Study Guide 10: An Introduction to Medical Statistics

Dr David Chinn,
Research, Innovation & Knowledge Department
Queen Margaret Hospital, Dunfermline, Fife.
david.chinn@nhs.scot 01383 623623 (ext 20955 Roy Halliday)
Alternative contact: Prof Frances Quirk frances.quirk@nhs.scot 01383 623623 (ext 20941)

Contents

Disclaimer

I am an epidemiologist, not a statistician. These notes are written from my experience of working in the field of medical research for over 40 years. I have sought to give what I hope is a clear and simple explanation of some rather complex statistical principles. I do not profess to be an expert in statistics and a 'proper' statistician reading this guide may take issue with some of my explanations. Accordingly, I would encourage the reader to refer to one of the many excellent introductory books available on statistics for further guidance; some titles are given in the references and further reading.

(1) Overview and learning outcomes

This guide is directed at those with no knowledge of statistics who require a better understanding of how statistics can be used to describe and summarise data. The level is very basic and we use examples throughout to illustrate the procedures and concepts. After reading this guide you should be able to:

- Describe different data types
- Describe measures of 'location' (also known as 'central tendency') and variability (spread) appropriate to the distribution of the data

- Describe the features of the Normal distribution and the t-distribution together with aspects of their application
- Understand the derivation and interpretation of the 'standard deviation'
- Derive and interpret confidence intervals for continuous data and proportions.

Associated NHS Fife study guides:

7	How to plan your data collection and analysis
11	How to calculate sample size and statistical power
12	How to choose a statistical test
13	How to make sense of numbers
14	An introduction to SPSS

(2) Introduction

Statistics is a fundamental tool in medical science. All health care practitioners, whatever their discipline, need a basic understanding of statistics as consumers of medical information. Statistics is concerned with estimation; we estimate what we think is the truth in a population by measuring something in what we hope is a representative sample from that population. Also, statistics is concerned with measuring variability within- and between-persons and the source and size of this variability. When two or more groups of individuals are compared there will always be a difference between them; the challenge is to decide if that difference is real or an artefact due to random variation.

Statistics are commonplace. Regularly we hear figures on average salaries, changes in the rate of inflation, percentage changes in house prices. In medicine we use common statistics such as length of hospital stay, average age of in-patients on a ward, readmission rates, hospital acquired infection rates, mortality rates, survival rates etc. Health care practitioners need a basic understanding of statistics to make sense of the information they are presented with, for example in journal articles and by drug reps and others when promoting their products. Published articles in 'quality' journals go through a rigorous peer review process. But, despite this, mistakes get through and papers are published which can be misleading. You cannot always trust what you read and, in consequence, everyone needs some basic knowledge of statistics. Even when presented accurately some knowledge is necessary to interpret them correctly!

Consider the following statements and set of true/false questions. Try to answer them but do not be concerned if you cannot answer them now. They will reappear at the end of this guide and, hopefully, you should have more confidence in answering them after reading it.

Quiz 1: True or false?

The Normal Distribution

1) is followed by many variables
2) is also called the Gaussian distribution
3) is followed by all measurements made in healthy people
4) is described by two parameters
5) is skew to the left

The Standard Normal Distribution
6) has mean = 1.0
7) has standard deviation = 0.0
8) has variance = 1.0
9) has the median equal to the mean

Quiz 2: The FEV_1 (a measure of lung capacity) of a group of women aged 20-25 follows a Normal distribution with mean of 3.0 litres, standard deviation 0.4 litres.

Which statements below are true?

1). The distribution of FEV is symmetric about the mean.
2). About 95% of the women have an FEV between 2.2 and 3.8 litres.
3). About 5% of the women have an FEV below 2.2 litres.
4). 50% of the women have an FEV below 3.0 litres.
5). The largest FEV will be 4.6 litres (mean + 4 SD)
6). A woman with an FEV below 2.2 litres is abnormal (unhealthy).
7). If the sample size was doubled the standard deviation would decrease.

(3) Descriptive and inferential statistics

Statistics is the science of assembling and interpreting numerical data. So, it has a role in designing a study, collecting information and in the analysis and interpretation of that information (or data). It is not just about 'number crunching.'

Descriptive statistics refers to the simple description of a sample in terms of some average value and a measure of the variability in the sample (which reflects the differences between different members of that sample). In usual practice we want to measure some aspect in a *population* (of all persons) by measuring that aspect in a *representative sample* taken from that population. A defined sample (of say 100 persons) is randomly chosen from all the persons making up the population (which may number, say, 5,000 in total). Summary statistics are then calculated from the sample to estimate what we believe reflects the true average and variability in the population from which the sample was drawn.

Inferential statistics make inferences, in the form of probabilities, between two or more populations from which representative samples are drawn. Some characteristics are measured in the samples and probability estimates made to test hypotheses of, for example, equivalent values (a null hypothesis) (more on this later). Inferential statistics can also include measuring inter-relationships between variables, for example, correlations, regression statistics etc.

(4) Types of data

There are essentially two types of data and measurement scales.

(1) Categorical data (all or none)

Categorical data (data in categories) are mutually exclusive (you can only be in one category or another) and may be *nominal* or *ordinal* in character. Data are described as nominal if they cannot be ordered because there is no natural order. Examples

include marital status, smoking habit, religion, eye colour, nationality, and vital status (dead or alive).

Data are described as ordinal where there is a ranked order but the size of the difference between adjacent categories is not identical. Consider the result of a road race. Runners are categorised as first, second, third etc. The winner may have run the race in 60 minutes, the person coming second may have taken 62 minutes (2 minutes behind), and the person in third place may have taken 75 minutes (13 minutes behind the person coming second). Hence, the difference between adjacent categories (1^{st}, 2^{nd}, 3^{rd} as in this example) is not the same. Another example is where patients are asked to report pain on a categorical scale of 'no pain', 'a little pain', 'a lot of pain', 'the worst imaginable pain'. We cannot assume the difference between being in 'no pain' and 'a little pain' is the same magnitude as the difference between being in 'a little pain' and 'a lot of pain'. Other examples of ordinal scales are the Borg scale of breathless scores (for example, in exercise studies) and those often used in patient satisfaction surveys where the response options on receiving a service may include, for example, 'Excellent', 'Good', 'Fair', 'Poor'.

(2) Interval or Continuous data

Interval or continuous data are derived from a count, or a standard measurement, and have a frequency distribution. The numbers can be discrete or continuous. Discrete data are integers (whole numbers) where the size of the difference between adjacent categories is identical (unlike ordinal data above). Examples include the number of children in a family (0, 1, 2, 3, 4, 5 etc), number of asthma attacks in a year, number of GP visits in a year, the number of beds on a hospital ward, or length of hospital stay (in days). The difference between 1 and 2 days (i.e. 1 day between adjacent categories) is the same as the difference between 6 and 7 days.

Continuous data can take any value within a range and includes measures such as standing height (stature), weight (technically *body mass*), haemoglobin level, and age. The data are measured in standard units, with clear meaning attached to the difference between measures whatever the magnitude of the measurement. For example, the difference between a body mass of 21-26 kg (i.e. 5 kg) is the same size as that for a difference between 65-70 kg.

There are differences in interpretation, however, between discrete and continuous data. We can calculate the average number of children in a family (for example 2.5) and the average body mass (for example, 68.9 kg) but the numbers have different interpretations. It is possible for a person to have a body mass of exactly 68.9 kg but not possible for a family to have 2.5 children!

Another form of data is the *ratio* which is also measured in standard units but the scale has a true zero which represents the total absence of the variable (for example, age, volume, time, mass).

The distinction between the different types of data is important because this has implications for graphically representing the data and choosing an appropriate summary descriptive measure and statistical test (Table 1).

Table 1. Data types and summary descriptive statistics

	Nominal	Ordinal	Interval	
			Discrete	Continuous
Example	blood group, eye colour, nationality, gender	symptom score	family size, length of stay	age, height, body mass, haemoglobin
Graphical summary	Bar chart Pie chart	Bar chart Pie chart	Bar chart Pie chart Histogram Boxplot Dotplot	Histogram Boxplot
Summary measures of location – *see section 5 below*	Mode	Mode	Mean Median Mode	Mean Median
Measures of spread (variability) – *see section 6 below*	-	Range, Inter-quartile range	Standard deviation, Range, Inter-quartile range	Standard deviation, Range, Inter-quartile range

Statistics are used to describe a lot of information about a sample in a few key numbers.

Any set of measurements has two important aspects:

(1) a measure of ‘location’ or ‘central tendency’ (an ‘average’ or ‘typical’ value) and,

(2) a measure of the spread of data values about that average value.

(5) Measures of location, central tendency

Measures of location are the mode, mean, and median.

The *mode* (symbol, *Mo*) is the value that occurs most often and is only really sensible with discrete data (Table 2). A set of numbers with one mode is called unimodal, that with two modes is called bimodal and that with more than two modes is multi-modal.

Table 2. Examples of the mode
(the value that occurs most often)

Previous pregnancies in women attending an antenatal clinic	Number
0	59
1	51
2	19
3	5
4	4
5	2
6	1
Mode =	***0***

Number of modules attended by students	Number
0	4
1	16
2	19
3	14
4	10
5	7
6	0
Mode =	***2***

The *mean* (symbol, $\bar{X}$) is the simple mathematical average calculated from the sum (Σ) of the individual numbers (*Xi,* where *i* refers to 'individual') divided by the number of cases (*n*) in the sample.

$$\text{Mean, } \bar{X} = \frac{\sum Xi}{n} \qquad \text{(equation 1)}$$

The mean is only useful for continuous data. Sometimes numerical codes are used to describe categorical data such as gender, religion, ethnicity etc. If, for example, in a data set, males were coded as '1' and females as '2' it does not make sense to calculate the 'average' gender. How, for example, can you interpret an average gender of 1.6?

The *median* (symbol, *Md*) is the middle value when data are ordered from lowest to highest (or vice versa). The median splits the sample in half (also called the 50th percentile).

Example data, the weekly income of 8 persons (£):

275, 286, 337, 276, 262, 347, 305, 312

Mean = sum of observations / number of observations

$$\text{Mean, } \bar{X} = \frac{\sum Xi}{n}$$

$$\text{Mean} = \frac{275+286+337+276+262+347+305+312}{8} = £\ 300$$

Median= middle value = (n+1)/2 th value

To derive the median first rank the observations from lowest to highest:

Rank:	1	2	3	4	5	6	7	8
Value:	262	275	276	286	305	312	337	347

median = (number of cases + 1)/2 = (8+1)/2 = 4.5th value.

Average values of ranks 4 and 5 = (286 + 305) / 2 = 295.50

The mean is a good measure of location (a 'typical' value or measure of central tendency) when the data are fairly well distributed and do not have extreme values (known as outliers, *see* Glossary). However, the mean can be markedly affected by a single outlier whereas the median is unaffected. Consider a change to the list of 8 weekly incomes above where the highest value changes from £347 to £747. The mean is now £350 but the median remains at £295.50. Only one value for income is now above the mean which therefore no longer represents a 'typical' value of 'central tendency.'

(6) Measures of spread, variability

Measures of spread are the Range, Variance, Standard Deviation, Percentiles, and Interquartile Range.

Range

The *range* is the difference between the highest and lowest values in a data set. In the income example above the range is the difference between £347 and £262 or £85. It has limited usefulness, particularly when there are outliers, and is little used.

Variance and standard deviation

The *variance* (symbol, s^2) is the sum (Σ) of the squared differences of individual values from the mean $(Xi - \bar{X})^2$ divided by the number of observations (n) minus 1.

The superscript 2 in $(Xi - \bar{X})^2$ means we take the square of the difference between Xi and $\bar{X}$. The square of a number is the number multiplied by itself, hence 2^2=2 x 2 = 4, and 4^2 = 4 x 4 = 16.

$$\text{Variance, } s^2 = \frac{\Sigma (Xi - \bar{X})^2}{(n-1)} \qquad \text{(equation 2)}$$

The *standard deviation* (symbol, *s*) is the square root of the variance and can be thought of as the <u>average difference of individual values from the mean</u>.

$$\text{Standard Deviation, } s = \sqrt{s^2} \qquad \text{(equation 3)}$$

An example of the calculation of variance and standard deviation (SD) using the income data above is given in Table 3. The individual values were (£): 275, 286, 337, 276, 262, 347, 305, 312 and the mean was £300. This part may seem confusing but persevere with it as consulting the table will help you see how the various statistics, particularly the standard deviation, are derived.

