Lecture 1 : Basic Statistical Measures

Jonathan Marchini

October 11, 2004

In this lecture we will learn about

- different types of data encountered in practice
- different ways of plotting data to explore its structure/shape
- different summary measures of the shape of a given set of data

1 A brief introduction to Statistics

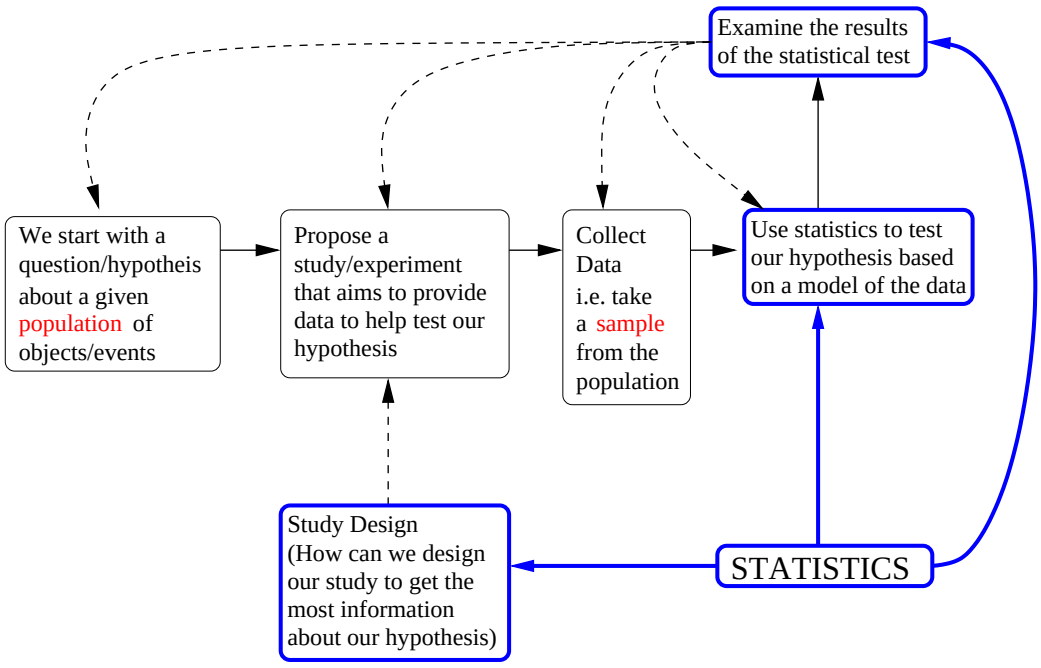


Figure 1: The scientific process and role of statistics in this process.

'The true logic of this world is the calculus of probabilities' - J C Maxwell

Statistics plays a very important role in all areas of science. Statistical models and methods allow us to take **samples** of data from our area of focus and answer questions about the **population** of objects or events that we are interested in studying.

Usually it is impossible to measure all the objects in the population we are interested in and so we have to make do with a sample. We study Statistics so we know how to make the best use of a given sample. Statistics can also inform us on how to go about designing an experiment and how to collect a sample of data. Figure 1 depicts the scientific process and the role statistics plays in this process. Notice how statistics enters into the process even before the data is collected and after a statistical test has been applied.

An Example Psychologists have long been interested in the relationship between stress and health. A focused question might involve the study of a specific psychological symptom and its impact on the health of the population. To assess whether the symptom is a good indicator of stress we need to measure the symptom and stress levels in a sample of individuals from the population. It is not immediately clear how we should go about collecting this sample, i.e. how we should design the study. Several questions present themselves

- Who should we include in the sample?
- How big should the sample be?
- Should we measure anything else?

To answer these questions we need to understand how we might go about analysing the data once collected, i.e. which statistical test or model we might use. Then we will be able to see how different study designs might affect the results. Thus statistics is involved right at the start of the process. Once a study design has been chosen and the data collected we will proceed to analyse the data. The process does not stop here. Close examination of the results may suggest several possible paths

- we might find that another test is more appropriate
- we might see that we need more data to reach a satisfactory conclusion
- we might find that our chosen study wasn't in fact the best of choices
- we might find the results suggest a more refined hypothesis and a further study

I hope this demonstrates the necessity of statistics in scientific work.

2 Types of data

The datasets that Psychologists and Human Scientists collect will usually consist of one or more observations on one or more "variables".

A **variable** is a property of an object or event that can take on different values.

For example, suppose we collect a dataset by measuring the hair colour, resting heart rate and score on an IQ test of every student in a class. The variables in this dataset would then simply be hair colour, resting heart rate and score on an IQ test, i.e. the variables are the properties that we measured/observed.

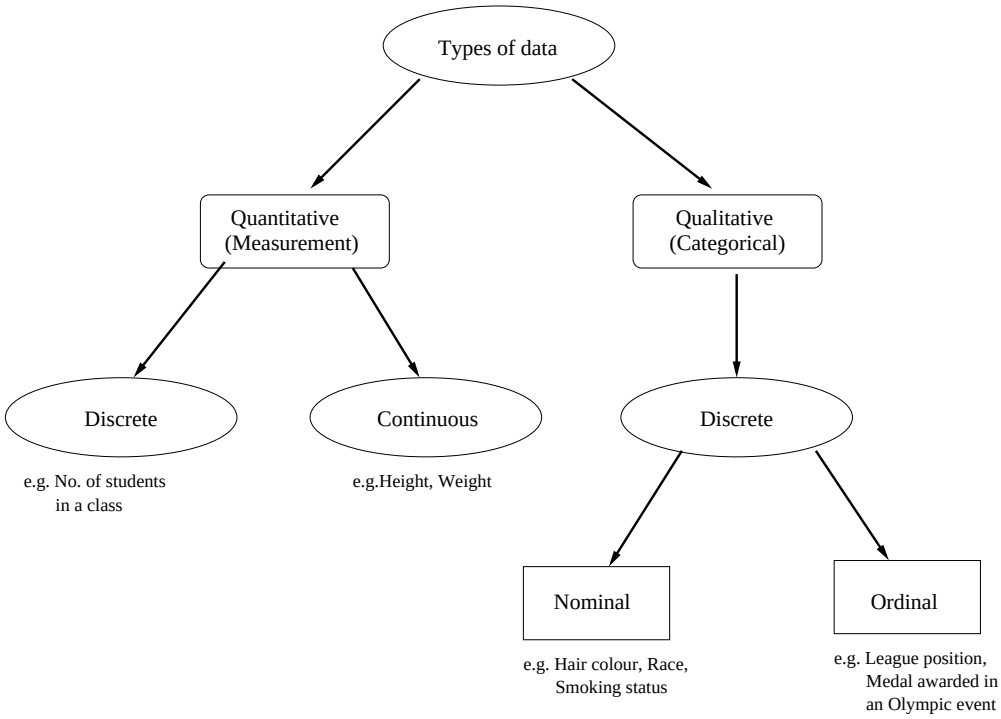


Figure 2: A summary of the different data types with some examples.

