

# Some big ideas in statistics

## 1. We learn through repeated observation

Every single child knows that if they drop a ball it will fall to the floor. They learn to expect a dropped ball to fall to the floor long before they ever hear anything of Isaac Newton or the law of gravity. Children (and people in general) learn basic information through repeated observation. Children can be confident that a dropped ball will fall to the floor because they have seen it happen in a similar way countless times before.

When we want to answer new questions on how the world works, the best strategy is often to try to get lots and lots of observations related to that question. If we want to know if rainfall has increased on a particular town in the last 50 years we look to compare records (observations) of rainfall 50 years ago with current records. If we want to know how safe aeroplanes are we look to see how many crashes there were observed across a large number of flights. If we want to know how long a particular type of light bulb will run before it blows, we wouldn't just test a single bulb – that one may have been unusual or faulty or different from normal in some way. We would test a large number of bulbs and see how long these bulbs normally run for.

## 2. Repetition generates data

If we measure how long each light bulb lasts for we will end up with as many numbers as bulbs. Some bulbs will last a relatively long time some will last a relatively short time and most observations will fall around about the middle somewhere. The time that each bulb lasts will be a number and each bulb we test will generate a number. If we test 50 bulbs we will be able to record 50 numbers. These numbers representing our repeated observations are called data.

We are not limited to one observation per bulb. We may also wish to record some information on location of the bulb (eg 1 to represent inside and 2 to represent outside) to see if location affects lifetime of bulbs. In this case, we will have 2 data items for each bulb, one for lifetime and the other for location.

## 3. Graphs help show the patterns in data

The question “how long does a light bulb of this type run before it blows” does not have a simple answer. Even if produced in the same factory and run in the same location, some bulbs will last a little longer than others will last a little shorter. No two items will ever be exactly identical. If we get a slightly different answer from each bulb we test, how can we have an answer to the question “how long do bulbs of this type run for?”. Giving 50 answers (one for each bulb) to the question is not very helpful.

One good strategy is to draw a picture or graph of the 50 bulb's observed lifetimes. (See below for discussion notes on various types of graphs.) A good graph will allow the reader to see how long most bulbs last and how much the lifetimes of bulbs vary between different bulbs. Such a graph will convey to the reader “most of the light bulbs lasted  $x$  amount of time, but lifetimes can vary from  $y$  to  $z$ ”.

*The best primary school students should get to here.*

#### **4. We can also use numbers to summarise these patterns**

A graph of data is useful as it allows us to immediately see the main characteristics of the data. For certain types of data we may be interested in the centre of the observed results, the spread of the observed results, what proportion of the responses fall into particular categories, etc. All of these main characteristics of the data can be reported as numbers. The centre of the observations can be recorded numerically as a mean or median, the spread of observations can be reported as a standard deviation and the percentage in any particular category is already a number (percentage). (See below for further discussion on these numeric measures).

An important advantage of a numeric (formula based) measure is that there is universal agreement on the value of that measure. Two people looking at a graph can quite legitimately have slightly different interpretations of where the centre of that graph is. Everyone calculating a mean from a particular data set should get exactly the same result – leaving no doubt or debate over the answer.

#### **5. We often can't measure everything – so we use a sample**

It is not possible to measure the lifetime of every light bulb the factory produces – there won't be any left to sell! It is not really possible to measure the average length of a particular species of cockroach – you would have to catch and measure every cockroach of that species in the world. How realistic is that!

For practical reasons we often measure just a sub-group (or sample) of the individual which we could measure if we had infinite time and money. For example, we test just some of the light bulb produced by the factory we are interested in and measure just some of the cockroaches that belong to the species we are interested in.

#### **6. A properly drawn sample can give us some information about the sample it was drawn from**

We have already noted that no two individuals are ever identical. Therefore, when we are using a sample, the each of individuals we are leaving out are different in some way to all of the individuals we are including. No member of the sample fully represents the excluded individuals. However, properly drawn sample can provide us with results which are similar to the results we would have obtained if we have measured every individual in the population.

Samples are used successfully every day. For example, when a chef is making a sauce she will take a little taste every now and again to check the flavour and consistency are correct. She doesn't have to eat all of the sauce to be sure that is good quality – a teaspoon full is usually sufficient. The only condition on using a sample in this way is that the sauce needs to be well stirred. If the sauce is burnt on the bottom and not stirred then the chef's sample from the top will not detect the problem.

Stirring the sauce is equivalent to randomising the selection for a sample. A sample consisting of individuals selected at random from the population of interest has the best chance of being similar to the population it was drawn from. If you sample only the easy to catch cockroaches, you may end up with a sample which is larger than the population of all cockroaches given bigger cockroaches are easier to catch.