Table 3. Calculation of variance and standard deviation

Xi *(units=£)*	$Xi - \bar{X}$	$(Xi - \bar{X})^2$ Called the 'Sum of Squares' (SS) (units=£^{2})
262	262 – 300 = -38	1444
275	275 – 300 = -25	625
276	276 – 300 = -24	576
286	286 – 300 = -14	196
305	305 – 300 = +05	25
312	312 – 300 = +12	144
337	337 – 300 = +37	1369
347	347 – 300 = +47	2209
	$\sum(Xi - \bar{X}) = 0$	$\sum(Xi - \bar{X})^2 = 6588$ Variance = SS / (n-1) **= 941.14** SD = $\sqrt{\mathbf{941.14}}$= **30.68**

Note, the sum of deviations ($Xi - \bar{X}$) is always zero, which is not helpful. They are squared to make them all positive. This figure is called the 'sum of squares' (or SS) and, because we have squared the differences the units are £ squared (£^{2}):

$$\sum(Xi - \bar{X})^2 = 6588$$

The SS measures the size of the scatter in the data, but not its direction.

The Variance (s^2) is the average of the squared deviations, $s^2 = \Sigma(Xi-\bar{X})^2$ / (n-1) = 941.14, but, as the units are £^{2} (pounds squared) this is not very useful. The Standard Deviation (s) is the square root of the variance to make the units the same as those of the mean.

$$s = \sqrt{\boldsymbol{Variance}} = \sqrt{\mathbf{941.14}} = 30.68 \text{ (£)}$$

The summary measures for the income data set are now mean = £300, Median = £295.50 and standard deviation = £30.68

The reason why we use 'n-1' and not 'n' is concerned with the fact that you cannot estimate the variance by making only one measurement from a set; instead, you must make at least 2 measurements. In addition, we use 'n-1' because it gives a better estimate of the variance of the *total* population from which the sample was drawn.

The standard deviation is a measure of the spread of data from a distribution and, as we have seen, is derived from the data itself. The mean and SD are good at summarising the data when the spread of data values is symmetrical, or approximately symmetrical, about the mean and there are no outliers (extreme values).

Quick formula for calculating the Variance and Standard Deviation

Tip: the convention is to always calculate the parts within the brackets first.

$$\text{Variance, } s^2 = \frac{1}{(n-1)} \times \left(\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right)$$

$$\text{SD, } s = \sqrt{s^2}$$

Weekly incomes, *Xi*	Xi^2
262	68644
275	75625
276	76176
286	81796
305	93025
312	97344
337	113569
347	120409
$\sum(Xi) = 2400$	$\sum(Xi^2) = 726588$

$$\text{Variance, } s^2 = 1/7 \times (726588 - (2400 \times 2400 / 8)) = 941.14$$

$$\text{SD, } s = \sqrt{s^2} = 30.68$$

Exercise: calculate the mean, and standard deviation (SD) of the following data set:

Xi	Xi^2	
2		
5		
6		
9		
7		
4		
2		
3		
6		n =
7		Mean, $\bar{X} = \sum Xi/n$ =
$\sum Xi$ =	$\sum Xi^2$ =	SD =

Answers: n= 10, $\bar{X}$= 5.1, SD = 2.33 (variance = 5.43)

Percentiles and the inter-quartile range

Consider listing the individual values in a dataset from the smallest to the highest value. The value that has 1% of the observations lying *below* it is called the first *percentile*. The value that has 10% of the observations lying below it is called the 10th percentile. Similarly, the 25th percentile and 75th percentile defines the values in the dataset that will have 25% and 75% of the observations lying below them, respectively. The 50th percentile is the median (that is the value that splits the sample in half). Percentile charts are used, for example, to define reference ranges in healthy subjects (Figure 1).

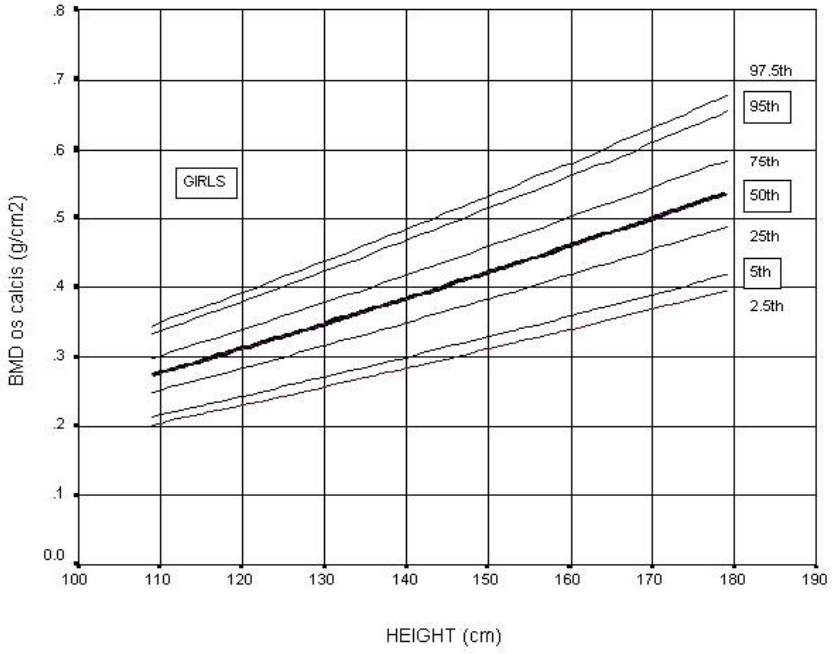


Figure 1. The use of percentiles to define the reference range for bone density at the os calcis (heel) in girls aged 5 – 18 years.
The thicker, centre line is the median (50th percentile).
[source: Chinn et al. *Arch Dis Child* 2005; **90**: 30–35]

The *inter-quartile range* (IQR) is the difference between the 25th and 75th percentiles (also called the first and third quartiles). The IQR encompasses the middle 50% of the data. The IQR is a useful descriptive measure but is variable from sample to sample and cannot be used for purposes of comparing groups, unlike the standard deviation. An example of the IQR is given later in Figure 4.

(7) The distribution of data (the frequency distribution or histogram)

Continuous data can be grouped into categories and the frequency of observations within each category plotted in a histogram which provides a useful visual summary, particularly of a large amount of data. Consider the standing height of women as recorded in the 1998 Scottish Health Survey. The women are categorised into one-centimetre blocks (Table 4), and the blocks then plotted as a histogram (Figure 2).

Table 4. Height Categories in 3,607 adult Women, Scottish Health Survey, 1998

Height range (one cm blocks)	Number of Women
140 - 141	2
141 - 142	5
142 - 143	5
143 - 144	8
144 - 145	6
etc ….	……
150 - 151	53
etc….	……
160 – 161	222
etc….	……
170 - 171	81

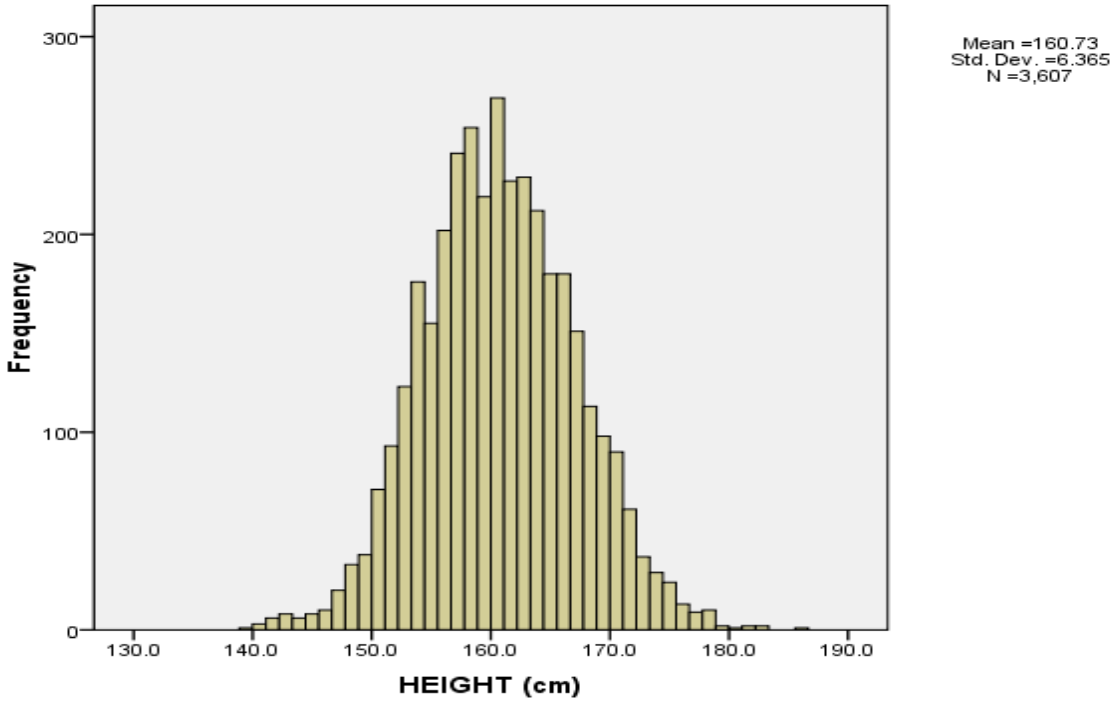


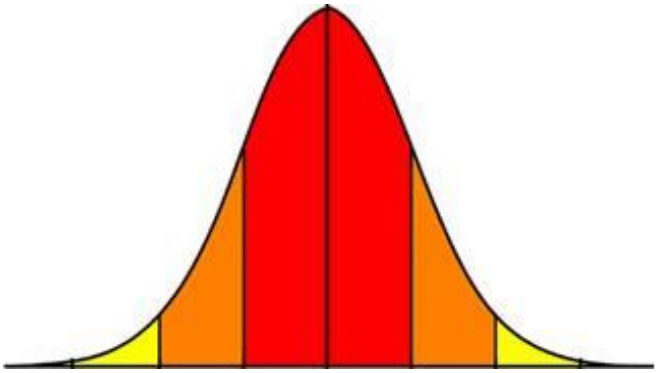
Figure 2. Height of 3,607 adult women recorded in the Scottish Health Survey, 1998.
Women were grouped into 1 cm blocks and the number of women (Frequency) in each block plotted as a histogram.
Mean = 160.7 cm, standard deviation = 6.36 cm

This distribution is bell-shaped and is referred to as an example of the 'Normal,' or Gaussian distribution (Figure 3). The properties of the Normal distribution are:

(1) Distribution is bell-shaped,
(2) Distribution is symmetrical (balanced) about the mean,
(3) The mean = median = mode,
(4) The mean plus or minus (+/-) 1 standard deviation encompasses 68% of the observations,
(5) The mean +/- 2 standard deviations encompass about 95% of the observations,

(6) 2½% of observations have a value = mean – 2 standard deviations,
(7) 2½% of observations have a value = mean + 2 standard deviations

From Figure 3 it is apparent that the large majority of women (about 68%) have a height in the range 154 to 167 cm, which is the mean +/- 1 SD, and about 95% have a height in the range 148 to 173 cm, which is the mean +/- 2 SD. Hence, about 5% of women will have a height outside this range, 2½% at either end of distribution, so 2½% will be shorter than 148cm and 2½% will be taller than 173 cm.



The two central areas together represent the mean +/- 1 SD (68% of the observations).

The four central areas together represent the mean +/- 2 SD (about 95% of the observations).

The ‘tails’ represent the extremes beyond the mean +/- 2 SD (2½% each side).