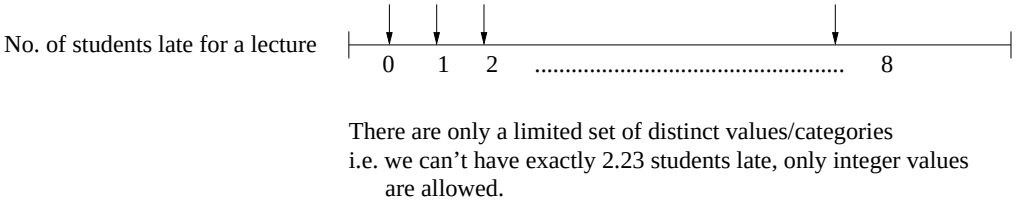
There are 2 main types of data/variable (see Figure 2)

(a) **Measurement / Quantitative Data** occur when we measure objects/events to obtain some number that reflects the quantitative trait of interest e.g. when we measure someones height or weight.

(b) **Categorical / Qualitative Data** occur when we assign objects into labelled groups or categories e.g. when we group people according to hair colour or race. Often the categories have a natural ordering. For example, in a survey we might group people depending upon whether they agree / neither agree or disagree / disagree with a statement. We call such ordered variables **Ordinal variables**. When the categories are unordered , e.g. hair colour, we have a **Nominal variable**.

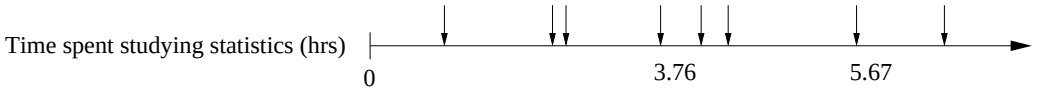
It is also useful to make the distinction between **Discrete** and **Continuous** variables (see Figure 3). Discrete variables, such as gender or number of peas in a pod, can take on only a limited set of values/categories. Continuous variables can take on (in theory) an unlimited set of values i.e. height.

Discrete Data



There are only a limited set of distinct values/categories
i.e. we can't have exactly 2.23 students late, only integer values are allowed.

Continuous Data



In theory there are an unlimited set of possible values!
There are no discrete jumps between possible values.

Figure 3: Examples of Discrete and Continuous variables.

3 Plotting Data

One of the most important stages in a statistical analysis can be simply to look at your data right at the start. By doing so you will be able to spot characteristic features, trends and outlying observations that enable you to carry out an appropriate statistical analysis. Also, it is a good idea to look at the results of your analysis using a plot. This can help identify if you did something that wasn't a good idea!

DANGER!! It is easy to become complacent and analyse your data without looking at it. This is a dangerous (and potentially embarrassing) habit to get into and can lead to false conclusions on a given study. The value of plotting your data cannot be stressed enough.

REMEMBER Data is messy! Use your massively parallel portable computer i.e. your brain, to untangle the mess by simply plotting your data.

Given that we accept the importance of plotting a dataset we now need the tools to do the job. There are several methods that can be used which we will illustrate with the help of the following dataset..

The baby-boom dataset

Forty-four babies (a new record) were born in one 24-hour period at the Mater Mothers' Hospital in Brisbane, Queensland, Australia, on December 18, 1997. For each of the 44 babies, The Sunday Mail recorded the time of birth, the sex of the child, and the birth weight in grams.

Whilst, we did not collect this dataset based on a specific hypothesis, if we wished we could use it to answer several questions of interest.

- Do girls weigh more than boys at birth?
- What is the distribution of the number of births per hour?
- Is birth weight related to the time of birth?
- Is gender related to the time of birth?
- If we assume that in the population a baby has an equal chance of being born a girl or a boy what is the probability that in random sample of 44 babies we observed the number of girls and boys in our sample.
- etc..

These are all questions that you will be able to test in a statistical manner by the end of this course. First though we can plot the data to view what the data might be telling us about these questions.

Time (mins since midnight)	Gender (1 = Girl, 2 = Boy)	Weight (g)
5	1	3837
64	1	3334
78	2	3554
115	2	3838
177	2	3625
245	1	2208
247	1	1745
262	2	2846
271	2	3166
428	2	3520
455	2	3380
492	2	3294
494	1	2576
549	1	3208
635	2	3521
649	1	3746
653	1	3523
693	2	2902
729	2	2635
776	2	3920
785	2	3690
846	1	3430
847	1	3480
873	1	3116
886	1	3428
914	2	3783
991	2	3345
1017	2	3034
1062	1	2184
1087	2	3300
1105	1	2383
1134	2	3428
1149	2	4162
1187	2	3630
1189	2	3406
1191	2	3402
1210	1	3500
1237	2	3736
1251	2	3370
1264	2	2121
1283	2	3150
1337	1	3866
1407	1	3542
1435	1	3278

Table 1: The Baby-boom dataset

3.1 Bar Charts

A Bar Chart is a useful method of summarising Categorical Data. We represent the counts/frequencies/percentages in each category by a bar. Figure 3.1 is a bar chart of gender for the baby-boom dataset. Notice that the bar chart has its axes clearly labelled.

