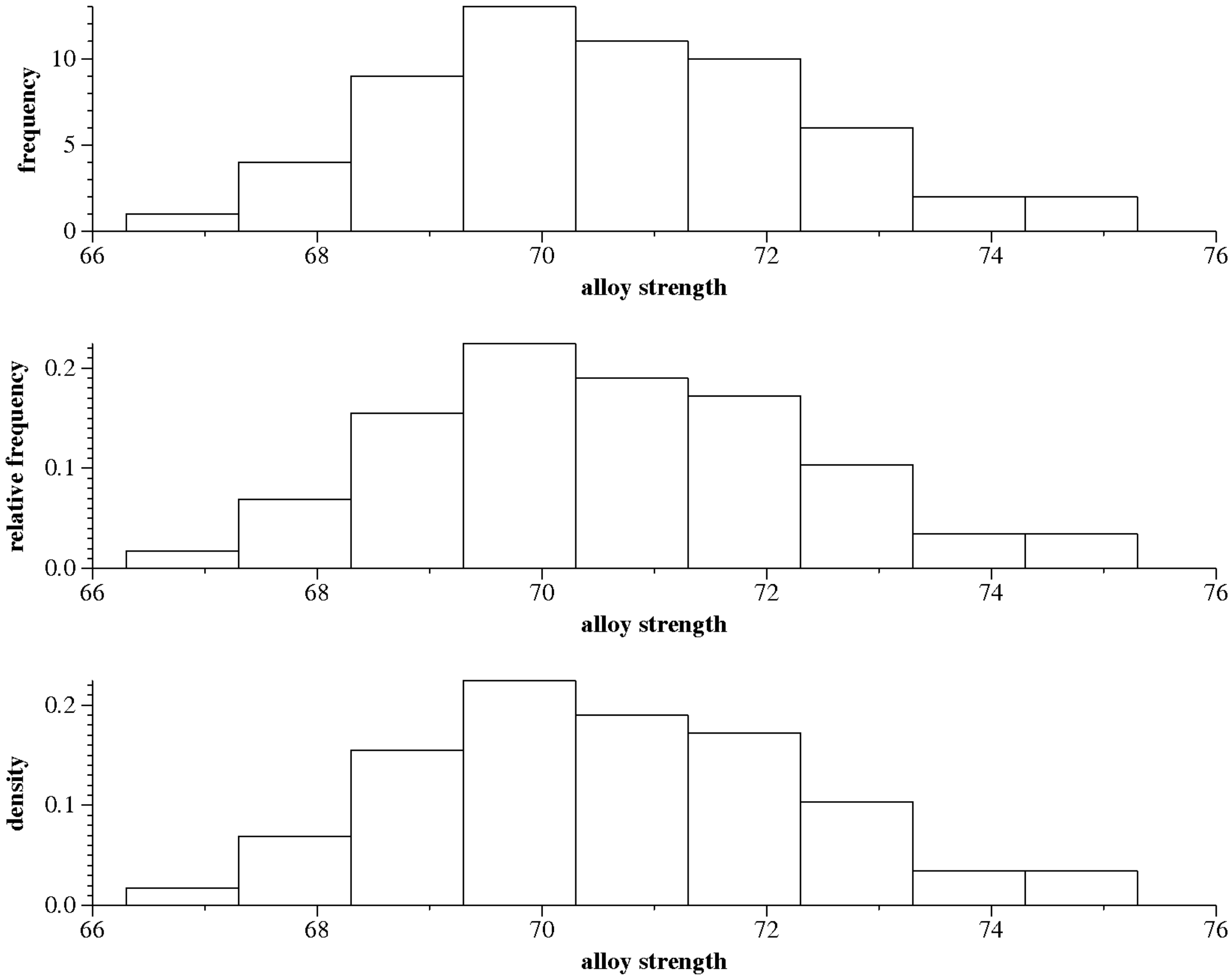
$$\text{height of bin } i = \frac{\text{relative frequency of bin } i}{\text{width of bin } i}$$

The area of each bin *i* = the relative frequency for bin *i*
The total area of a density distribution is always unity (= 1)

ONLY with density distributions are bins of different widths permissible
(although different bin widths should still be avoided when possible)



Note: if the bin limits remain the same, the shape of the distribution (frequency, relative frequency, density) remains unchanged.



Quartiles and Percentiles

In the same way that we use the **median** to find the center point of a data set

½ of the data values lie below the median, ½ of the values lie above

we can fine the **first quartile Q_1**, the data value that has ¼ of the data points lying below it (and 1 – ¼ = ¾ of the points lying above it). Since ¼ of the data corresponds to 25% of the data, Q_1 is also referred to as the **25th percentile $P_{0.25}$** of the data set.