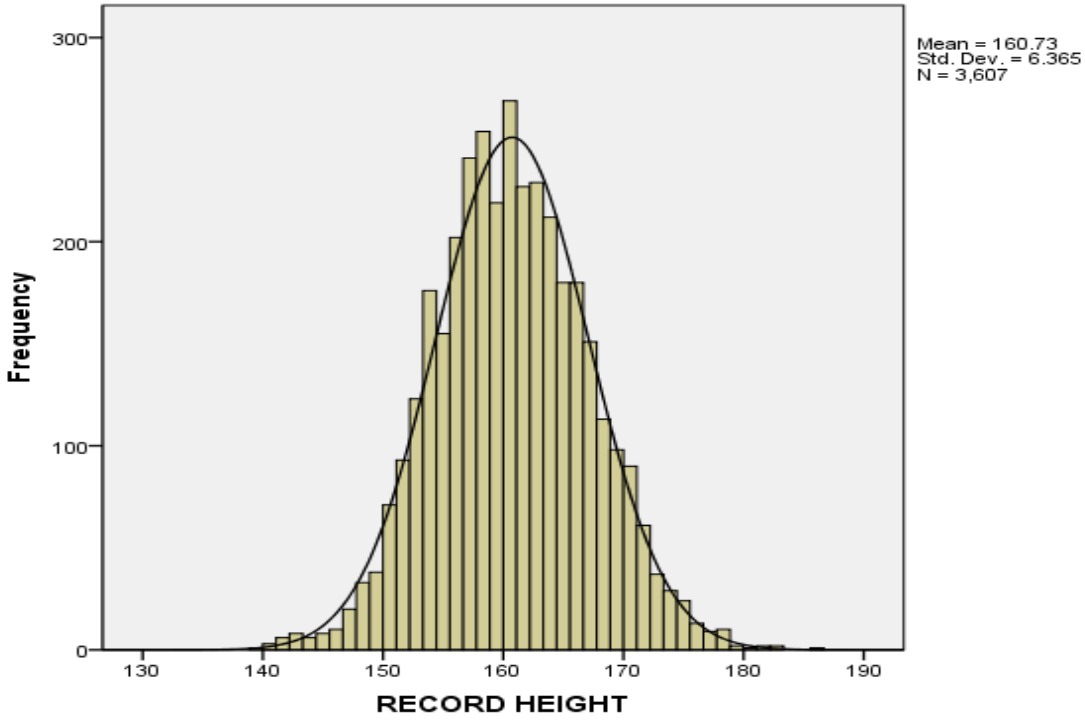
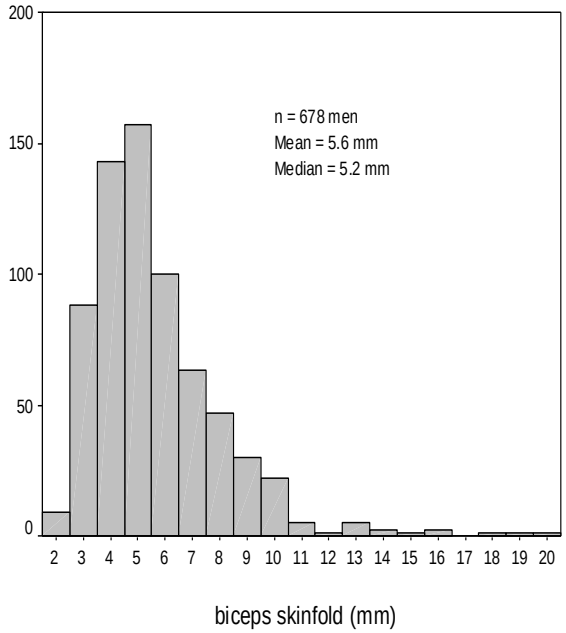


Figure 3. Upper plot: the ‘Normal’ distribution.
Lower plot: the height of 3,607 adult women recorded in the Scottish Health Survey, 1998 now with the (theoretical) ‘Normal’ curve superimposed.

When the distribution is skewed (that is not as a bell shape) the median is a better measure of location than the mean ('typical' values) because it is less influenced by outliers, as shown in the income example above. Also, the IQR is a better measure of the spread of observations than the SD for the same reason (Figure 4).

The data in Figure 4 has an IQR of 4.1 to 6.8 mm (25th to 75th percentiles). This means that 50% of the observations have values in this range (a difference of 2.7 mm). The maximum value is 20.4 mm and the 75th percentile is 6.8 mm which means that the top 25% of observations lie between 6.8 and 20.4 mm (a difference of 13.6 mm) and an indication from the numbers alone that the distribution of observations is likely to be skewed.



Biceps skinfold (mm)	
n	678
Mean	5.6
Median	5.2
SD	2.3
Minimum	2.2
Maximum	20.4
Percentiles:	
25%	4.1
50%	5.2
75%	6.8
IQR	4.1 to 6.8

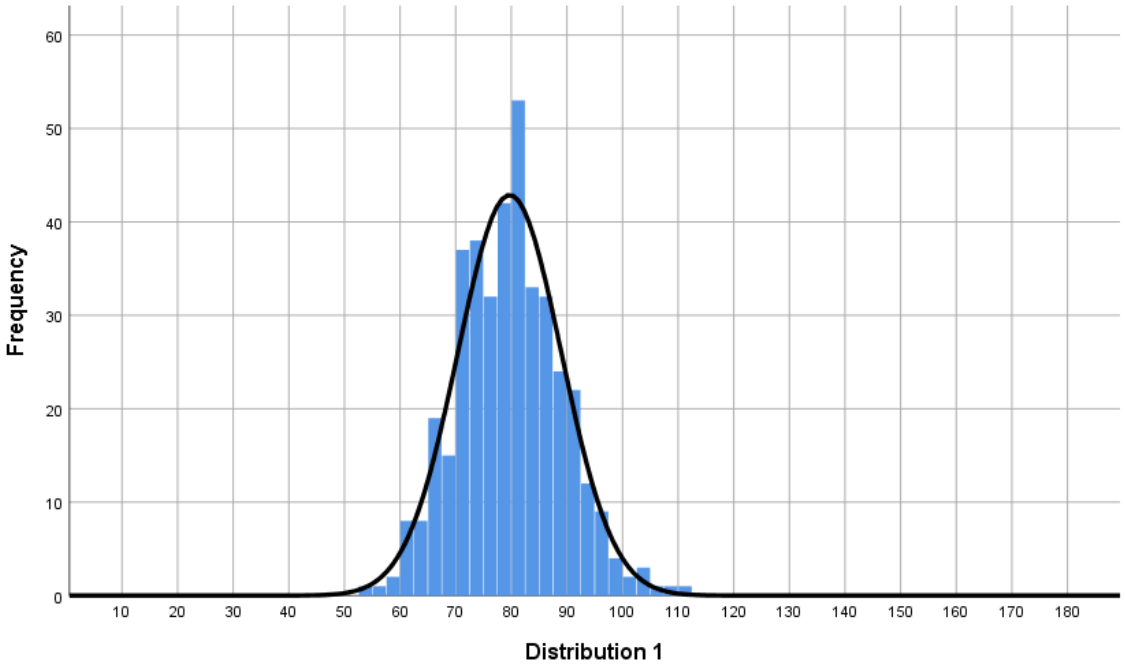
Figure 4. An example of a skewed distribution: the biceps skin fold thickness in 678 men

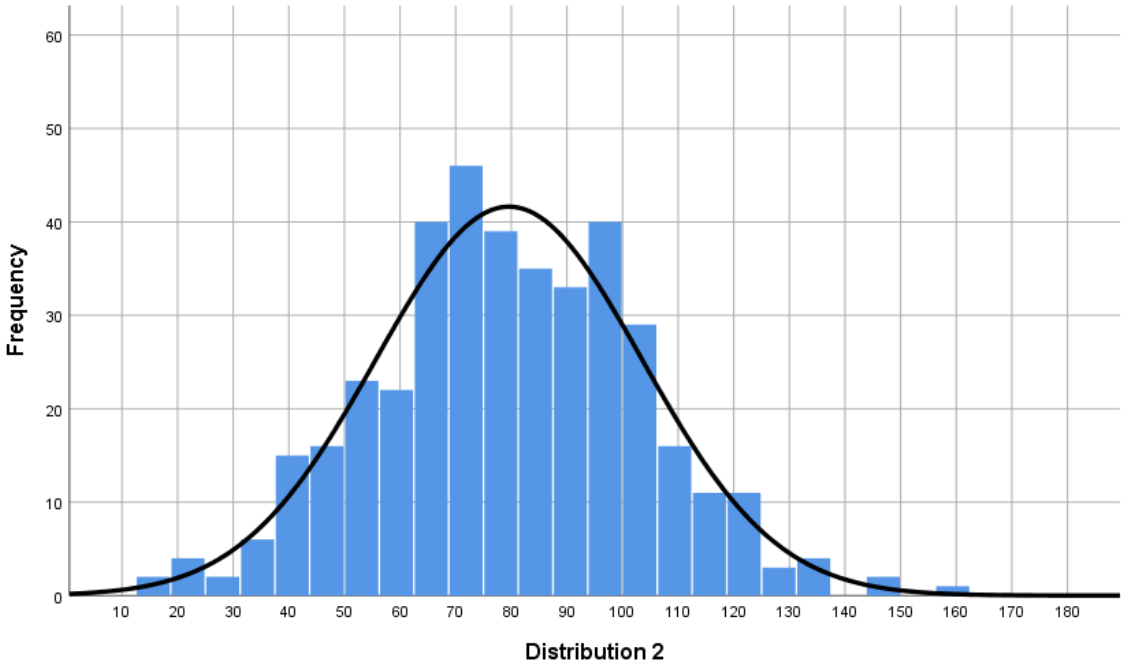
Quiz 3: True or false?

1) Marital status is a categorical variable
2) The mean is a better measure of central tendency when the distribution of data is skewed (i.e. does not conform to a 'bell shape')
3) The variability of a set of data with a skewed distribution is best described by the standard deviation
4) The standard deviation of a set of data is derived from the individual data values

Answers in Appendix 3

Quiz 4:
The 2 distributions below have the same mean (=80) and are plotted with the same scales. Which distribution has the greater variance and standard deviation?
Answer in Appendix 3





(8) The Standard Normal (Gaussian) Distribution

The shape of the distribution of variables has been studied for over 250 years. In the 19th Century Adolphe Quetelet, a Belgian mathematician and Sir Francis Galton, an English scientist had noted that many characteristics (such as weight, standing height, chest circumference) when measured on a large number of people and plotted as a histogram showed the same pattern, namely that of a bell-shape. Their work led to the concept of the Standard Normal distribution (Figure 5).

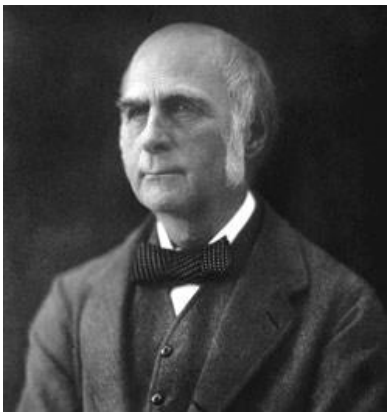
The Standard Normal distribution is a **theoretical** distribution with the following properties:

Bell-shaped
Unimodal (one 'hump'),
Symmetrical (balanced) about the mean,
Mean = Median = Mode = 0
Standard Deviation = 1

The distribution is designated N(0,1) where 'N' refers to 'Normal', '0' refers to its mean and '1' to its standard deviation. The distribution is characterised completely by these two parameters, the mean and standard deviation (SD).

The distribution is used to derive probability estimates. For example, if a person is chosen *at random* from a population there is a 68% chance that his/her value for a measurement that is Normally or approximately Normally distributed will lay within the range of values that is 1 SD below the mean and 1 SD above the mean (denoted as mean +/- 1 SD, refer back to Figure 3, the example of height in adult women). Similarly, there is a 95% chance that his/her value will lay within the range of values that is 1.96 SD (or approximately 2 SD) below the mean and 1.96 SD above the mean (denoted as mean +/- 1.96 SD). In consequence, there is only a 2½% chance that his/her value will be *smaller* than the mean – 1.96 SD or *greater* than the mean + 1.96 SD.