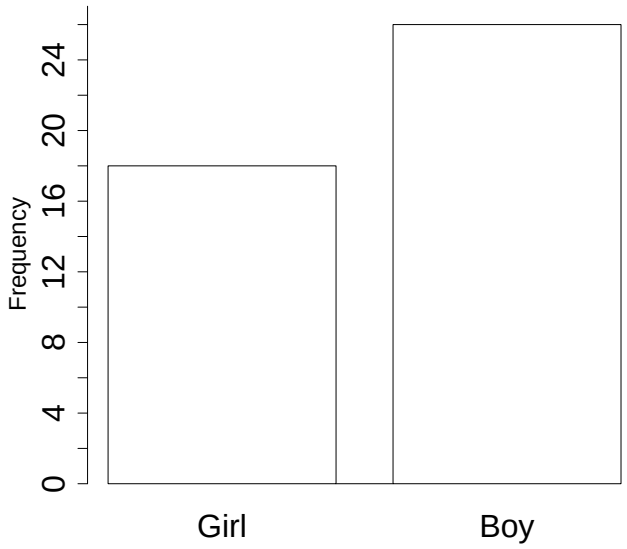


Figure 4: A Bar Chart showing the gender distribution in the Baby-boom dataset.

3.2 Histograms

An analogy

'A Bar Chart is to Categorical Data as a Histogram is to Measurement Data'

A histogram shows us the "distribution" of the numbers along some scale. A histogram is constructed in the following way

(i) Divide the measurements into categories

(ii) Determine the number of measurements within each category.

(iii) Draw a bar for each category whose heights represent the counts in each category.

The 'art' in constructing a histogram is how to choose the number of categories and the boundary points of the categories. For "small" datasets, it is often feasible to simply look at the values and decide upon sensible boundary points.

For the baby-boom dataset we can draw a histogram of the birth weights (Figure 5). To draw the histogram I found the smallest and largest values

smallest = 1745 largest = 4162

There are only 44 weights so I decided on 6 categories

1500-2000 2000-2500 2500-3000 3000-3500 3500-4000 4000-4500

Using these categories works well, the histogram shows us the shape of the distribution and we notice that distribution has an extended left ‘tail’.

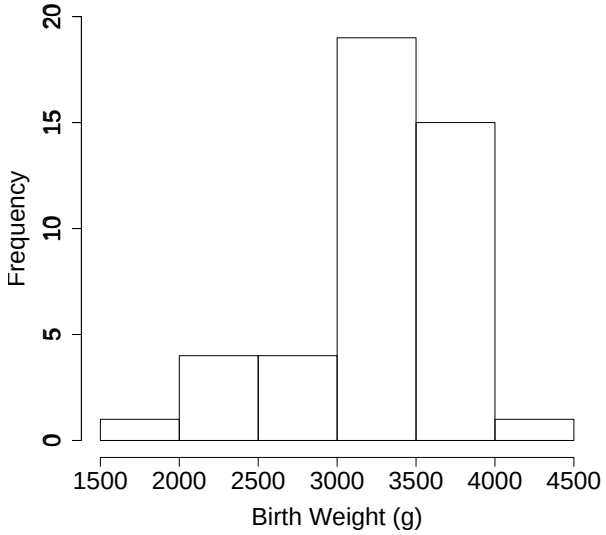


Figure 5: A Histogram showing the birth weight distribution in the Baby-boom dataset.

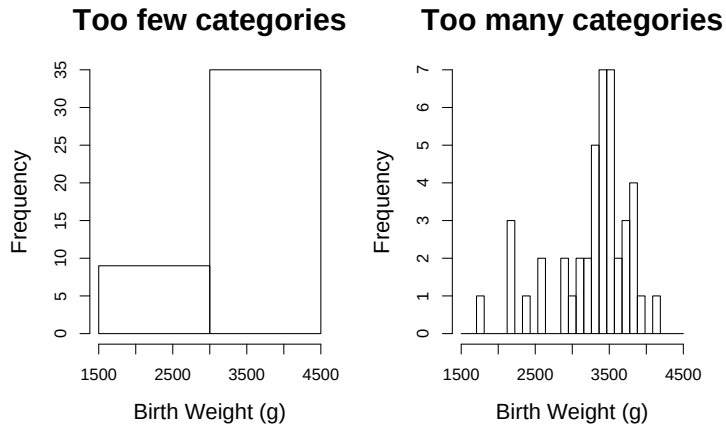


Figure 6: Histograms with too few and too many categories respectively.

Too few categories and the details are lost. Too many categories and the overall shape is obscured by too many details (see Figure 6).

3.3 Cumulative and Relative Cumulative Frequency Plots and Curves

A **cumulative frequency plot** is very similar to a histogram. In a cumulative frequency plot the height of the bar in each interval represents the total count of observations within interval and *lower than* the interval (see Figure 7)

In a **cumulative frequency curve** the cumulative frequencies are plotted as points at the upper boundaries of each interval. It is usual to join up the points with straight lines (see Figure 7).

Relative cumulative frequencies are simply cumulative frequencies divided by the total number of observations (so relative cumulative frequencies always lie between 0 and 1). Thus **relative cumulative frequency plots and curves** just use relative cumulative frequencies rather than cumulative frequencies. Such plots are useful when we wish to compare two or more distributions on the same scale.

Consider the histogram of birth weight shown in Figure 5. The frequencies, cumulative frequencies and relative cumulative frequencies of the intervals are given in Table 2.

Interval	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500
Frequency	1	4	4	19	15	1
Cumulative Frequency	1	5	9	28	43	44
Relative Cumulative Frequency	0.023	0.114	0.205	0.636	0.977	1.0

Table 2: Frequencies and Cumulative frequencies for the histogram in Figure 5.

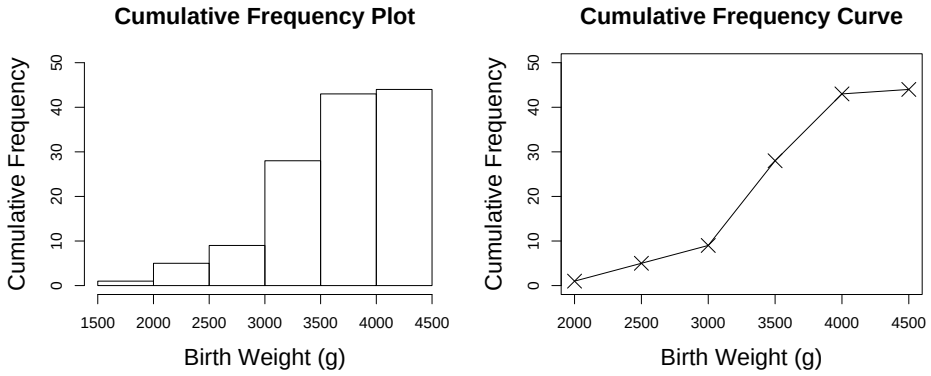


Figure 7: Cumulative frequency curve and plot of birth weights for the baby-boom dataset.