We can generalize this to the 100 p'th percentile $P_{0.p}$ **,** the data value that has 100 p% of the data points lying below it (and 100 (1− p) % of the data points lying above it).

The **50'th percentile $P_{0.50}$** is known as the **second quartile Q_2** and is, in fact, identical to the **median** of the data set.

The **75'th percentile $P_{0.75}$** is known as the **third quartile Q_3** .

To compute percentile $P_{0.p}$ in a data set of n values:

1. order the data smallest to largest
2. compute the product ***n p***
   if *np* **is not** an integer, round it up to the next integer value and find the corresponding data point
   if *np* **is** an integer *k*, take the average of the *k*'th and (*k*+1)'st data values.

e.g. The $\boldsymbol{Q_1}$, $\boldsymbol{Q_3}$ and $\boldsymbol{Q_3}$ values for the alloy strength data set are:

$\boldsymbol{Q_1}$ (p = 0.25) $np = 0.25 \cdot 58 = 14.5$ round to 15. From the Stem-Leaf table, $\boldsymbol{Q_1} = x_{15} =$ 69.4

$\boldsymbol{Q_2}$ (p = 0.50) $np = 0.50 \cdot 58 = 29$ Average the 29’th and 30’th values. From the Stem-Leaf table, $\boldsymbol{Q_2} = (x_{29} + x_{30})/2 = (70.5 + 70.6)/2 = 70.55$ as previous computed for the median value of this data set

$\boldsymbol{Q_3}$ (p = 0.75) $np = 0.75 \cdot 58 = 43.5$ round to 44. From the Stem-Leaf table, $\boldsymbol{Q_3} = x_{44} =$ 71.8

As previously implied, the **range** of the data is

range = maximum − minimum

= $x_n - x_1$ when the data are ordered smallest to largest

The “middle half” of the data lies in the **interquartile range**

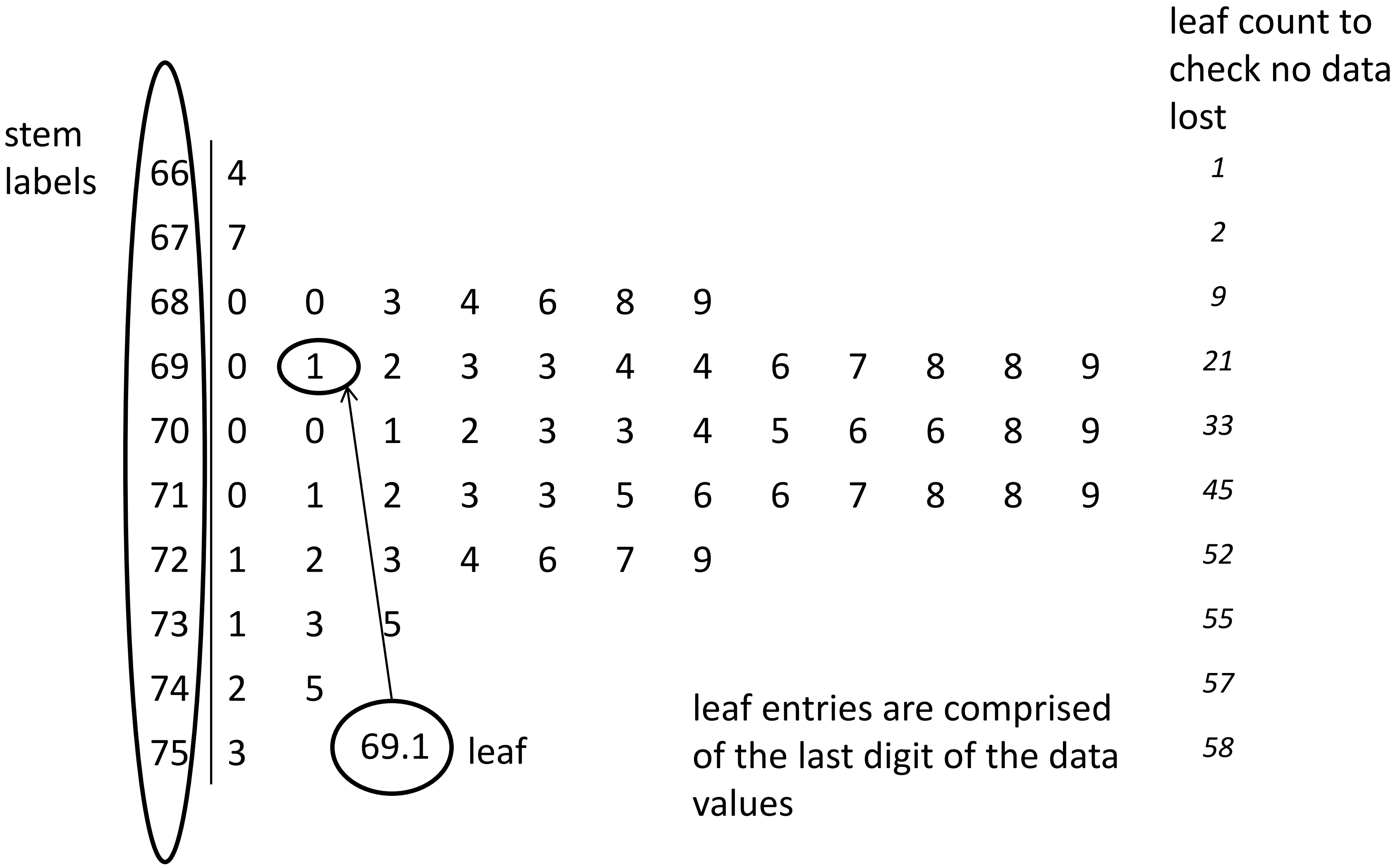
interquartile range = $\boldsymbol{Q_3 - Q_1}$