Adolphe Quetelet 1796-1874



Sir Francis Galton 1822-1911

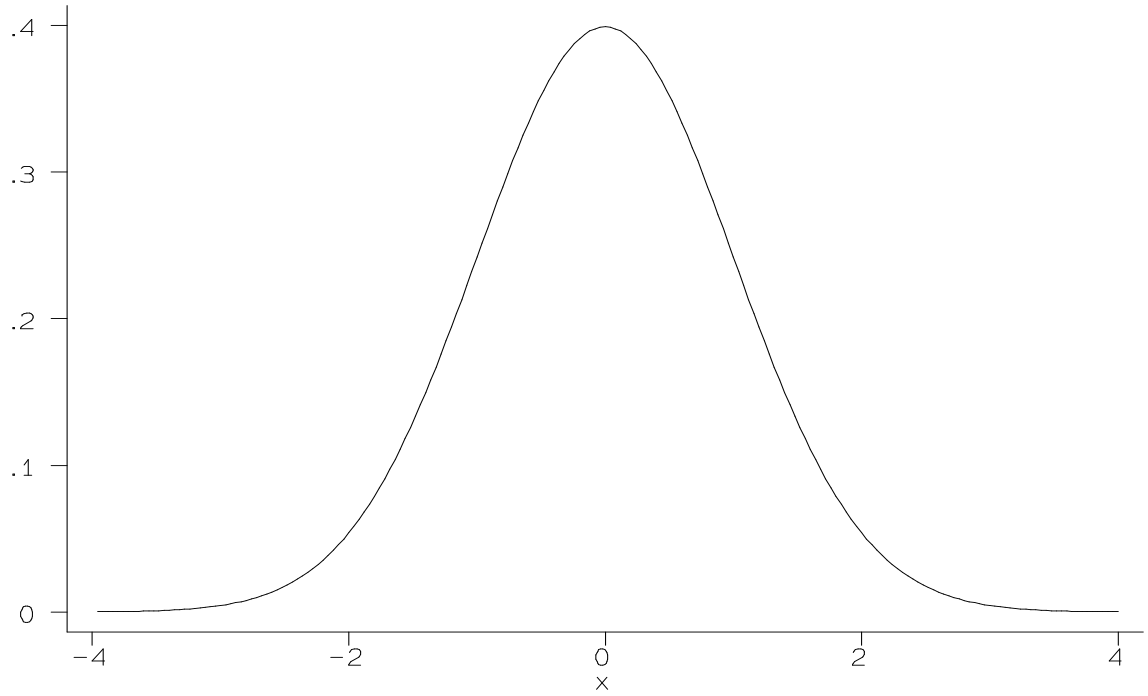


Figure 5. The Standard Normal distribution, N(0,1). Mean = 0, Standard Deviation = 1
68% within ± 1 SD 95% within ± 1.96 SD 99% within ± 2.58 SD

It is often useful to know how many standard deviations a given value is from the mean. This value is referred to as the Z-score. We calculate a Z-score to fit it into the Normal distribution (N(0,1)) using the mean and SD of the measure of interest. The Z-score is:

Z-score= (observed value – population mean) / Standard Deviation (equation 4)

Example: What is the Z-score for height in an adult male 167.4 cm tall? First, we need to know the mean and SD of height in the male population (*see* Figure 6 for an example of population values).

The estimated average height of men in Scotland (from the Scottish Health Survey of 1998) is 174.2 cm with an SD of 7.2 cm. So, for our 167.4 cm tall man, his Z-score is:

Z-score= (167.4 – 174.2) / 7.2 = -1.0

Hence, this man is 1 standard deviation *below* the population mean. Next, we need to look up the probability distribution in the statistical tables for the Standard Normal distribution (Appendix A). The table in Appendix A lists the areas in the tail of the Standard Normal distribution. Find the value of Z that is 1.0; look down the first column to row marked ‘1’ then read along to column headed ‘0’. The value listed is 0.159 which, because the Z-score is negative means that 15.9% of men are *shorter* than this particular man, and hence 84.1% are taller.

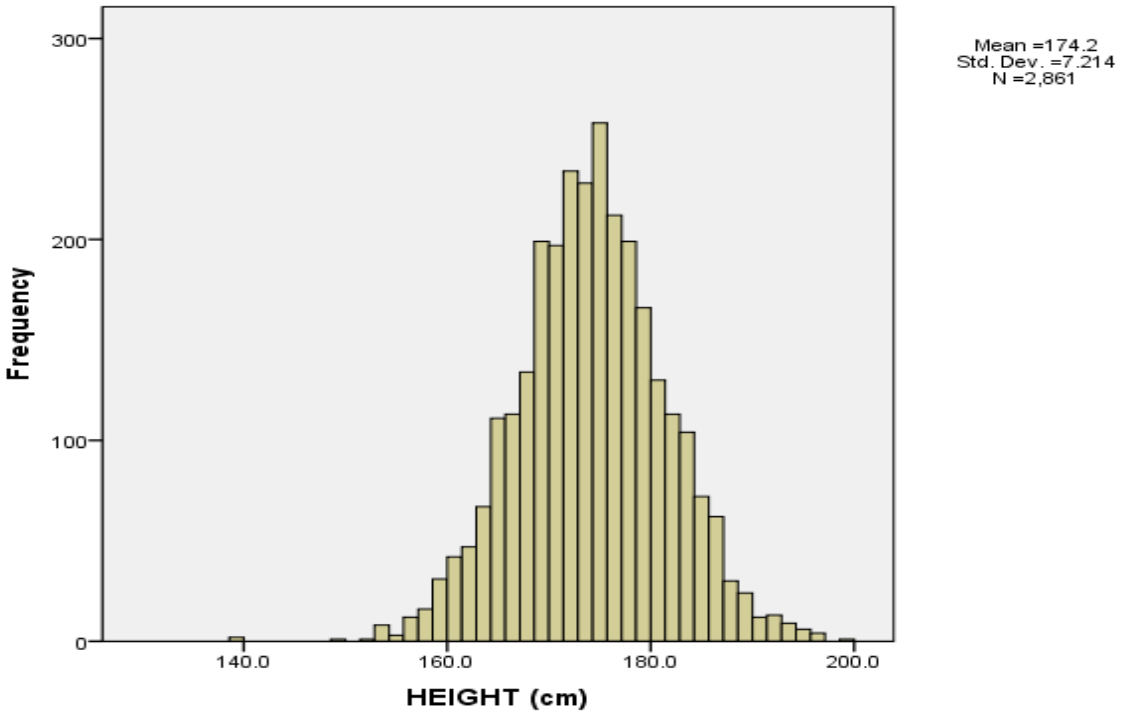


Figure 6. Height of 2,861 adult men recorded in the Scottish Health Survey, 1998. Mean = 174.2 cm, standard deviation = 7.2 cm

Exercise: try calculating the Z-score for your own height using these values for the population mean and standard deviations:
Women: mean = 160.7 cm, SD = 6.4 cm
Men: mean = 174.2 cm, SD = 7.2 cm
Now look up the scores in Appendix A to find out where you lie in the distribution. What proportion of the population is shorter than you? What proportion of the population is taller than you?

Some interesting facts about Adolphe Quetelet and Sir Francis Galton:

Quetelet developed an index of 'adiposity' (the 'Quetelet' index) which is a person's weight in kilograms divided by the square of their height in metres [= weight / (height)^2]. This is better known these days as the Body Mass Index (BMI).

Galton was a cousin to Charles Darwin. Galton's work on the markings on our finger tips led to the use of finger print identification of individuals.

To use the Standard Normal distribution we actually need to know the *true population mean* (symbol=μ) and *true SD* (symbol=σ). Because they are population values they are called *parameters*. In the height example above, we have assumed the estimates are true parameters whereas, in reality, they are only sample values, and therefore estimates. However, we seldom know the true values and, instead, have to estimate them by taking a large enough sample that we hope is representative of the population. Then, we use a 'working' distribution, called the t-distribution, which is an approximation of the Standard Normal distribution (the theoretical distribution). The probabilities listed in the t-distribution table take into account the levels of uncertainty associated when relying on a smaller number of observations (*explained further below*).

(9) The t-distribution

The t-distribution is bell-shaped but flatter and wider than the Standard Normal distribution. It differs from the Standard Normal distribution because sampling variation means that when the number of observations is low there is greater uncertainty and more likelihood of departure of the distribution from that expected in the Normal distribution. The t-distribution is more spread out than the Normal distribution and the amount of spread depends on the sample size. Generally, the more observations we have the more confident we can be about the shape of the curve and hence the t-distribution approximates the Standard Normal distribution as the number of observations increases.

If the number of observations is 30 or more then the shape of the t-distribution is very similar to the Standard Normal distribution (Figure 7). Here the number of SD units needed to define the 95% range of values is 2.04 (remember, for the Standard Normal distribution this value is 1.96). In Figure 7 we refer to t(29) to denote that we are using estimates based on 30 observations. The value of 29 in the brackets refers to the *'degrees of freedom'* (df) which is a statistical concept.

Degrees of freedom (df) refers to the number of sample values that are free to vary. In a sample, all but one value is free to vary, and the degrees of freedom is then n-1 where n is the number of observations. For example, consider a set of four values with the mean of 5 and a sum of 20. If you are asked to 'invent' the individual four values then you are only 'free' to invent three of them as the fourth must ensure the sum adds to 20 (note, it can be a negative number). Hence, in this example, the degrees of freedom are 3.

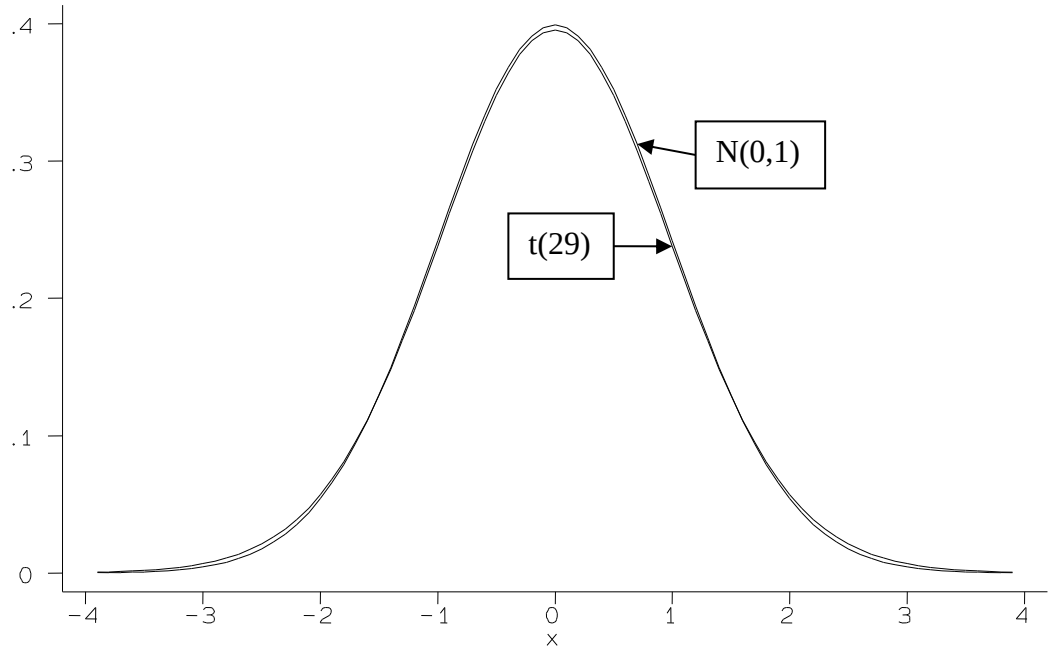


Figure 7. The t-distribution based on 30 observations, t(29), 'x' refers to the number of standard deviations, 95% of observations are within ± 2.04 SD
The overlap with the Standard Normal Distribution N(0,1) is very good.

When the number of observations is only 10 the shape of the t-distribution departs more from that of the Standard Normal distribution (Figure 8). Here the number of SD units needed to define the 95% range of values is 2.26 (and the degrees of freedom are 9).