3.4 Dot plot

A Dot Plot is a simple and quick way of visualising a dataset. This type of plot is especially useful if data occur in groups and you wish to quickly visualise the differences between the groups. For example, Figure 8 shows a dot plot of birth weights grouped by gender for the baby-boom dataset. The plot suggests that girls may be lighter than boys at birth.

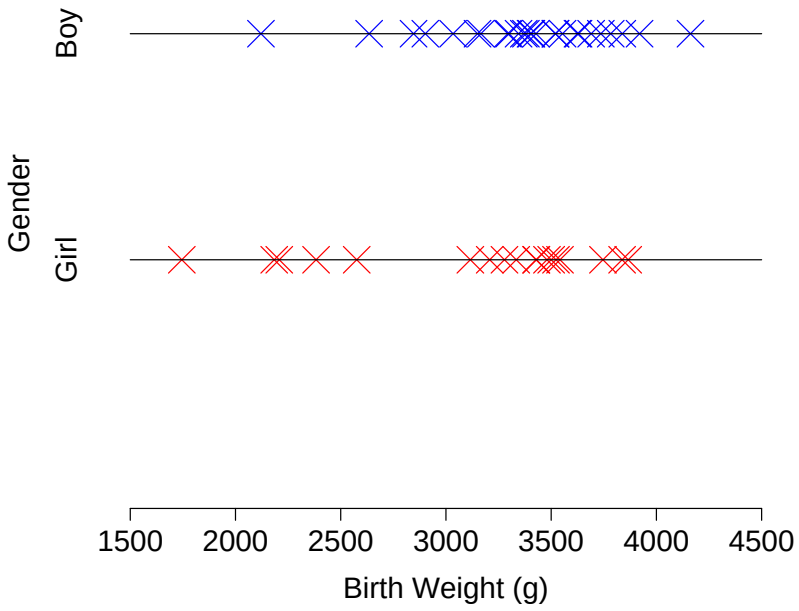


Figure 8: A Dot Plot showing the birth weights grouped by gender for the baby-boom dataset.

3.5 Scatter Plots

Scatter plots are useful when we wish to visualise the relationship between two measurement variables.

To draw a scatter plot we

- Assign one variable to each axis.
- Plot one point for each pair of measurements.

For example, we can draw a scatter plot to examine the relationship between birth weight and time of birth (Figure 9). The plot suggests that there is little relationship between birth weight and time of birth.

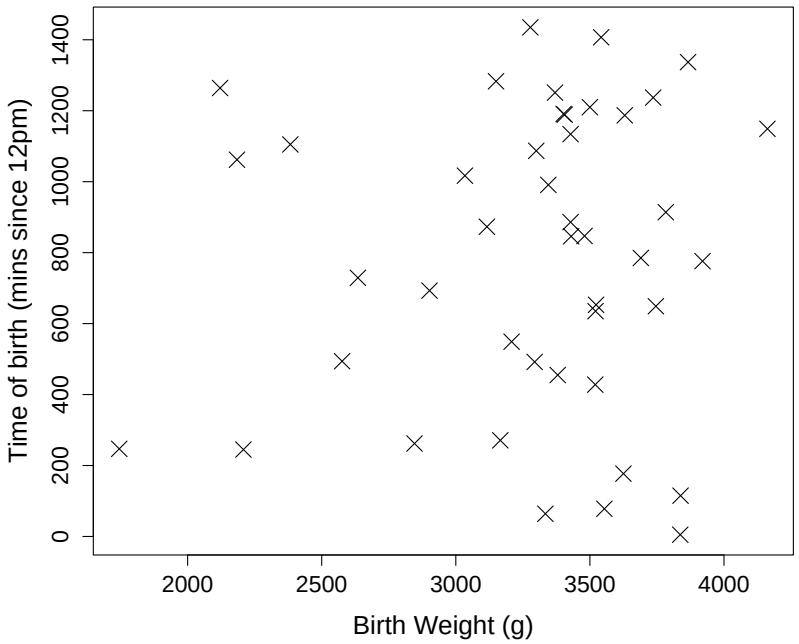


Figure 9: A Scatter Plot of birth weights versus time of birth for the baby-boom dataset.

3.6 Box Plots

Box Plots are probably the most sophisticated type of plot we will consider. To draw a Box Plot we need to know how to calculate certain "summary measures" of the dataset covered in the next section. We return to discuss Box Plots in Section 5.

4 Summary Measures

In the previous section we saw how to use various graphical displays in order to explore the structure of a given dataset. From such plots we were able to observe the general shape of the "distribution" of a given dataset and compare visually the shape of two or more datasets.

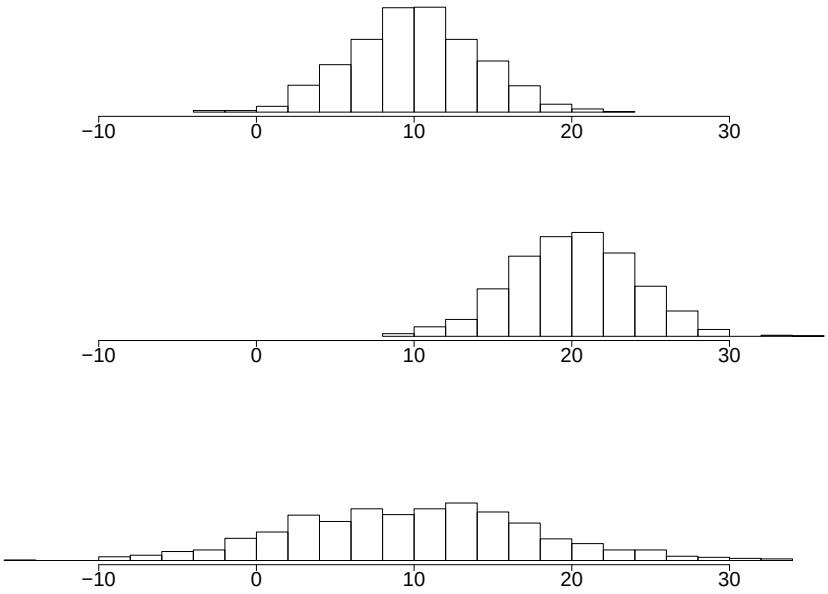


Figure 10: Comparing shapes of histograms

Consider the histogram in Figure 10. Comparing the first and second histograms we see that the distributions have the same shape or spread but that the center of the distribution is different. Roughly, by eye, the centers differ in value by about 10. Comparing first and third histograms we see that the distributions seem to have roughly the same center but that the data plotted in the third are more spread out than in the first. Obviously, comparing second and the third we observe differences in both the center and the spread of the distribution.