*The best junior high school students should get to here*

## **7. Things can go wrong in sampling, measurement and interpretation**

Technically, samples should be randomly drawn from the whole population of interest. This minimises opportunities for biases. Bias occurred when the sample is systematically different from the population of interest. That is, if you wanted to know opinions of year 8 and 9 children on some particular issue, you would try not to just interview your friends. Your friends may be all girls – in which case the information you are collecting does not represent the opinions of all year 8 and 9 children. The opinions of boys would not be represented.

Even if selected completely at random, a sample may not be representative of the population it was drawn from. Randomly select 10 people from your street and there is a chance that all 10 will be females (and males will go unrepresented in your population). Select 1000 people at random from your town and there is a chance (albeit minute) that all 1000 may be female.

Non-response can also make samples biased (different to the population they should have been drawn from). If all the 17 year olds in your survey refuse to answer a question, then the answers you get provide no information on the opinions of 17 year olds.

Measurements can go wrong too. Poorly posed questions can lead to misleading answers. For example “You agree with <issue> don’t you” will tend to elicit a ‘yes’ response no matter what the issue because people tend to agree when asked to. Measurements of how high a ball bounces will differ according to whether you are looking at the top, bottom or centre of the ball. Measurement of how long a light bulb will last may only be accurate to the nearest day – it is hard to imagine someone will be watching the bulb continuously; the observer may only check the bulbs once per day.

A link between shoe size and reading ability may be wrongly interpreted as “study causes your feet to grow larger” (whereas there is probably a third hidden variable here – older age causing both larger feet and improved reading ability).

It is always important to review all work asking are there any errors or potential weaknesses in my sampling, my measurement or my interpretation of these results. All statistical results should be accompanied by an appropriate discussion of the strengths and weakness of the methods used.

*The best senior high school students should get to here.*

## **8. We can use probability models to estimate the most likely characteristics of a population using a random sample selected from that population. This is called statistical inference.**

Statistical inference are a group of mathematical techniques which are used to generalise information obtained from samples back to the wider populations which we are actually interested in. For example, when conducting a survey of students in the playground we are not specifically interested in the answers from the 20 children who actually answered the survey. We are actually interested in what the answers from these 20 children can tell us about all students at the school or all young people in the town. The tools of statistical inference are beyond the scope of the current competition (they represent the majority of the content in university statistics courses). However, better high school students will consider how any sample of individuals they have looked at relate to the wider population which they are really interested in.

*No student is expected to apply formal statistical inference methods.*

## Some tools for data analysis

Methods of statistical analysis may be divided into ‘Descriptive Analyses’ and ‘Statistical Inference’. Descriptive analyses can be conducted on data collected from whole populations and on data collected from samples. Statistical inference

### Descriptive Analyses

Descriptive analyses summarise the main features of the data. Raw data is too complex and too detailed to be of any direct use – we need ways of extracting the information we are looking for from otherwise difficult to read data sets.

Example: Here is a data set I collected from a class of university students. There are 19 students in the class. Each row of data relates to one individual student. That is, the first student in the data set is a 30 year old female who is enrolled in the Business Faculty and received a mark of 79%. The columns contain ‘variables’. The variables are the characteristics of the individuals which I actually measured. The second column is the variable *Age in years*. I asked each of the 19 students their age and recorded it in this column. The third column records whether each of the students is a male or female, etc. *Age, gender, faculty enrolled in* and *marks* vary from individual to individual (hence the name “variables”). The first column is not a variable; it is a label. Labels are not characteristics – the label tells me nothing about the individual’s abilities or attributes. Labels are just a tag I put on to that individual for my own convenience.

Student	Age in years	Gender	Faculty	Mark (%)
1	30	Female	Business	79
2	26	Male	Business	78
3	21	Male	Business	62
4	19	Female	Business	59
5	25	Male	Engineering	58
6	33	Female	Sciences	75
7	35	Male	Sciences	71
8	29	Male	Business	82
9	41	Male	Engineering	100
10	25	Male	Business	69
11	24	Female	Business	56
12	31	Male	Business	72
13	24	Female	Business	80
14	31	Male	Engineering	87
15	21	Male	Business	87
16	25	Male	Business	83
17	24	Male	Engineering	73
18	23	Male	Business	90
19	31	Female	Business	78

It can be difficult to see all the patterns in the raw data. You can see that there are 19 students in the class. You can see that there are more males than females in the class and the majority of the students are enrolled in the Business Faculty. If you look closely you can see that students’ ages range from 19 to 41 years. However, even with this small data set, it is very difficult to answer more complex questions like:

- What proportion of students are less than 30 years of age?
- Do older students get better marks?
- Do females get better marks?

There are a range of statistical tools available to help us summarise the important patterns in the data set. In order to select the right tool for the right job we first need to understand the difference between categorical and quantitative variables.

Quantitative variables are measured on a numeric scale. For example, *age* is measured in years and years is a numeric scale. You can do maths on numeric scales and get answers that make some sense.