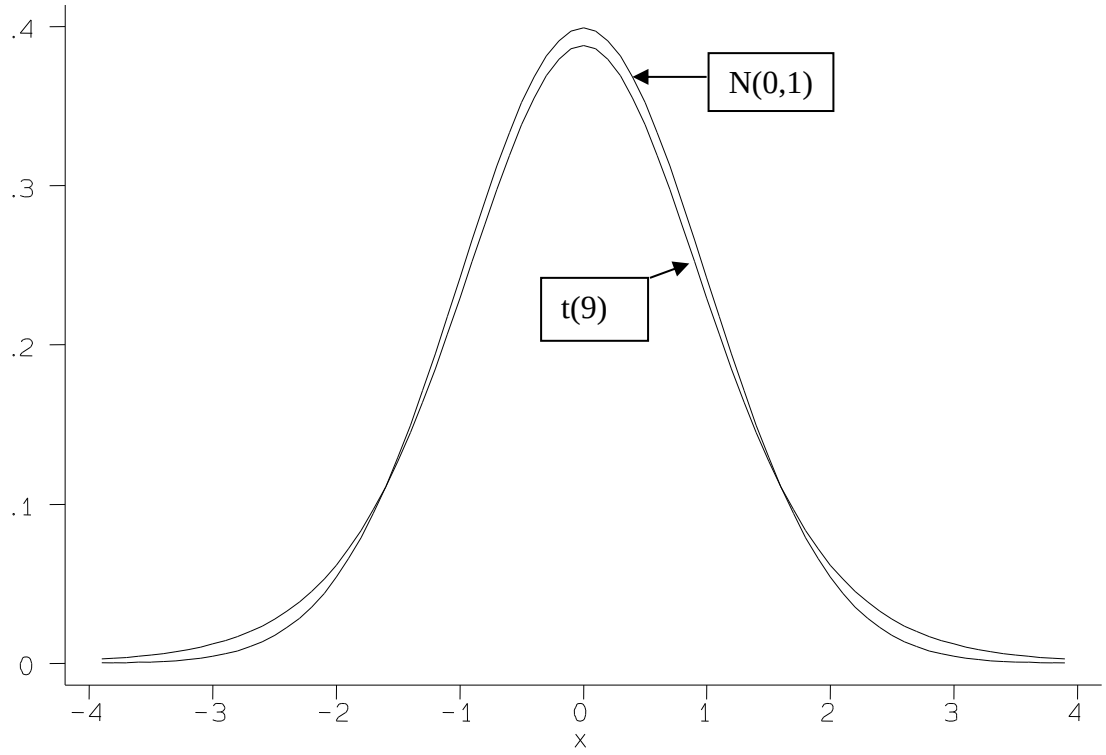


Figure 8. The t-distribution based on 10 observations, t(9), 'x' refers to the number of standard deviations, 95% of observations are within ± 2.26 SD
The overlap with the Standard Normal Distribution N(0,1) is less good than with t(29) above.

If the number of observations is only 2 then the shape of the t-distribution departs markedly from that of the Standard Normal distribution due to greater uncertainty in the accuracy of estimates of the true values of the mean of the population from which the 2 values were chosen (Figure 9). Here the number of SD units needed to define the 95% range of values is 12.7 (and the degrees of freedom are 1).

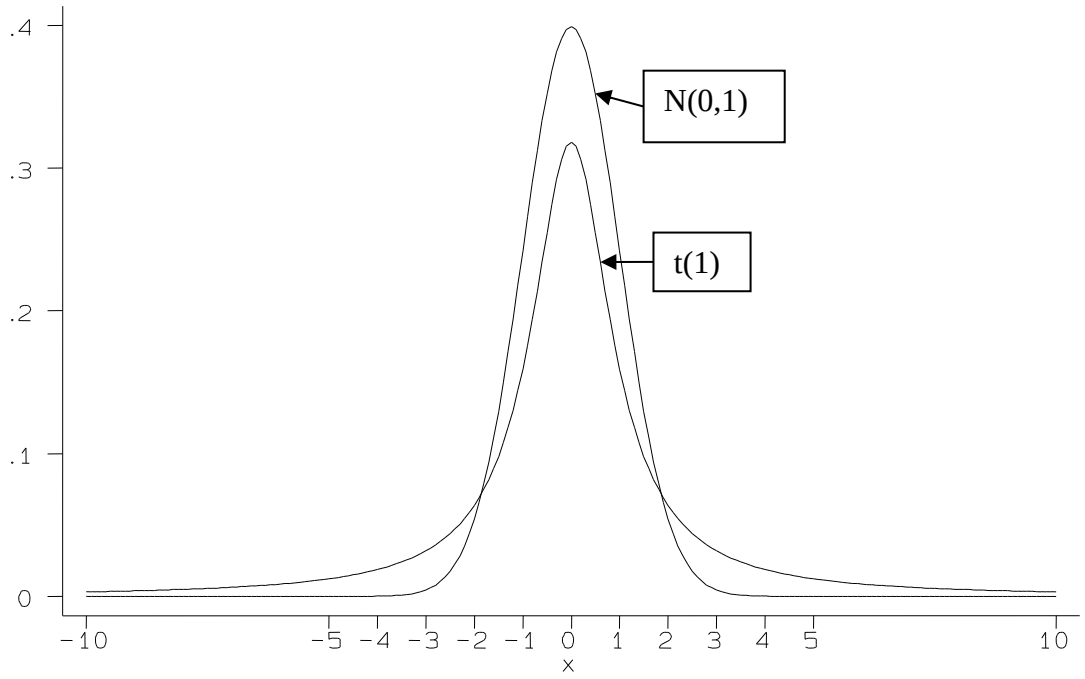


Figure 9. The t-distribution based on 2 observations, t(1), 'x' refers to the number of standard deviations, 95% of observations are within ± 12.7 SD
The overlap with the Standard Normal Distribution N(0,1) is very poor.

As a general rule, use the t-distribution for estimates of the number of SDs to define the 95% range of values from a given distribution of a sample, particularly if the number of observations is less than 30.

As 'n' becomes larger the t-distribution becomes more like the Standard Normal distribution until it reaches infinity (*symbol,* ∞) where both distributions are identical (Table 5). Statistics packages take into account the number of observations and, in reality, provided 'n' is large (>30) then most statistical methods are robust against moderate departures from 'normality'.

Table 5. Critical values for the t-distribution
('n' = number of observations, 'df' = degrees of freedom)

		Number of Standard Deviations to define a given interval		
Interval:		90%	**95%**	99%
n: 2	df: 1	6.31	**12.71**	63.66
10	9	1.83	**2.26**	3.25
30	29	1.70	**2.04**	2.76
101	100	1.66	**1.98**	2.63
∞	∞	1.65	**1.96**	2.58

The distribution of many variables in medicine is bell-shaped. Examples of variables which have a Normal (or approximately Normal) distribution are standing heights in adulthood, and blood pressure, haemoglobin concentration and lung capacity in *healthy* people. However, there are examples of variables which do not fit with a Normal, bell-shaped distribution. These include skin fold measurements (*see* Figure 4 above) and length of stay (Figure 10) both of which are skewed to the right (positive skew). Some measures are skewed to the left (negative skew) and these include

gestational age at birth (Figure 11). Many ‘quality of life’ measures can be positively or negatively skewed depending on how they are scored. This can present problems when analysing data as the shape of the distribution determines the choice of summary measures (mean versus median, SD versus IQR) and the statistical test chosen when comparing groups (*see* NHS Fife Study Guide 12: ‘How to Choose a Statistical Test’).

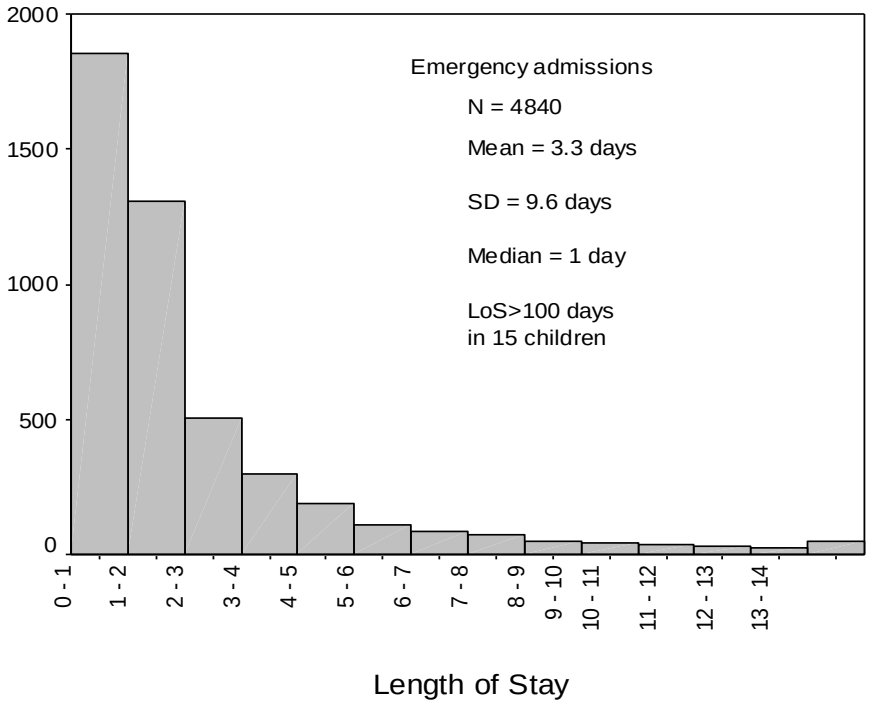


Figure 10. Length of stay (days) in 4840 children admitted to hospitals in one Hospital Trust over 3 years. An example of a positively skewed distribution. The mean value of 3.3 days is not a good measure of a ‘typical’ (or average) value when the median is only 1 day (and hence 50% of children are admitted for a day and 50% for longer than a day)

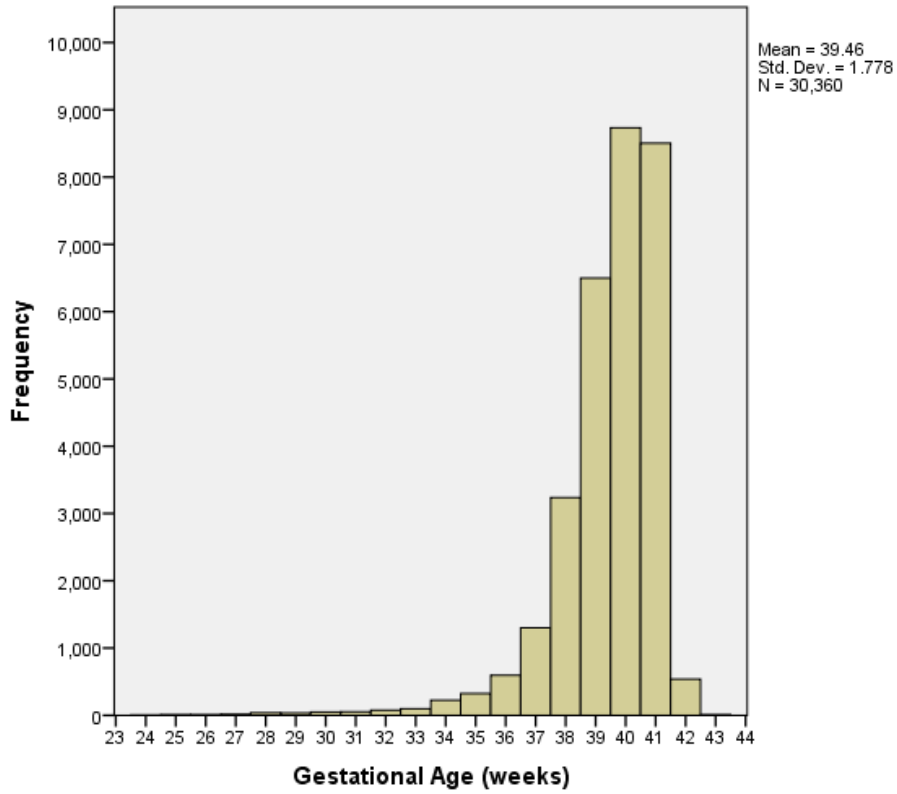


Figure 11. Gestational age in 30,360 births in Fife, 2003 - 2012. An example of a negatively skewed distribution.

Other examples of measures that have a slightly skewed distribution are weight in women (Figure 12) and age at first pregnancy (Figure 13).