In the previous paragraph we made simple observations and used the expressions 'seem to have have roughly the same' and 'more spread out'. To do better than these imprecise statements we can calculate measures of

(i) the 'center' point of the data.

(ii) the 'spread' of the data.

These two characteristics of a set of data (the center and spread) are the simplest measures of its shape. Once calculated we can make precise statements about how the centers or spreads of two datasets differ. Later we will learn how to go a stage further and 'test' whether two variables have the same center point.

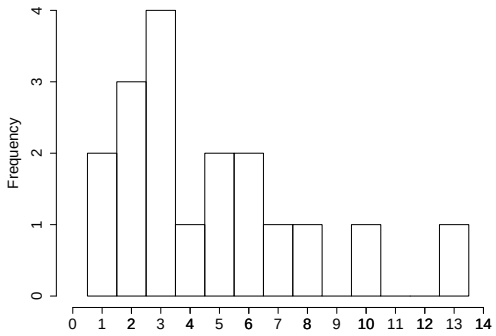
4.1 Measures of location (Measuring the center point)

There are 3 main measures of the center of a given set of (measurement) data

(i) **The Mode** of a set of numbers is simply the most common value e.g. the mode of the following set of numbers

$$1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6, 7, 8, 10, 13$$

is 3. If we plot a histogram of this data



we see that the mode is the peak of the distribution and is a reasonable representation of the center of the data. If we wish to calculate the mode of continuous data one strategy is to group the data into adjacent intervals and choose the modal interval i.e. draw a histogram and take the modal peak. This method is sensitive to the choice of intervals and so care should be taken so that the histogram provides a good representation of the shape of the distribution.

The Mode has the advantage that it is always a score that actually occurred and can be applied to nominal data, properties not shared by the median and mean. A disadvantage of the mode is that there may two or more values that share the largest frequency. In the case of two modes we would report both and refer to the distribution as *bimodal*.

(ii) **The Median** can be thought of as the 'middle' value i.e. the value for which 50% of the data fall below when arranged in numerical order. For example, consider the numbers

$$15, 3, 9, 21, 1, 8, 4,$$

When arranged in numerical order

$$1, 3, 4, \boxed{8}, 9, 15, 21$$

we see that the median value is 8. If there were an even number of scores e.g.

$$1, 3, \boxed{4,8}, 9, 15$$

then we take the midpoint of the two middle values. In this case the median is $(4 + 8)/2 = 6$. In general, if we have N data points then the **median location** is defined as follows:

$$\text{Median Location} = \frac{(N+1)}{2}$$

For example, the median location of 7 numbers is $(7+1)/2 = 4$ and the median of 8 numbers is $(8+1)/2 = 4.5$ i.e. between observation 4 and 5 (when the numbers are arranged in order).

A major advantage of the median is the fact that it is unaffected by extreme scores (a point it shares with the mode). We say the median is **resistant** to outliers. For example, the median of the numbers

$$1, 3, 4, \boxed{8}, 9, 15, 99999$$

is still 8. This property is very useful in practice as *outlying* observations can and do occur (Data is messy remember!).

(iii) **The Mean** of a set of scores is the sum[1] of the scores divided by the number of scores. For example, the mean of

$$1, 3, 4, 8, 9, 15 \quad \text{is} \quad \frac{1+3+4+8+9+15}{6} = 6.667 \quad \text{(to 3 dp)}$$

In mathematical notation, the mean of a set of n numbers $x_1, \ldots, x_n$ is denoted by $\bar{x}$ where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{or} \qquad \bar{x} = \frac{\sum x}{n} \qquad \text{(in the formula book)}$$

See the appendix for a brief description of the summation notation $(\sum)$

The mean is the most widely used measure of location. Historically, this is because statisticians can write down equations for the mean and derive nice theoretical properties for the mean, which are much harder for the mode and median. A disadvantage of the mean is that it is not resistant to outlying observations. For example, the mean of

$$1, 3, 4, 8, 7, 15, 99999$$

is 14323.57, whereas the median (from above) is 8.

Sometimes discrete measurement data are presented in the form of a frequency table in which the frequencies of each value are given. Remember, the mean is the sum of the data divided by the number of observations. To calculate the sum of the data we simply multiply each value by its frequency and sum. The number of observations is calculated by summing the frequencies.

For example, consider the following frequency table
We calculate the sum of the data as

$$(2 \times 1) + (4 \times 2) + (6 \times 3) + (7 \times 4) + (4 \times 5) + (1 \times 6) = 82$$

[1] The total when we add them all up

Data (x)	1	2	3	4	5	6
Frequency (f)	2	4	6	7	4	1

Table 3: A frequency table.

and the number of observations as

$$2 + 4 + 6 + 7 + 4 + 1 = 24$$

The the mean is given by

$$\bar{x} = \frac{82}{24} = 3.42 \quad \text{(2 dp)}$$

In mathematical notation the formula for the mean of frequency data is given by

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \qquad \text{or} \qquad \bar{x} = \frac{\sum fx}{\sum f}$$

4.1.1 The relationship between the mean, median and mode

In general, these three measures of location will differ but for certain datasets with characteristic 'shapes' we will observe simple patterns between the three measures (see Figure 11).

(a) If the distribution of the data is **symmetric** then the mean, median and mode will be very close to each other.

$$\text{MODE} \approx \text{MEDIAN} \approx \text{MEAN}$$

(b) If the distribution is **positively skewed** i.e. the data has an extended right tail, then

$$\text{MODE} < \text{MEDIAN} < \text{MEAN}$$

(c) If the distribution is **negatively skewed** i.e. the data has an extended left tail, then

$$\text{MEAN} < \text{MEDIAN} < \text{MODE}$$

4.1.2 The mid-range

There is actually a fourth measure of location that can be used (but rarely is). The **Mid-Range** of a set of data is half way between the smallest and largest observation i.e. half the **range** of the data. For example, the mid-range of

$$1, 3, 4, 8, 9, 15, 21$$

is (1 + 21) / 2 = 11. The mid-range is rarely used because it is not resistant to outliers and by using only 2 observations from the dataset it takes no account of how spread of the data.