That is, if the student 1 is 30 years of age and student 2 is 26 year of age then student 1 is  $30-26=4$  years older than student 2.

Categorical variables are measured on categorical scales. For example *gender* is measured in categories (male or female). You can't do maths on categorical scales. For example, Male+Femal? does not have a mathematical answer.

In the example above *age* and *mark* are quantitative variables and *gender* and *faculty* enrolled in are categorical variables.

Once you have figured out which of your variables are categorical and which are quantitative you can move on to selecting the appropriate statistical tool. The choice of which tool to use where is summarised in the following tables:

Statistical tools for summarising a single variable

	Categorical variables	Quantitative variables
Graphical summaries	Bar chart Pie chart	Histogram Boxplot
Numerical summaries	Frequency counts	Mean and standard deviation Median and quartiles

Statistical tools for summarising the relationship between two variables

	Categorical and categorical	Categorical and quantitative	Quantitative and quantitative
Graphical summaries	Complex bar charts	Side-by-side boxplots Multiple histograms	Scatterplots
Numerical summaries	Frequency tables	(ANOVA)	(Correlation) (Regression)

Some of these statistical tools are described below.

### Frequency counts

This is simply a count of the number of times each outcome occurs. For example a frequency count on gender in the example above is:

- 13 males and
- 6 females

Sometimes people prefer to present percentages rather than raw frequencies. These are called relative frequencies. In this example the relative frequencies are:

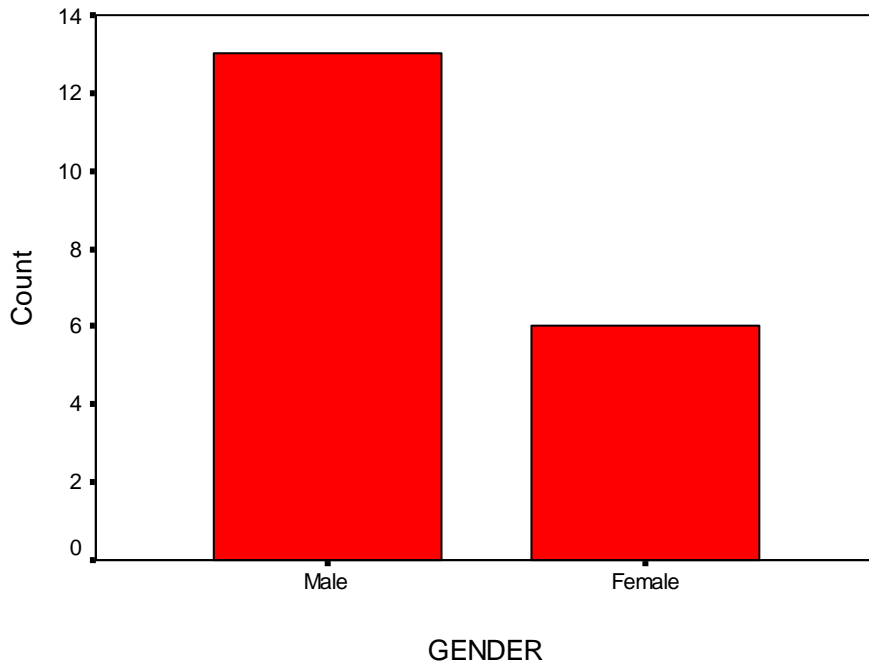
- 68% males (ie  $13/19 \times 100\% = 68\%$ ) and
- 32% females

If using relative frequencies, make sure you also mention the total number of individuals in your data set (the sample size). 1 out of 10 and 100 out of 1000 both have relative frequency (10%) even though the study on 10 individuals provides far less repetition (and hence less confidence) than the larger study.

### Bar charts

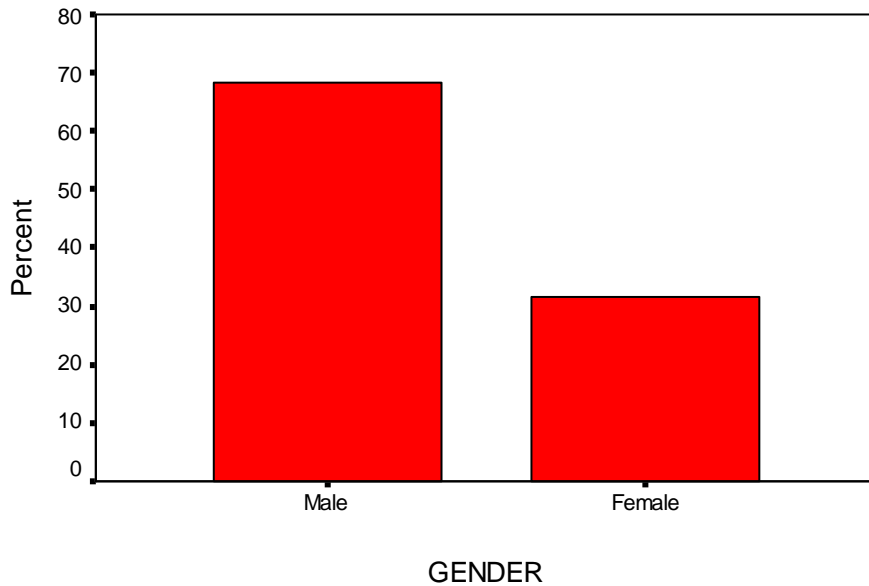
Bar charts are graphical representations of the frequency counts. Examples of bar charts for gender are:

Bar chart for gender in Uni data set



Relative frequencies of gender in Uni data

(n=19)



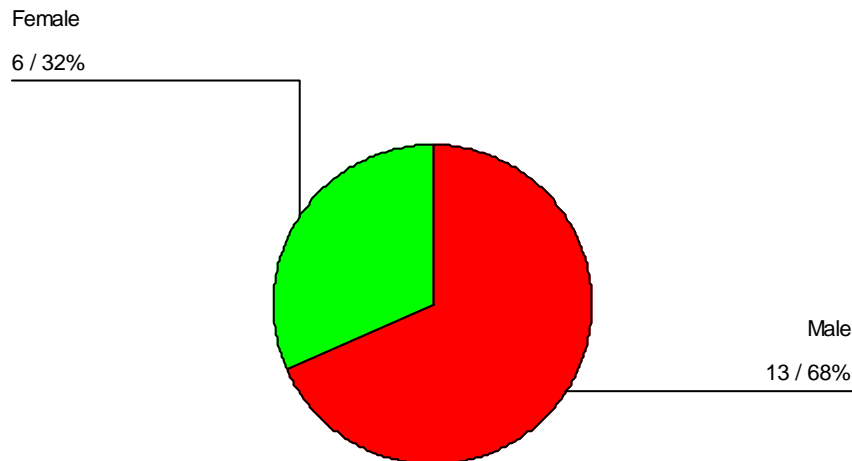
The categories (male and female in this case) are displayed on the horizontal axis and the frequencies or relative frequencies are given by the heights of the bars. All titles and scales are vital. Without the titles and scales, the reader will have difficulty in understanding the information being presented. Notice the inclusion of the sample size (n=19) when using relative frequencies.

## Pie charts

Pie charts are also constructed from frequency counts or relative frequencies. In this case the whole sample size (19 in our example) is represented by the circle and the frequencies or relative frequencies are depicted as segments of the whole. For example:

### Pie chart for gender in Uni data set

Frequency counts and relative frequencies (n=19)



Again, the titles and labels are just as important as the 'pie' itself.

## Histogram

To create a histogram:

- divide the quantitative scale up into say 5 to 12 separate intervals
- count the frequency in each of intervals
- draw a 'bar chart' of frequencies

There is an important difference between histograms and bar charts. Bar charts have gaps between the bars (representing the categorical scale) but histograms have no gaps between the bars.

I will summarise *Age in years* using a histogram. As there is a relatively small sample size (n=19) five intervals, each 5 years wide will be sufficient in this case. The 5 year intervals I am going to use are:

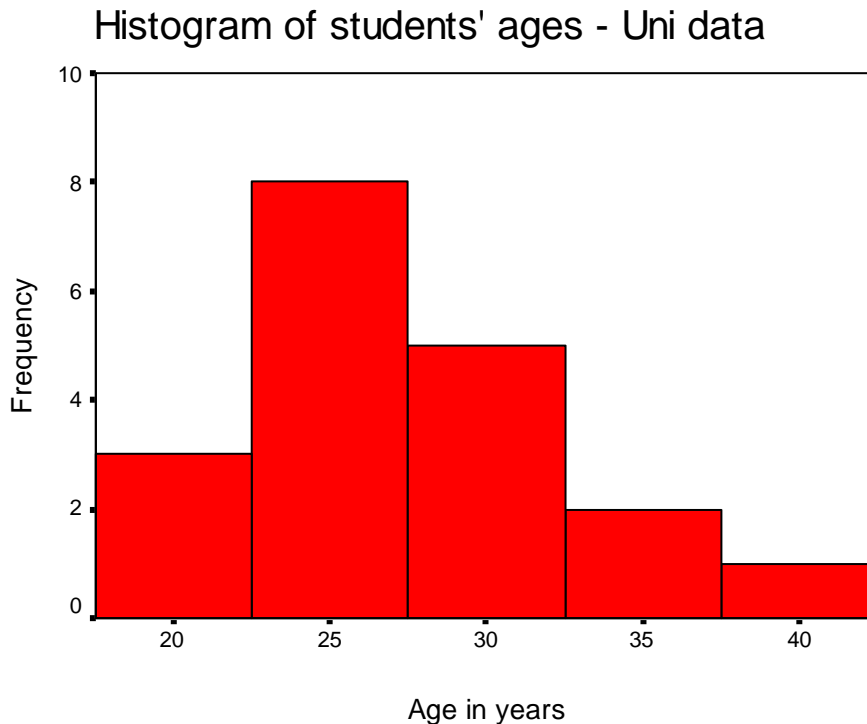
- 17.5 years to less than 22.5 years
- 22.5 years to less than 27.5 years
- 27.5 years to less than 32.5 years
- 32.5 years to less than 37.5 years
- 37.5 years to less than 42.5 years

These definitions look unusual but have a particular purpose – the centres of the five intervals are 20, 25, 30, 35 and 40 years respectively. It is important that each interval contains exactly the same number of years – each is 5 years wide in this case.