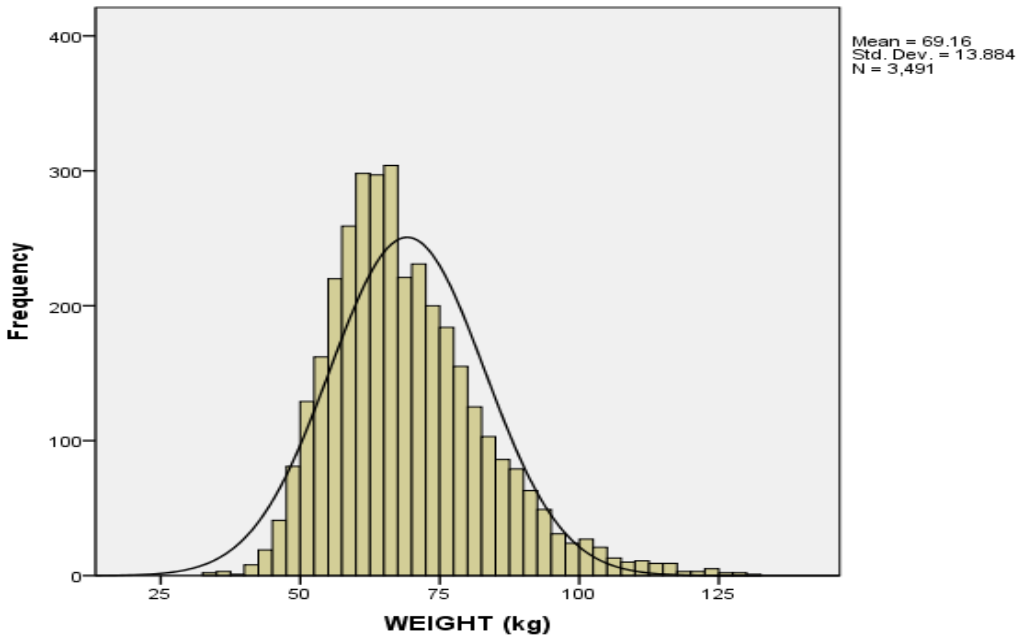
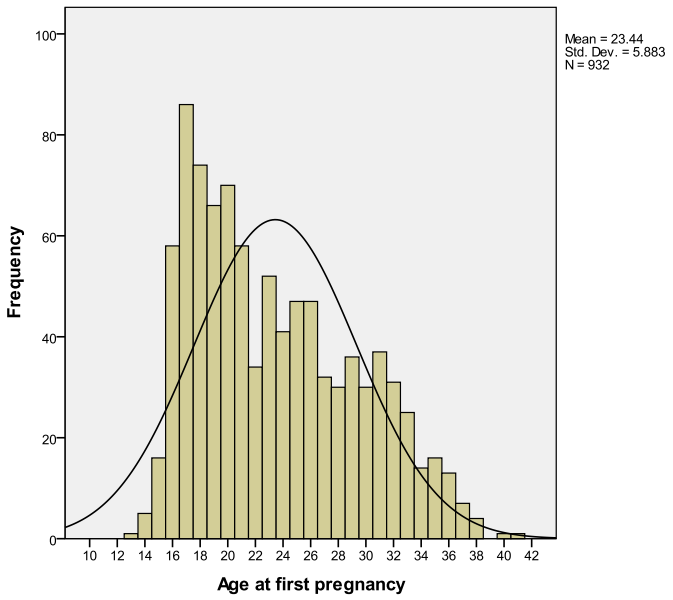


Figure 12. Weight (kg) of 3,491 adult women (Scottish Health Survey 1998), overlaid with the Standard Normal curve to show where the data depart from it.



Statistics

Age at first pregnancy

N	Valid	932
	Missing	7
Mean		23.44
Median		22.00
Std. Deviation		5.883
Minimum		13
Maximum		41
Percentiles	25	18.00
	50	22.00
	75	28.00

Figure 13. Age at first pregnancy in 932 women, overlaid with the Standard Normal curve.

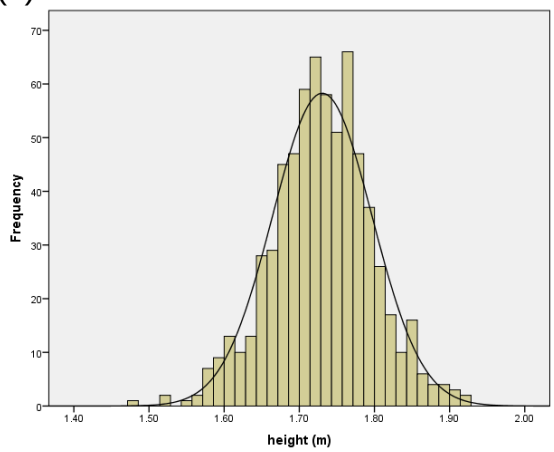
The summary statistics show the effect on the mean and median when the distribution of the data does not conform to the Normal distribution. The mean (mathematical average) is 23.4 but this value is less good as a 'typical' result because the median is only 22.0, suggesting that 50% of the women are actually younger than this age. Furthermore, the first 25% of women are aged 13 (the minimum) to 18 (25th percentile, so 5 years difference) whereas the upper 25% are aged 28 (75th percentile) – 41 (the maximum, so 13 years difference).

(10) How to check if a distribution is normal?

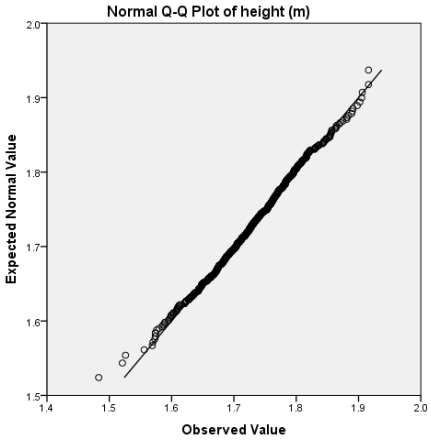
Visual checks

First, generate a histogram, *box and whisker plot* or one of the *probability plots* that can be run with statistics packages (Figures 14 and 15). A histogram should look bell-shaped. A box and whisker plot has a shaded box where the lower and upper boundaries represent the 25th and 75th percentiles, respectively. The thicker line across the shaded box represents the median. The 'whiskers' can represent the minimum or maximum values or, when there are *outliers* they mark a value that is 1.5 times the inter-quartile range (IQR). Outliers, when marked, are values that lie between 1.5 IQRs and 3 IQRs from the end of the box. *Extreme* values are those that lie greater than 3 IQRs from the end of the box. For a variable that is Normally distributed the median line in the shaded box should appear in the middle with the length of the whiskers the same on either side of the box. A probability plot such as the Q-Q plot (Q stands for quantile) is a plot of each observed value against that expected if the distribution of the variable fitted perfectly with the Normal distribution. For a variable that is Normally distributed the points should lie along the *line of identity* where the observed value equals that expected.

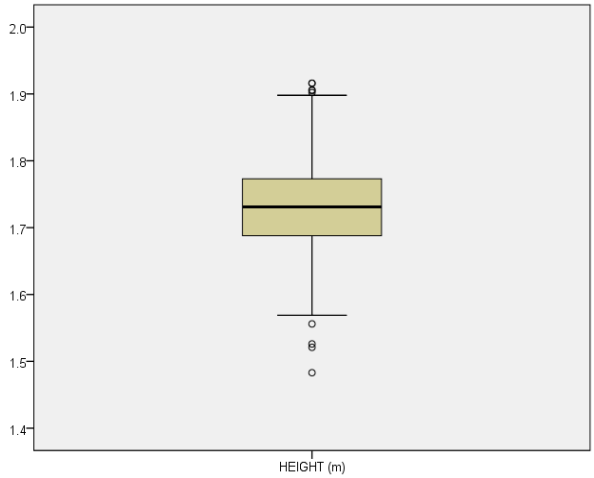
(a)



(b)



(c)

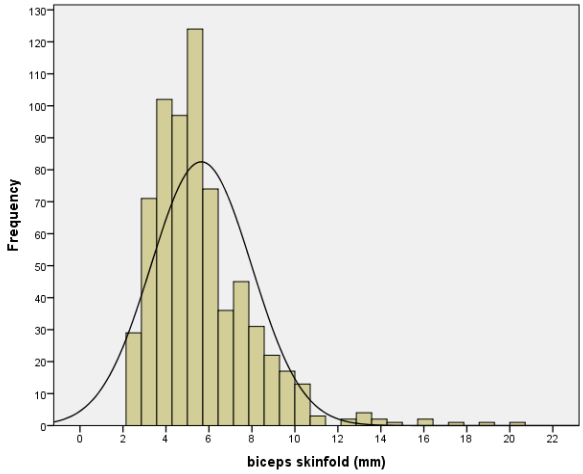


		Height (m)
N	Valid	678
Mean		1.7305
Median		1.7310
Std. Deviation		.06630
Skewness		-.111
Kurtosis		.351
Minimum		1.48
Maximum		1.92
Percentiles	25	1.6880
	50	1.7310
	75	1.7730

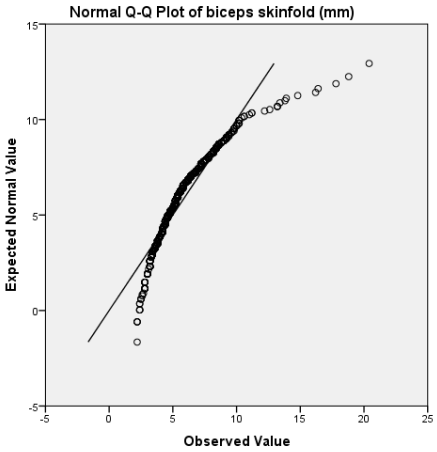
Figure 14. Histogram (a), Normal Q-Q plot (b), Box and whisker plot (c) and summary statistics of a Normally distributed variable: Height in 678 men.

Note the mean and median are identical, as would be expected when the distribution of the variable conforms to a Normal distribution (bell-shaped).

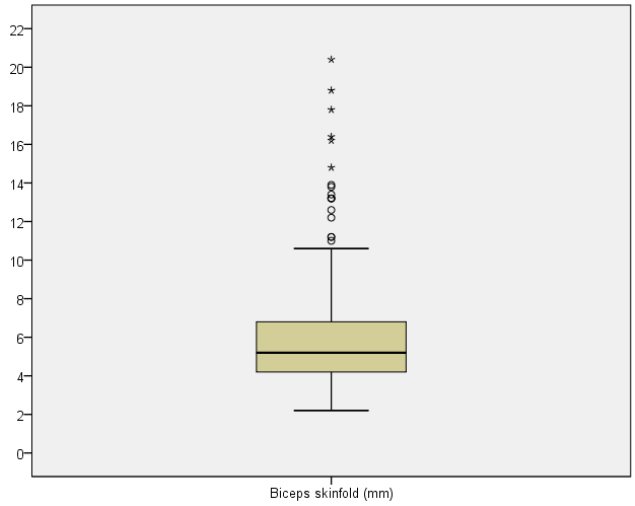
(a)



(b)



(c)



		biceps skinfold (mm)
N	Valid	678
Mean		5.642
Median		5.200
Std. Deviation		2.3417
Skewness		1.841
Kurtosis		6.023
Minimum		2.2
Maximum		20.4
Percentiles	25	4.175
	50	5.200
	75	6.800

Figure 15. Histogram (a), Normal Q-Q plot (b), Box and whisker plot (c) and summary statistics of a variable that has a skewed distribution: Biceps skin fold thickness in 678 men.

The circles above the upper whisker are individual outliers defined as values between 1.5 IQRs and 3 IQRs from the upper boundary of the shaded box (the 75th percentile). The stars are individual extreme values defined as values greater than 3 IQRs from the upper boundary of the shaded box.

Note the thick black line (the median) across the box is not in the middle and the whiskers are of different lengths. Compare with Figure 14 above.

Note the mean and median are not identical, as might be expected when the distribution of the variable is skewed and does not conform to a Normal (bell-shaped) distribution.