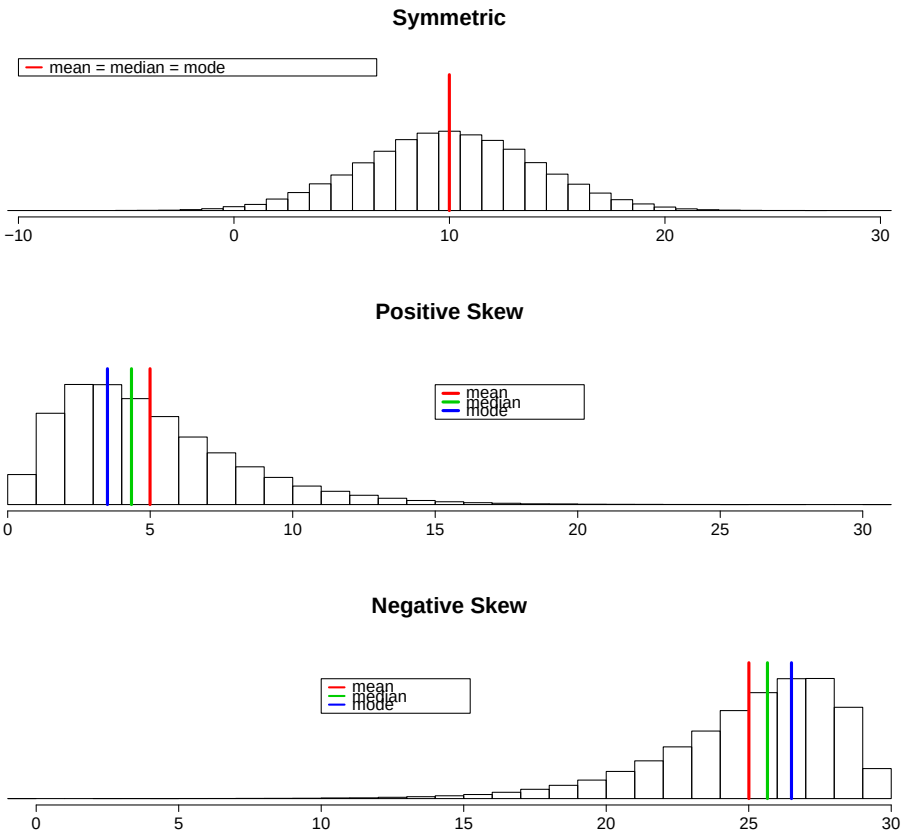


Figure 11: The relationship between the mean, median and mode.

4.2 Measures of dispersion (Measuring the spread)

4.2.1 The Interquartile Range (IQR) and Semi-Interquartile Range (SIQR)

The Interquartile Range (IQR) of a set of numbers is defined to be the range of the middle 50% of the data (see Figure 12). The Semi-Interquartile Range (SIQR) is simply half the IQR.

We calculate the IQR in the following way:

- Calculate the 25% point (**1st quartile**) of the dataset. The location of the 1st quartile is defined to be the $\left(\frac{N+1}{4}\right)$th data point.
- Calculate the 75% point (**3rd quartile**) of the dataset. The location of the 3rd quartile is defined to be the $\left(\frac{3(N+1)}{4}\right)$th data point[2].
- Calculate the IQR as

$$\text{IQR} = \text{3rd quartile - 1st quartile}$$

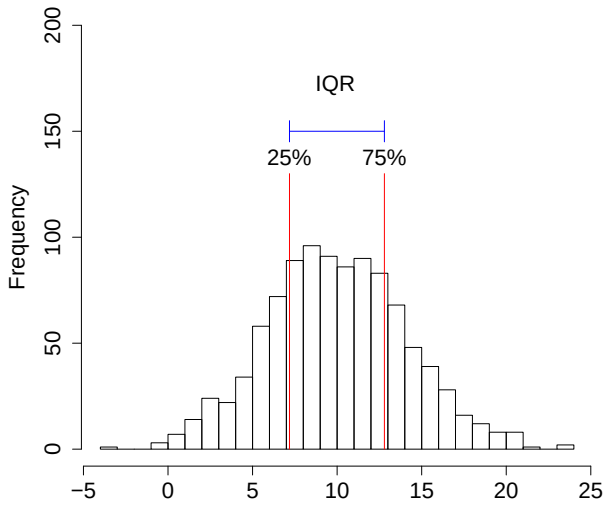


Figure 12: The Interquartile Range.

Example 1 Consider the set of 11 numbers (which have been arranged in order)

$$10, 15, 18, 33, 34, 36, 51, 73, 80, 86, 92$$

The 1st quartile is the $\frac{(11+1)}{4} = 3$rd data point = 18
The 3rd quartile is the $\frac{3(11+1)}{4} = 9$th data point = 80

$$\Rightarrow \text{IQR} = 80 - 18 = 62$$
$$\Rightarrow \text{SIQR} = 62 / 2 = 31.$$

[2]The **2nd quartile** is the 50% point of the dataset i.e. the median.

4.2.2 The Mean Deviation

To measure the spread of a dataset it seems sensible to use the 'deviation' of each data point from the mean of the distribution (see Figure 13). The deviation of each data point from the mean is simply the data point minus the mean.

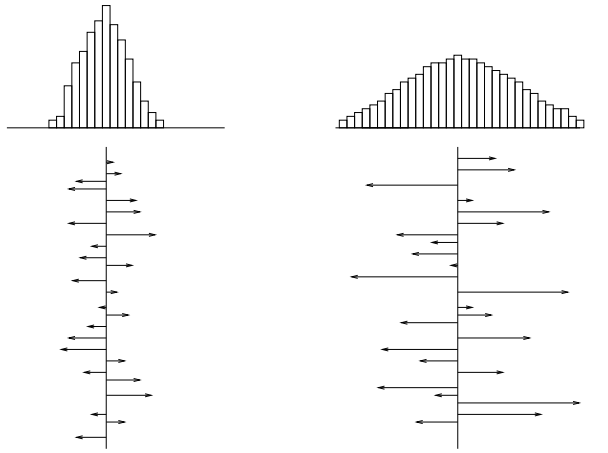


Figure 13: The relationship between spread and deviations..