The frequency count for each interval is:

- interval centred on 20 years: 3 students
- interval centred on 25 years: 8 students
- interval centred on 30 years: 5 students
- interval centred on 35 years: 2 students
- interval centred on 40 years: 1 student

This information can be used to construct a histogram like the following.



Notice that I have labelled the mid-point of each interval. Alternatively, you could label each interval with the range of values it contains (eg label the first interval 17.5 to 22.5 years). There are no gaps between the bars. The chart is fully labelled. Again, relative frequencies (percentages) could be used instead of actual counts.

#### Mean and Standard deviation

Graphs contain more information, but there are times (particularly when it comes to statistical inference below) when single number summaries of the data are more convenient. The mean and standard deviation are both single number summaries of the data. The mean is a measure of the centre of the data and the standard deviation is measure of the spread. (Looking at the histogram above you can see that the centre of the data is probably somewhere between 25 and 30 years and that students ages spread somewhere from say 17.5 years and 42.5 years).

The 'mean' is what people generally call the 'average'. To work out the mean of *age* you add all the ages together and divide by the number of individuals you have data for. In this example the mean of *age* is:

$$\bar{x} = \frac{30 + 26 + 21 + 19 + 25 + 33 + 35 + \dots}{19} = 27.26$$

The standard deviation is, roughly, the average of how far each observation is from the mean. The higher the standard deviation, the more spread out the data are.

Calculating the standard deviation is difficult and not recommended. We recommend you use a scientific calculator or computer to help you work it out. In this case the standard deviation of *age* is 5.516 years.

Aside:

If you are keen, here's how it's done: In this example, individual 1 is 30 years of age. This is  $30 - 27.26 = 2.74$  years away from the mean of *age*. Individual 2 is 26 years of age and hence  $26 - 27.26 = -1.26$  years away from the mean. The fact that some distances from the mean are positive (above the mean) and some are negative (below the mean) causes us problems. To make all the distances positive we square them: so  $2.74^2 = 7.5076$ ,  $-1.26^2 = 1.5876$ , etc. Once all of the distances have been squared and are all positive, they are added together to get a total squared distance score. To get an 'average' squared difference score we divide by one less than the sample size (ie divide by 18 in this case) – it takes a fair bit of work to explain why, so we usually stick with 'trust us' for the first year or two at University. Finally, to correct for the fact that we squared all of the distances, the last step is to take the square root of the answer.

## Median and Quartiles

If you were to check back on the example data set, you would find that 11 of the 19 students in the data set were younger than the mean of *age* and only 8 were older. In this case we can see the mean isn't doing very well at describing the centre of the data. A different, and sometimes better, method for describing the centre of the data is to find the middle point or median.

To find the median, we first sort the data from smallest to largest. For 'Age in years', this will give us:

19, 21, 21, 23, 24, 24, 24, 25, 25, **25**, 26, 29, 30, 31, 31, 31, 33, 35, 41

You can see where the middle (median) is now. It is the 25 years which is highlighted. There are 9 data points below this one and 9 data points above it. (If you have an even number of data points the median is the point half way between the middle two data points.)

The first quartile is the point that has one quarter of the data below it and three quarters above it. Perhaps the easiest way to find it is to find the median of the lower half of the data. That is:

19, 21, 21, 23, **24**, 24, 24, 25, 25, 25, 26, 29, 30, 31, 31, 31, 33, 35, 41

If we just look at the nine data points below the median, the first quartile lies at the centre of those nine data points. In this case the highlighted 24 years has 4 data points below it and 4 above.