Checking the numbers

Next, compare the mean and median, and the SD in relation to the mean. The mean should equal the median for a variable that is Normally distributed. When the data values can only be positive the data are likely to be skewed if the mean is less than twice the SD. However, you need to be aware that even if the mean is more than twice the SD this does not guarantee that the data are Normally distributed!

Another approach is to look at the relationship between the mean and SD when the data are split into groups. The test is to see if the SD of the subgroups is independent of the mean (that is, does the SD remain about the same even if the mean of the different subgroups varies) or is the SD proportional to the mean (where the SD increases as the mean increases). A proportional relationship suggests the data are not Normally distributed.

Example of non-Normally distributed data: Urinary cotinine excretion (µg/mg) related to cigarette consumption (Altman & Bland *BMJ* 1996; **313**: 1200) – *these data were taken from a paper that had reported the mean and Standard Error (SE) of the mean but when Altman & Bland calculated the SD it was clear that the data were not Normally distributed because the mean was less than twice the SD and a negative value for urinary cotinine is not possible.*

Another indication was from the proportional relationship of the SD to the mean. Note how the SD (and SE) increases with increasing mean values. In this study the authors were not justified in their choice of the t-test to analyse their raw data because the measurements were not Normally distributed. Instead, the data should have been log transformed before using a t-test (more on this later).

Gigs / day	n	mean	SE	SD
1-9	25	0.31	0.08	0.40
10-19	57	0.42	0.10	0.75
20-29	99	0.87	0.19	1.89
30-39	38	1.03	0.25	1.54
>40	28	1.56	0.57	3.02

(11) What to do if a distribution is not Normal?

First, you must ask, does it matter? If you are comparing two groups you will want to use a test of statistical significance. Using a t-test assumes the distribution of the data is Normal, or approximately Normal and the variance (spread of data) is about the same in each group. The t-test is called a *parametric* test. To use a t-test on data that are not Normally distributed (skewed either positively or negatively) you must *transform* the data by subjecting it to a mathematical function so that it does fit a Normal, or approximately Normal distribution. Data can be transformed by taking logarithms (to base 10) or calculating the reciprocal (1/x) for positively skewed, and square (x^2) or cube (x^3) functions for negatively skewed data. But watch out for zero and negative values as these cannot be transformed using logarithms. Using a t-test on data that are not Normally distributed can lead to the wrong conclusions! (*See* NHS Fife Study Guide 13: 'How to make sense of numbers' for an example in which using a t-test inappropriately gave the wrong interpretation)

Some notes on logarithms: Any positive number can be expressed as a logarithm. Logarithms are calculated to the base 10 or base *e* where *e* is a constant (=2.71828….). Hence, a positive number, *x*, expressed as a logarithm to base 10 is 10^x. For example, the logarithm to base 10 of 100 is 2, where $100 = 10^2$. The logarithm of 1000 is 3, where $1000 = 10^3$ (or 10 x 10 x 10). In this account we are only concerned with logarithms to base 10. A logarithm cannot be derived for a negative number or zero.

The decision whether to transform the data and use a parametric test depends on the question you are asking. For example, if you want to know simply if two groups differ (yes or no) then you can use a *non-parametric* test which makes no assumptions about the distribution of the data. However, if you want to know *by how much* the two groups differ then you will need to transform the data and use a parametric test such as the t-test.

An example of a transformation is that for the skin fold data from Figure 4 (page 13). The data values have been converted to logarithms (log-transformed) which changes the shape of the distribution to fit more closely with that for a Normal distribution (Figure 16).

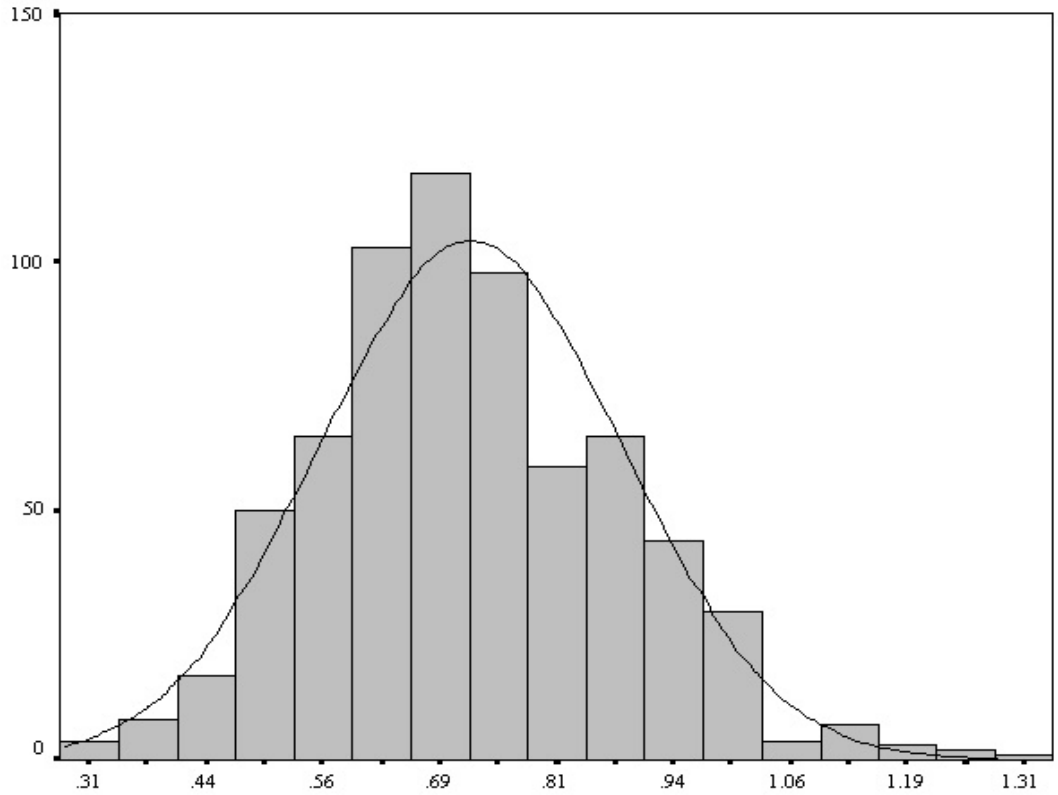


Figure 16. The skinfold thickness data from Figure 4 after log-transformation.

(12) Confidence Intervals for a Continuous Variable

Statistics is concerned with estimation and describing 'uncertainty'. A population is unique; a sample taken from it is not. Many different samples of varying sizes may be taken from a population. We take a sample that we hope is representative of the

population from which it is drawn and derive the summary statistics (mean, median and SD) which, together with the number of observations the sample is based on are used to derive a range of values (called the confidence interval) in which we believe the true values of the mean (μ) and SD (σ) in the population are likely to lie (Table 6).

Table 6. Relationship between a sample and the population from which it is drawn

	Reality, '=Truth'	Estimate
Population:	Size = N	Sample: size 'n'
Mean	μ	$\bar{x}$
Variance	σ^2	SD^2 (symbol s^2)
SD	σ	SD, (symbol *s*)
		↓
		Standard Error of the mean, SE ($\bar{x}$) = SD /$\sqrt{n}$
		↓
		Confidence Interval of the estimated mean

The sample mean ($\bar{x}$) is unlikely to be exactly the same as the population mean (μ). A different sample would give a different estimate of the population mean and the difference would be due to *sampling variation*. The distribution of the mean values of a variable taken from many samples of a population will be Normally distributed, provided the variable in the population is also Normally distributed. We could calculate the *standard deviation of the mean values* and this, in fact, is referred to as the *standard error of the mean* SE ($\bar{x}$). The SE is a measure of the accuracy of the estimated mean (*see* Glossary). But, we do not need to take many samples and can calculate the SE from a single sample:

Standard Error of the mean, SE ($\bar{x}$) = SD of sample /$\sqrt{n}$ (equation 5)

where 'n' is the number of observations in the sample.

The SE thus calculated from a single sample is an estimate of the precision of the sample mean. The size of the SE depends on the degree of variation in the population and on the size of the sample, the larger the sample, the smaller the SE.

The SE is used to derive a *confidence interval* (CI) in which we believe the true mean (μ) for a population is likely to lie. It has a probability (for example, 95%) attached to it to give it an element of precision. It is most common to report the 95% confidence interval but other intervals (e.g. 90%, 99%) can also be cited.

The process is to select a representative sample of size 'n' from the population and calculate its mean and SD, then calculate the SE (= sample SD /$\sqrt{n}$) and derive the confidence interval:

Confidence interval = sample mean ± t_α x SE ($\bar{x}$) (equation 6)

t_α is the *number of SDs* from the Normal distribution needed to capture the desired percentage of observations. For a 95% CI the value of t_α is 1.96 (see Figure 5), hence equation 6 becomes:

95% CI = sample mean ± 1.96 x SE ($\bar{x}$) (equation 7)

The 95% CI is a range of values in which we are *95% confident that the true mean (μ) for a population is likely to lie.*

Using 1.96 to define the interval is acceptable when you have a sufficiently large sample but the situation is different when you have much smaller samples. Then, you use the t-distribution to define the interval because this will be larger when based on small numbers due to the greater level of uncertainty in the assumptions. If the sample size is 20, t_α = 2.093, and if the sample size is 30, t_α= 2.045 (*see* Appendix B). As an example, a study was undertaken to estimate the number of hours of pain relief in 7 patients with arthritis. The hours of relief were 2.2, 2.4, 4.9, 3.3, 2.5, 3.7, and 4.3. The mean was 3.33 hours and the SD 1.03 hours. The SE of the mean was 1.03/ $\sqrt{7}$ = 0.39 hours. The degrees of freedom are 6 (based on 7 observations, so n-1) and the t_α = 2.45 (*see* Appendix B). Hence the 95% CI is:

95% CI = 3.33 ± 2.45 x 0.39 = **2.4 to 4.3 hours**

Interpretation: we are 95% confident that the *true mean* number of hours of pain relief of the *population of patients with arthritis* (from which the sample was drawn) lies between 2.4 and 4.3 hours.

If we had used 1.96 to define the interval (t_α) for 95% confidence in this sample of only 7 patients we would have obtained an incorrect, narrower range of values (actually 2.6 to 4.1 hours).

If we wished to derive a *99% confidence interval* in the above example, we would need to use t_α = 3.71 (see Appendix B).

99% CI = 3.33 ± 3.71 x 0.39 = **1.9 to 4.8 hours**

Interpretation: we are 99% confident that the *true mean* number of hours of pain relief of the *population of patients with arthritis* lies between 1.9 and 4.8 hours.

As a general precaution, if the sample size is less than 30 use the t-distribution to define the interval (t_α). If the sample size is more than 30 you can use either the t-distribution or the Standard Normal distribution to define t_α as the difference between them makes little difference to the estimated confidence interval (compare the values in Appendix B where the degrees of freedom exceed 30). Statistics packages should automatically use the correct distribution to define the interval for deriving confidence intervals.

(13) Worked Example on Confidence Intervals for a Continuous Variable

Consider the situation where we have a defined population and have a complete set of measurements for a variable (e.g. height) on all members of that population. Such a situation may arise for recruits in the army, for fire fighters, police officers, ambulance personnel and other groups which have a medical examination prior to entry into that profession. In this case the mean and standard deviation calculated from the full set of measurements represent *true* values for these parameters. Figure 17 shows the distribution of height for such a population. Let us assume they are female recruits in the army (n=3919). They all have had a medical examination in which their height was measured. The mean and SD of their height is 160.8 cm and 6.4 cm, respectively. These are the true mean (μ) and true SD (σ).