e.g. The range for the alloy strength data set is 75.3 – 66.4 = 8.9

The interquartile range is $\boldsymbol{Q_3 - Q_1}$ = 71.8 – 69.4 = 2.4

Unlike the dot diagram, which displays every value, frequency distributions suppress information on individual data values.

Stem-Leaf displays provide a tabular way of organizing the display of all data values



leaf count to check no data lost

stem labels

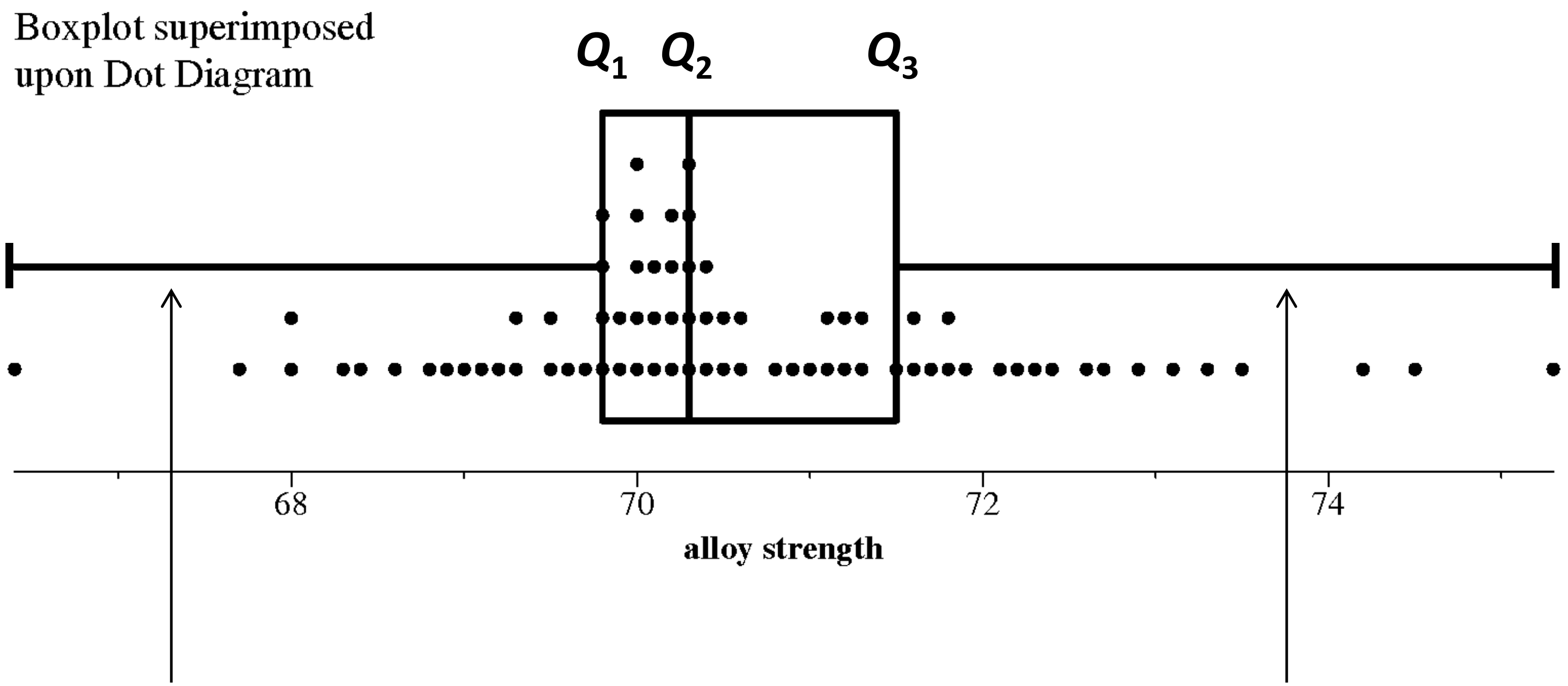
66	4												*1*
67	7												*2*
68	0	0	3	4	6	8	9						*9*
69	0	1	2	3	3	4	4	6	7	8	8	9	*21*
70	0	0	1	2	3	3	4	5	6	6	8	9	*33*
71	0	1	2	3	3	5	6	6	7	8	8	9	*45*
72	1	2	3	4	6	7	9						*52*
73	1	3	5										*55*
74	2	5											*57*
75	3												*58*