The third quartile is the point that has three quarters of the data below it and one quarter above it. Using the equivalent method:

19, 21, 21, 23, 24, 24, 24, 25, 25, 25, 26, 29, 30, 31, **31**, 31, 33, 35, 41

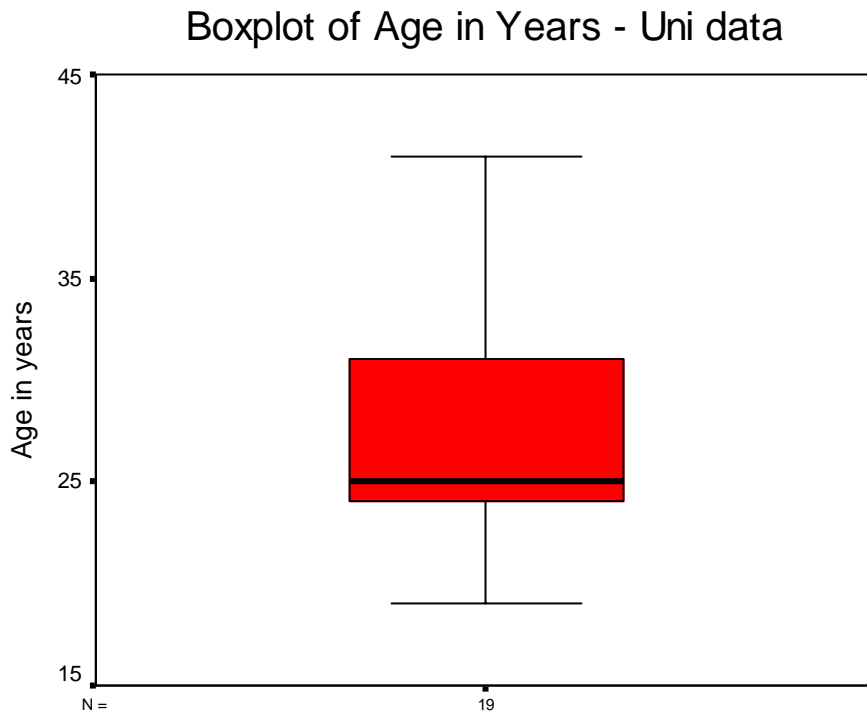
The third quartile is 31 years of age.

The quartiles describe the spread of the middle 50% of the data.

Using the median and quartiles we can say that the centre of the age data is 25 years and the middle 50% of observations lie between 24 and 31 years.

## Boxplot

Medians and quartiles are sometimes depicted graphically using a boxplot. Here, for example, is the boxplot of *age in years* for the university data.



The dark line within the box represents the median (25 years). The bottom and top of the box show the first and third quartiles (24 and 31 years) and the whiskers extend up to the maximum age (41 years) and down to the minimum age (17 years).

## Frequency tables

Suppose we wished to summarise the relationship between two categorical variables – lets look at *gender* and *faculty* for example. To summarise this relationship we construct a table with one variable across the top and the second variable down the side as follows:

**Frequency table of gender by faculty**

		Gender		Total
		Male	Female	
Faculty	Business	8	5	13
	Science	1	1	2
	Engineering	4	0	4
Total		13	6	19

It makes no difference which variable you put across the top and which one you put down the side. Using the table above we can see that of the total of 19 students, 8 are male business students, 5 are female business students, etc. Actually it appears that there is a relationship between *faculty* and *gender* in that the Faculty of Engineering has no females.

Column (or row) percentages will often assist in highlighting relationships. For example:

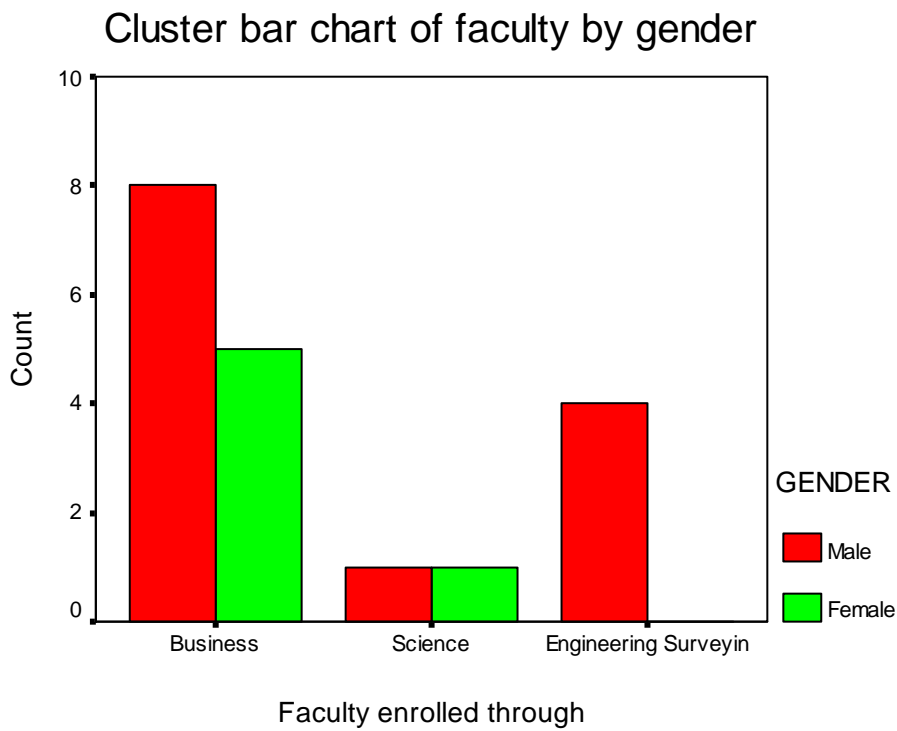
**Frequency table of gender by faculty**