For example, for deviations of the set of numbers

$$10, 15, 18, 33, 34, 36, 51, 73, 80, 86, 92$$

which have mean 48 are given in Table 4.

Data x	Deviations $x - \bar{x}$	|Deviations| $\lvert x - \bar{x} \rvert$	Deviations^2 $(x - \bar{x})^2$
10	10 - 48 = -38	38	1444
15	15 - 48 = -33	33	1089
18	18 - 48 = - 30	30	900
33	33 - 48 = -15	15	225
34	34 - 48 = -14	14	196
36	36 - 48 = -12	12	144
51	51 - 48 = 3	3	9
73	73 - 48 = 25	25	625
80	80 - 48 = 32	32	1024
86	86 - 48 = 38	38	1444
92	92 - 48 = 44	44	1936
Sum = 528 $\sum x = 528$	Sum = 0 $\sum(x - \bar{x}) = 0$	Sum = 284 $\sum \lvert x - \bar{x} \rvert = 284$	Sum = 9036 $\sum(x - \bar{x})^2 = 9036$

Table 4: Deviations, Absolute Deviations and Squared Deviations.

The Mean Deviation of a set of numbers is simply mean of deviations.

In mathematical notation this is written as

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

At first this sounds like a good way of assessing the spread since you might think that large spread gives rise to larger deviations and thus a larger mean deviation. In practice, the mean deviation is *always* zero. The positive and negative deviations cancel each other out exactly. Even so, the deviations still contain useful information about the spread, we just have to find a way of using the deviations in a sensible way.

4.2.3 Mean Absolute Deviation (MAD)

We solve the problem of the deviations summing to zero by considering the *absolute values* of the deviations. The absolute value of a number is the value of that number with any minus sign removed, e.g. $|-3| = 3$. We then measure spread using the mean of the absolute deviations, denoted (MAD).

This can be written in mathematical notation as

$$\frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n} \quad \text{or} \quad \frac{\sum|x - \bar{x}|}{n}$$

Note The second formula is just a short hand version of the first (See the Appendix).

We calculate the MAD in the following way (see Table 4 for an example)

- Calculate the mean of the data, $\bar{x}$
- Calculate the deviations by subtracting the mean from each value, $x - \bar{x}$
- Calculate the absolute deviations by removing any minus signs from the deviations, $|x - \bar{x}|$.
- Sum the absolute deviations, $\sum|x - \bar{x}|$.
- Calculate the MAD by dividing the sum of the absolute deviations by the number of data points, $\sum|x - \bar{x}|/n$.

From Table 4 we see that the sum of the absolute deviations of the numbers in Example 1 is 284 so

$$\text{MAD} = \frac{284}{11} = 25.818 \quad \text{(to 3dp)}$$

4.2.4 The Sample Variance (s^2) and Population Variance (σ^2)

Another way to ensure the deviations don't sum to zero is to look at the *squared* deviations i.e. the square of a number is always positive. Thus another way of measuring the spread is to consider the mean of the squared deviations, called the *variance*

If our dataset consists of the whole population (a rare occurrence) then we can calculate the population variance σ^2 (said 'sigma squared') as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad \text{or} \qquad \sigma^2 = \frac{\sum(x - \bar{x})^2}{n}$$

When we just have a sample from the population (most of the time) we can calculate the sample variance s^2 as

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

Note We divide by $n-1$ when calculating the sample variance as then s^2 is a 'better estimate' of the population variance σ^2 than if we had divided by n. We will see why later.

For frequency data (see Table) the formula is given by

$$s^2 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i - 1} \qquad \text{or} \qquad s^2 = \frac{\sum f(x - \bar{x})^2}{\sum f - 1}$$

We can calculate s^2 in the following way (see Table 4)

- Calculate the deviations by subtracting the mean from each value, $x - \bar{x}$
- Calculate the squared deviations, $(x - \bar{x})^2$.
- Sum the squared deviations, $\sum(x - \bar{x})^2$.
- Divide by $n-1$, $\sum(x - \bar{x})^2/(n-1)$.

From Table 4 we see that the sum of the squared deviations of the numbers in Example 1 is 9036 so

$$s^2 = \frac{9036}{11-1} = 903.6$$

4.2.5 The Sample Standard Deviation (s) and Population Standard Deviation (σ)

Notice how the sample variance in Example 1 is much higher than the SIQR and the MAD.

$$\text{SIQR} = 31 \qquad \text{MAD} = 25.818 \qquad s^2 = 903.6$$

This happens because we have squared the deviations transforming them to a quite different scale. We can recover the scale of the original data by simply taking the square root of the sample (population) variance.

Thus we define the sample standard deviation s as

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

and we define the population standard deviation σ as

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

Returning to Example 1 the sample standard deviation is

$$s = \sqrt{903.6} = 30.05 \quad \text{(to 2dp)}$$

which is comparable with the SIQR and the MAD.

5 Box Plots

As we mentioned earlier a Box Plot (sometimes called a Box-and-Whisker Plot) is a relatively sophisticated plot that summarises the distribution of a given dataset.

A box plot consists of three main parts

1. A box that covers the middle 50% of the data. The edges of the box are the 1st and 3rd quartiles. A line is drawn in the box at the median value.
2. Whiskers that extend out from the box to indicate how far the data extend either side of the box. The whiskers should extend no further than 1.5 times the length of the box, i.e. the maximum length of a whisker is 1.5 times the IQR.
3. All points that lie outside the whiskers are plotted individually as outlying observations.