69.1 leaf

leaf entries are comprised of the last digit of the data values

Stem-Leaf displays are a useful way to analyze class grades, especially for computing median and quartile values

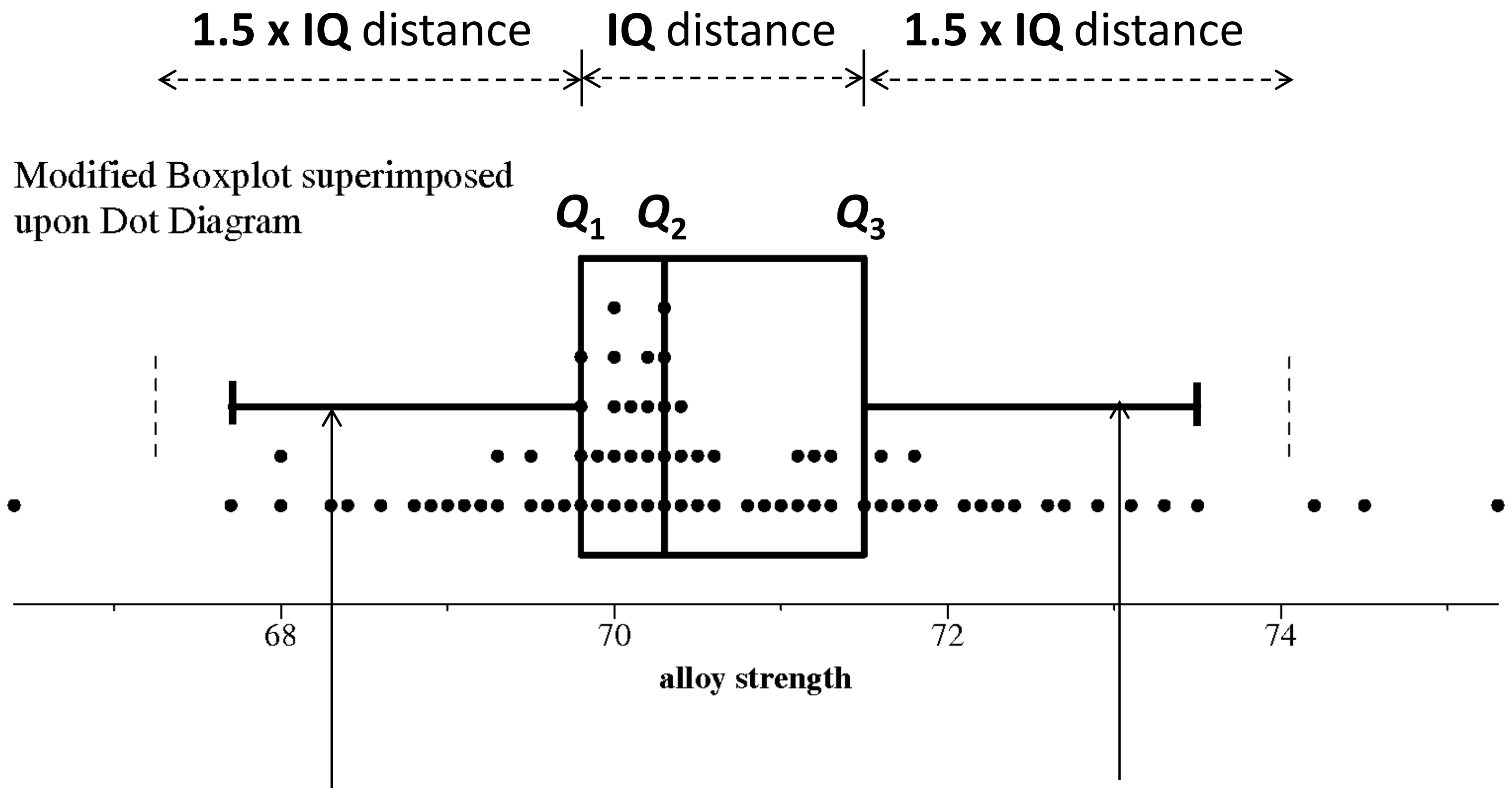
Boxplots: (aka box-whisker plot) plot quartile information