			GENDER		Total
			Male	Female	
Faculty enrolled through	Business	Count	8	5	13
		%	61.5%	83.3%	68.4%
	Science	Count	1	1	2
		%	7.7%	16.7%	10.5%
	Engineering	Count	4	0	4
		%	30.8%	.0%	21.1%
Total	Count	13	6	19	
	%	100.0%	100.0%	100.0%	

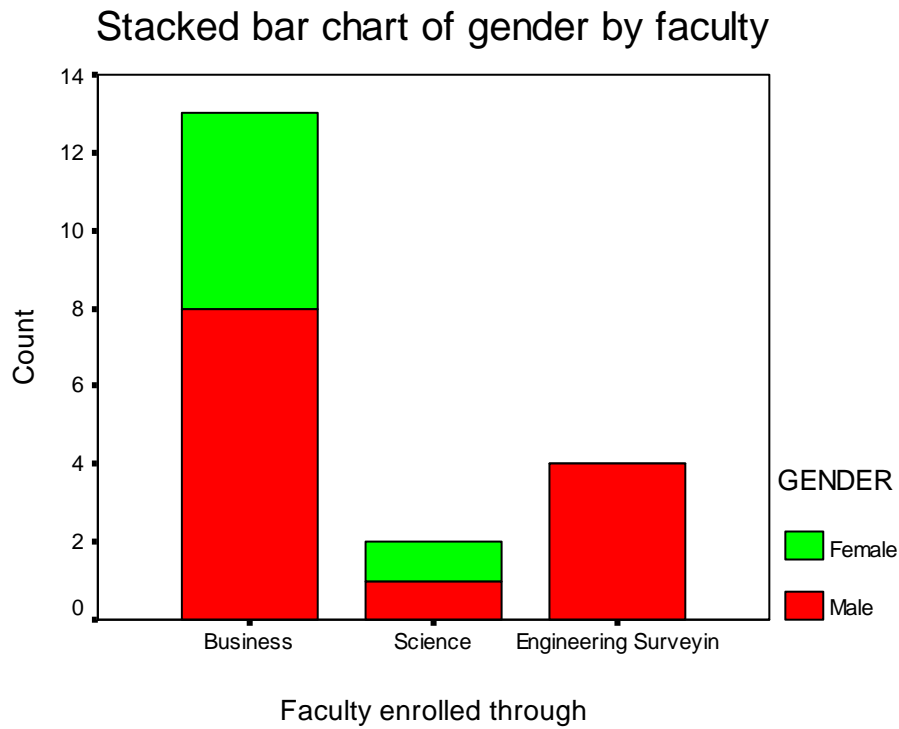
Percentages are calculated for each column separately (such that each column represents 100%). We can see that 83% of all females have enrolled through business and 0% have enrolled through Engineering. A lower proportion of males have enrolled through Business and 30% have enrolled through Engineering.

**Clustered or stacked bar charts**

This is a clustered bar chart of two categorical variables:



This is a staked bar chart of two categorical variables:



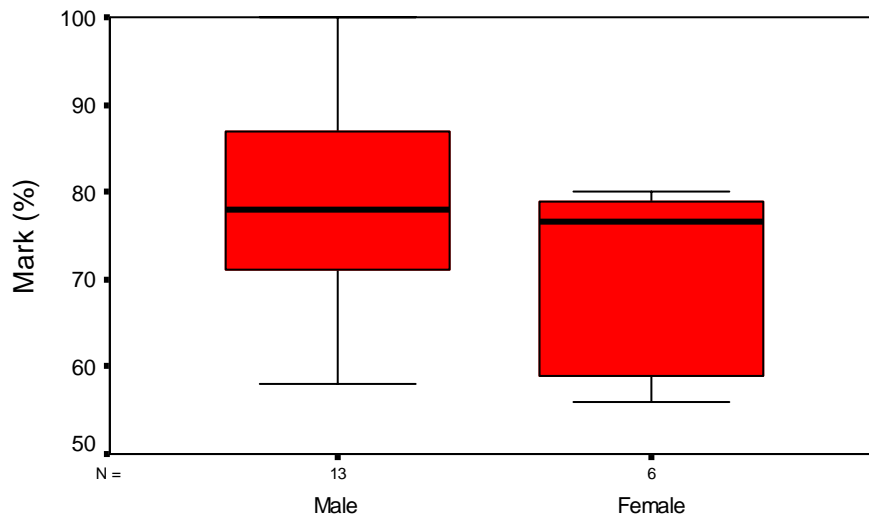
The counts are taken from the frequency table. They may be converted to percentages and presented as relative frequencies.