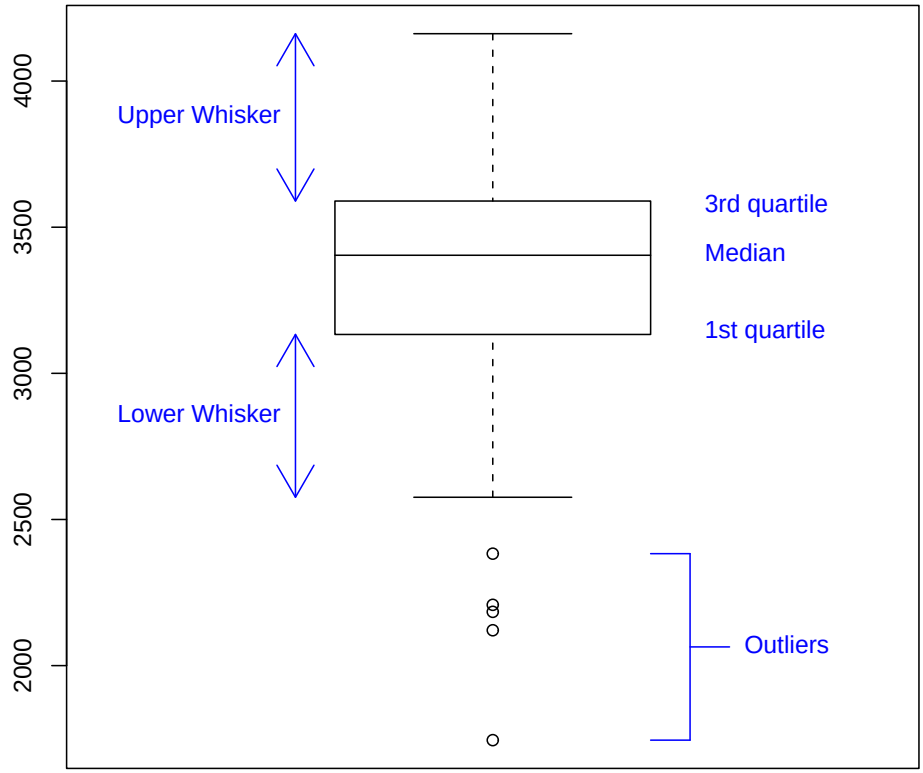


Figure 14: A Box Plot of birth weights for the baby-boom dataset showing the main points of plot.

Plotting box plots of measurements in different groups side by side can be illustrative. For example, Figure 15 shows box plots of birth weight for each gender side by side and indicates that the distributions have quite different shapes.

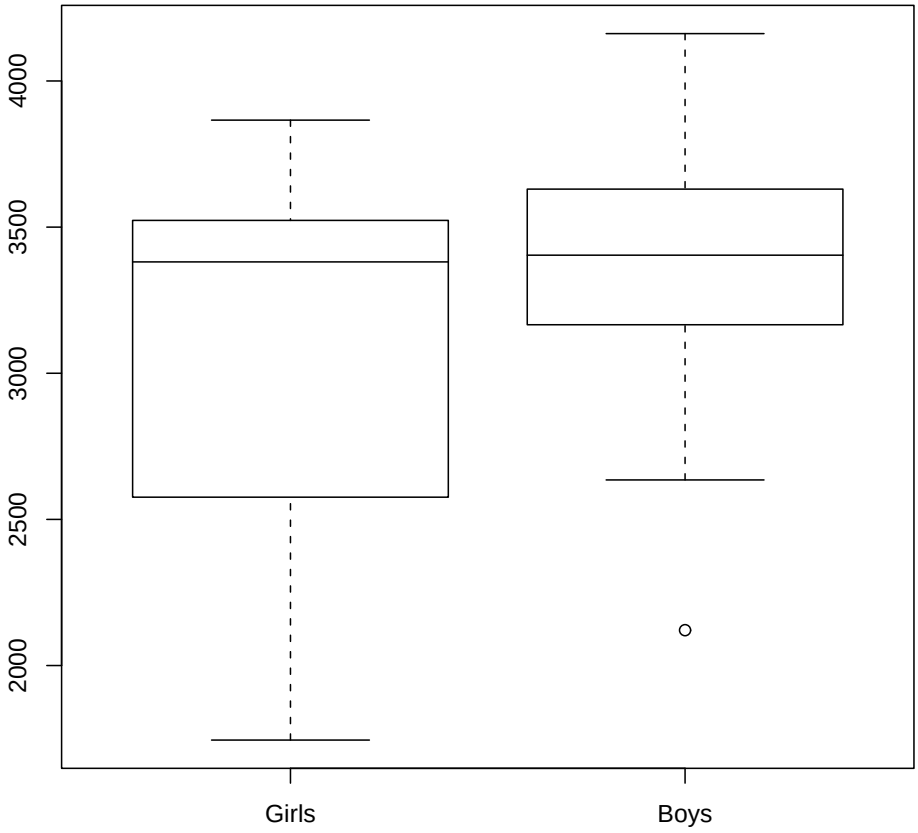


Figure 15: A Box Plot of birth weights by gender for the baby-boom dataset.

6 Appendix

6.1 Mathematical notation for variables and samples

Mathematicians are lazy. They can't be bothered to write everything out in full so they have invented a language/notation in which they can express what they mean in a compact, quick to write down fashion. This is a good thing. We don't have to study maths every day to be able to use a bit of the language and make our lives easier. For example, suppose we are interested in comparing the resting heart rate of 1st year Psychology and Human Sciences students. Rather than keep on referring to variables 'the resting heart rate of 1st year Psychology students' and 'the resting heart rate of 1st year Human Sciences students' we can simple denote

X = the resting heart rate of 1st year Psychology students
Y = the resting heart rate of 1st year Human Sciences students

and refer to the variables X and Y instead.

In general, we use capital letters to denote variables. If we have a sample of a variable the we use small letters to denote the sample. For example, if we go and measure the resting heart rate of 1st year Psychology and Human Sciences students in Oxford we could denote the p measurements on Psychologists as

$$x_1, x_2, x_3, \ldots, x_p$$

and the h measurements on Human Scientists as

$$y_1, y_2, y_3, \ldots, y_h$$

6.2 Summation notation

One of the most common letters in the Mathematicians alphabet is the Greek letter **sigma** $(\sum)$, which is used to denote summation. It translates to "add up what follows". Usually, the limits of the summation are written below and above the symbol. So,

$$\sum_{i=1}^{5} x_i \qquad \text{reads "add up the } x_i\text{s from } i = 1 \text{ to } i = 5\text{"}$$

or

$$\sum_{i=1}^{5} x_i = (x_1 + x_2 + x_3 + x_4 + x_5)$$

If we have some actual data then we know the values of the x_is

$$x_1 = 3 \quad x_2 = 6 \quad x_3 = 1 \quad x_4 = 7 \quad x_5 = 6$$

and we can calculate the sum as

$$\sum_{i=1}^{5} x_i = (3 + 2 + 1 + 7 + 6) = 19$$

If the limits of the summation are obvious within context the the notation is often abbreviated to

$$\sum x = 19$$