whisker extending to
minimum data value

whisker extending to
maximum data value

Boxplot variation:

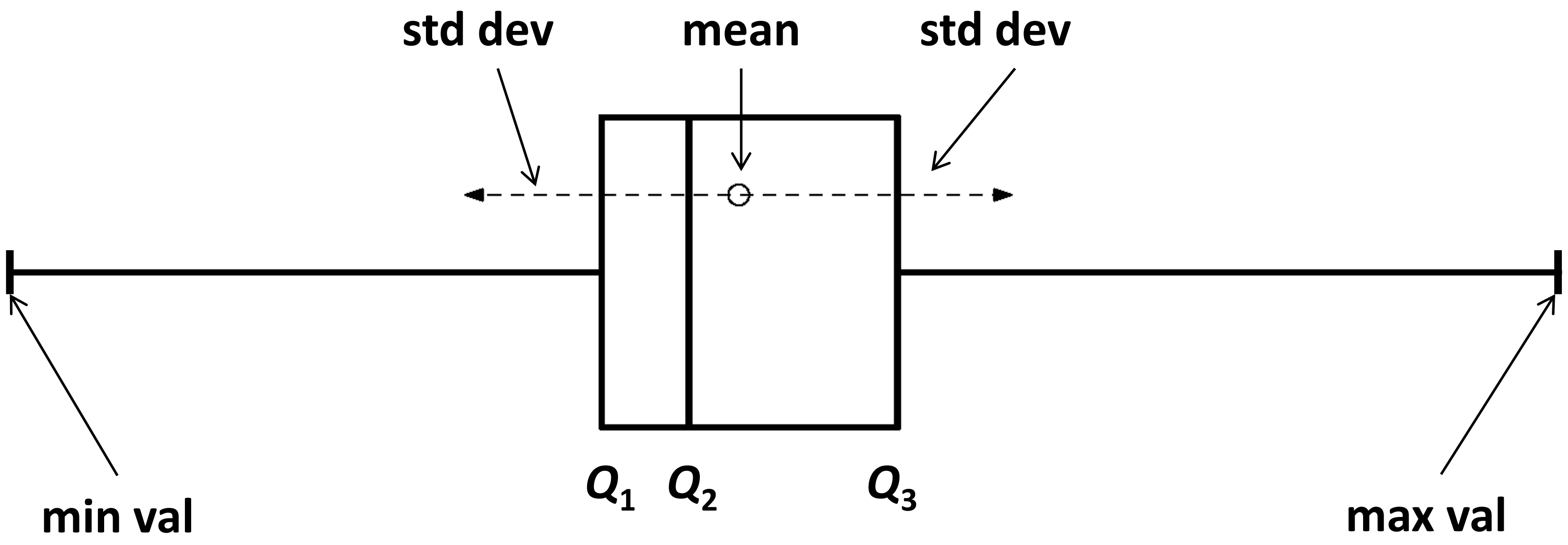


1.5 x IQ distance **IQ** distance **1.5 x IQ** distance

whisker extending to smallest data value lying above Q_1 − IQ distance

whisker extending to largest data value lying below Q_3 + IQ distance

Boxplot combined with sample mean and standard deviation



Chapter 2 summary

- dot diagrams
- distributions (histograms): bins, range, bin limits, bin marks, bin interval
- distributions of: frequency, relative frequency, percent frequency, cumulative frequency
- distribution graphs, density graphs
- stem-and-leaf displays
- data quantifiers: sample mean, sample variance, sample standard deviation
  Quartiles, percentiles, median, interquartile range
- boxplots