#### Side-by-side boxplots and multiple histograms

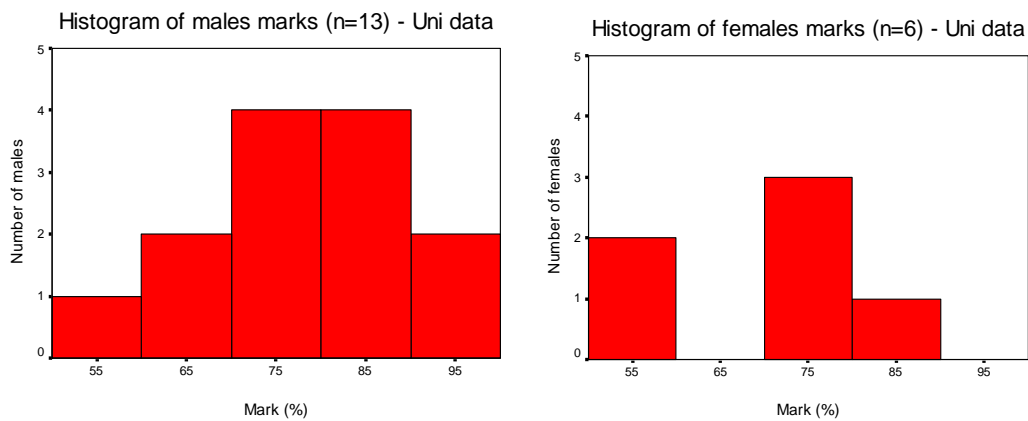
When we wish to look at the relationship between a categorical variable and a quantitative variable we first separate our data into the categories. That is, if looking at the relationship between *gender* and *marks*, we first separate marks into two groups – *marks* for males and *marks* for females. We then do the normal boxplot or histogram on *marks* within each group:

## Side-by-side boxplots of marks by gender

### Uni data



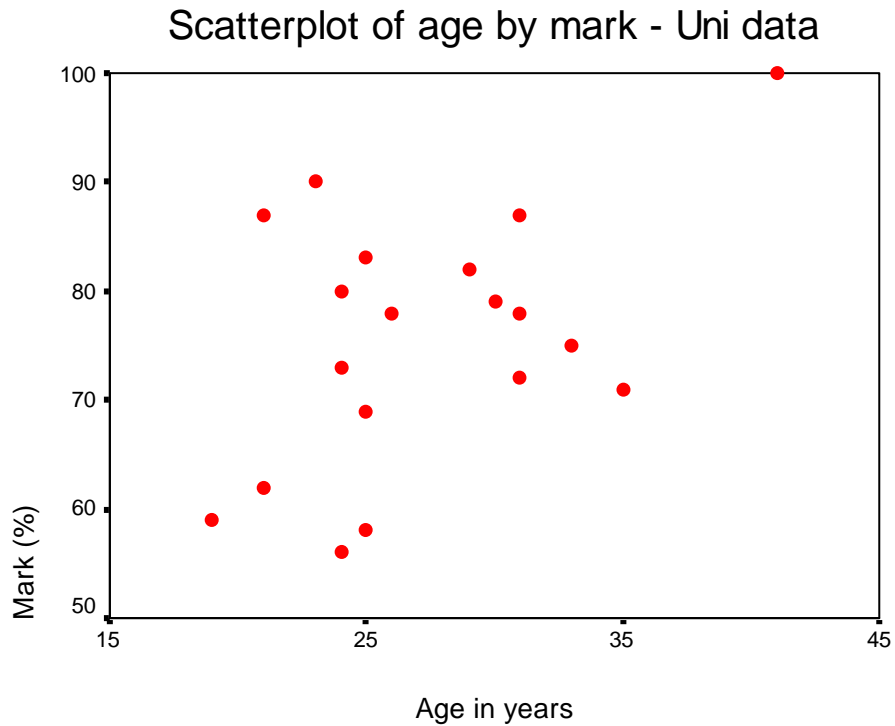
### GENDER



Notice that the boxplots or histograms have to be plotted on exactly the same scales (on both the vertical and horizontal axes) to allow direct comparison of results in the different categories.

### Scatterplots

When looking at the relationship between two quantitative variables we again put one variable on the horizontal axis and one on the vertical axis (for our purposes here, it doesn't really matter which variable is put where). Both the horizontal and vertical axes are numerical scales and each individual; has their result plotted at the appropriate height on both axes. Here is a scatterplot of *age* against *mark*:



Consider the point at the top right of the graph. Looking straight down below this point on the horizontal axis we see that this student was 41 years of age. Looking straight across to the vertical axis we see the student obtained a mark of 100%.

Looking at this scatterplot, there may be a general trend in the points from the bottom left to the top right (there are very few points at bottom right or top left). Such a trend would imply that older students tend to get better marks than younger students.