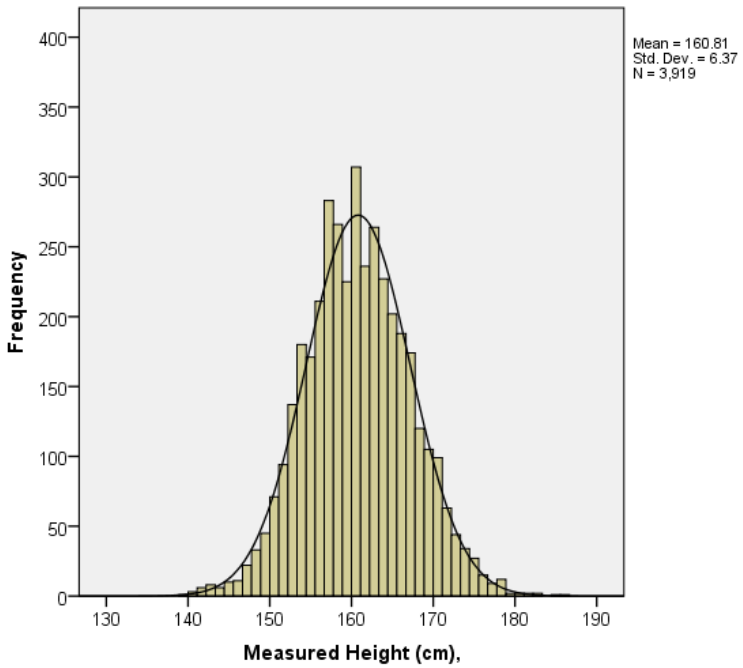


Figure 17. The Height (cm) of Female Recruits in the Army (n=3919).
Note: this is a fictitious example

Suppose the records of height are lost, including the summary measures, and new uniforms are being ordered for which the Ministry of Defence needs to know the true mean and SD of the height of the population of recruits before an order can be placed. Instead of repeating the height measurements of every recruit (expensive!) the MOD decides to take a random sample of a smaller number of recruits and use the results to estimate the true mean (μ) and true SD (σ).

A random sample of 32 recruits was identified. The mean height was 162.0 cm, the SD was 7.3 cm and the 95% CI (calculated from Table 6 and equation 6) was 159.4 – 164.6 cm. Hence, we are 95% confident that the true mean height of the population of 3,919 recruits lies between 159.4 and 164.6 cm. The width of the interval was 5.2 cm.

The 32 recruits were returned to the ranks and another random sample, this time of 114 recruits was selected. The mean height was 159.6 cm, the SD was 6.1 cm and the 95% CI was 158.4 – 160.8 cm. The width of the interval was narrower at 2.4 cm because it was based on a larger number of observations. With a sample of 366 recruits the 95% CI was even narrower (Table 7). This makes sense as the more observations you have the more confident you can be about the accuracy of the limits in which the true population mean is likely to lie.

The relationship between sample size and width of the 95% confidence interval can be seen in Figure 18 where we took multiple random samples of varying sizes and plotted the confidence intervals in relation to the true mean (160.8 cm). The confidence intervals included the true value in every case though one sample (n=114) had an upper limit which just included the true mean. Remember, these estimates are only associated with 95% confidence, so they are not guaranteed to include the true mean in every case. There is a probability of 5% that the 95% CI from any single sample will *not* contain the true mean. Remember also, that statistics are concerned with estimation and describing ‘uncertainty’!

The relationship between the width of the 95% confidence interval and sample size is further explored in Figure 19. There comes a time when the width of the interval (derived from the standard error, = SD/$\sqrt{n}$) reduces by only a small amount with increasing sample size. The added effort to obtain more data may not be justified as the confidence interval will not be much narrower. The size of the desired confidence interval, and hence the size of the standard error, is one of the characteristics needed when undertaking a power calculation to determine the sample size required for a particular study. By first specifying a desired standard error you can solve the equation to calculate the desired sample size, n = $(SD/SE)^2$. More details are given in NHS Fife Study Guide 11: 'How to calculate sample size and statistical power'.

Table 7. Relationship between estimated mean height measured from a sample and the true mean height of the population (of female recruits) from which it is drawn.

	Reality, 'Truth'	*Estimate Height from Sample (cm)*			
Population:	N=3919	Sample: size 'n':	32	114	366
Mean (cm)	160.8 (μ)	$\bar{x}$	162.0	159.6	161.1
SD (cm)	6.4 (σ)	SD	7.3	6.1	6.1
		↓			
		SE ($\bar{x}$) = SD/$\sqrt{n}$	1.29	0.57	0.32
		↓			
		95% CI of $\bar{x}$	159.4-164.6	158.4-160.8	160.4-161.7
		Width of interval	5.2	2.4	1.3

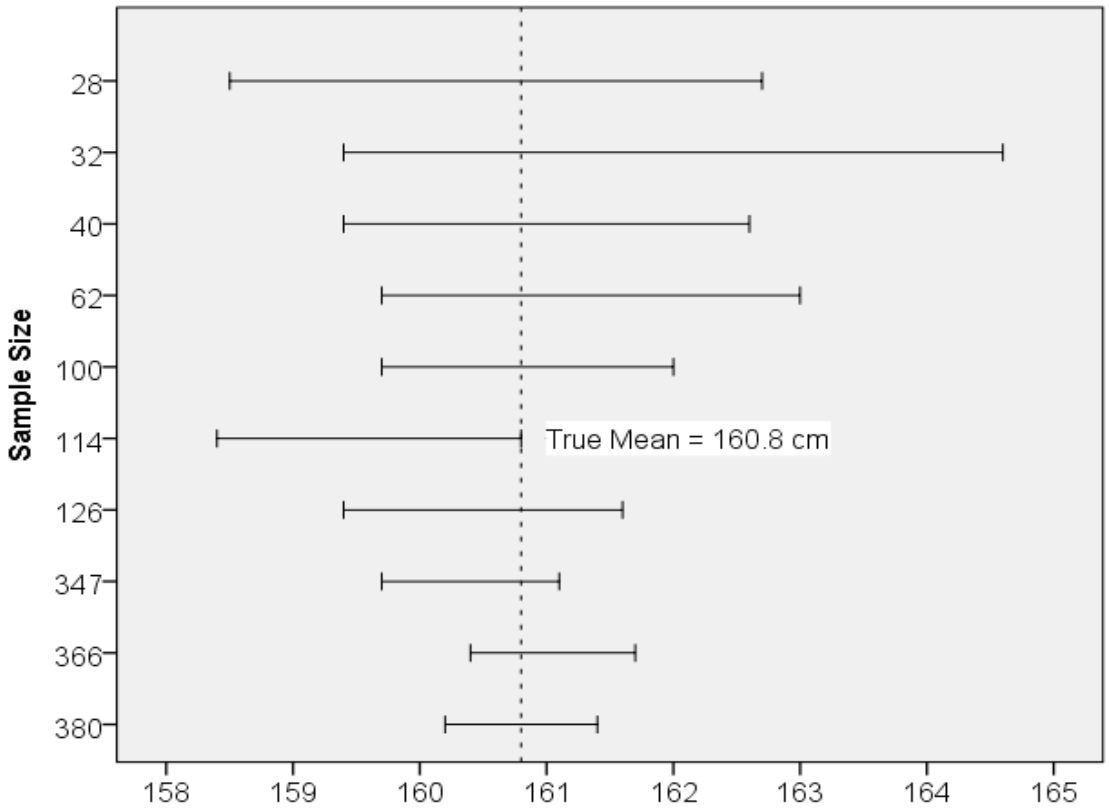


Figure 18. The 95% confidence intervals plotted against sample size. Multiple random samples taken from a (mythical) population of 3,919 values for height in female recruits. Note how the width of the interval becomes smaller with increasing sample size.

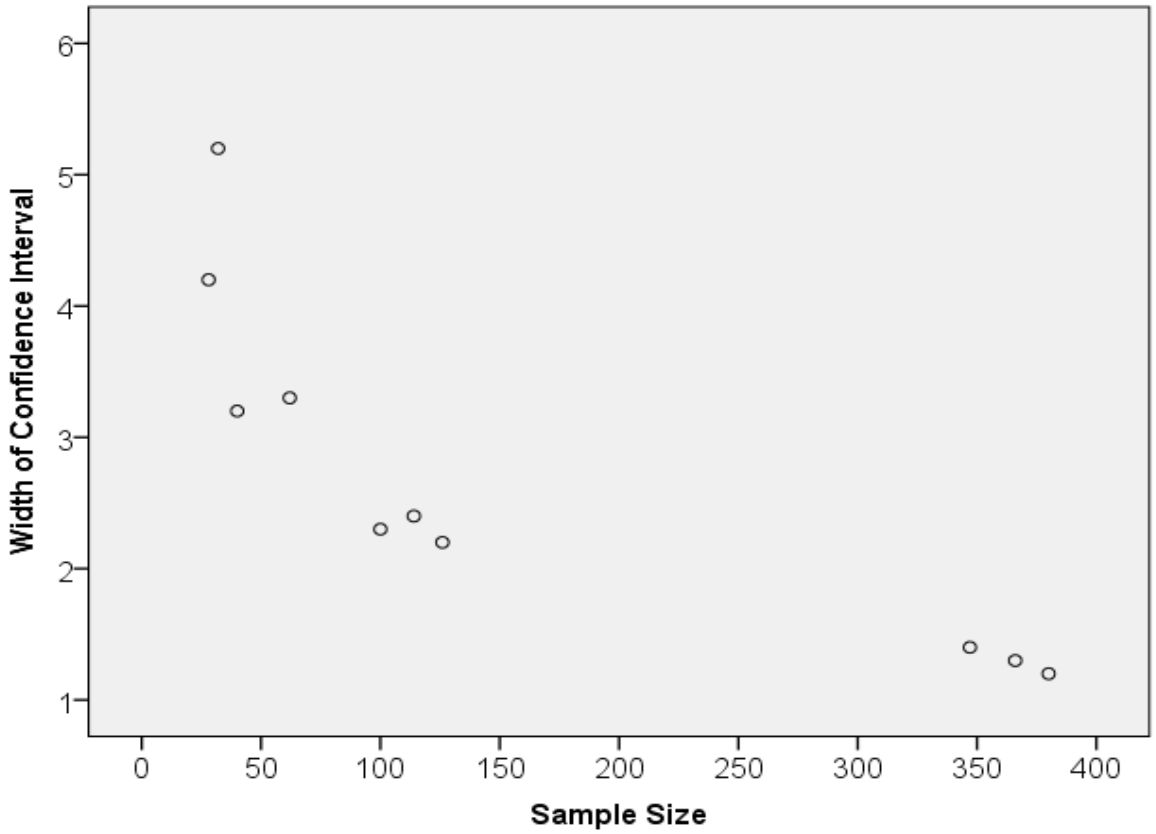


Figure 19. The relationship between the width of the 95% confidence intervals and sample size plotted in Figure 18 above.

(14) Confidence interval for a proportion (percentage).

A confidence interval can be calculated for any estimated parameter including a proportion (or percentage). The sampling distribution for a proportion follows a *Binomial* distribution (Binomial = only 1 of 2 choices). When the sample size is fairly large the sampling distribution of proportions from repeated samples is approximately Normal around the true mean for the population from which the samples are drawn. Hence, we can use the intervals (t_α) from the Standard Normal distribution to derive the confidence intervals. The sample size must be large enough for it to be safe to assume the Normal distribution applies. In addition, if the proportion itself is very small, and the sample size also is small, it is not appropriate to assume a Normal approximation. The rule of thumb is to use the Normal distribution when **both** ($n \times P$) and $n(1-P)$ are greater than 5, where n is the number of observations and P is the proportion. If these criteria are not met the calculations get much more complicated and we have to use the Binomial distribution to determine the precise confidence intervals (and beyond this simple guide).

As above, a narrow confidence interval means your estimate is precise whereas a wide confidence interval means it is less precise. A wide confidence interval can occur because it is based on a small number of observations, or that there is simply a large degree of variation in the individual data. To calculate a confidence interval (CI) for a proportion you need:

- The proportion (or percentage) itself
- The number of observations
- The standard error of the proportion (or percentage)
- The number of standard errors needed to encompass the interval chosen (e.g. a 95% CI needs 1.96 SEs)

For a proportion (P), the variance = P (1-P) / n (equation 8)

and SE (P) = $\sqrt{(P\ (1\text{-}P)/n)}$ (equation 9)

95% CI = P $\pm$ 1.96 x SE (P) (equation 10)

For a percentage (P%), the variance = P% (100-P%) / n (equation 11)

and SE (P%) = $\sqrt{(P\%\ (100\text{-}P\%)/n)}$ (equation 12)

95% CI = P% $\pm$ 1.96 x SE (P%) (equation 13)

Example: An audit of emergency admissions at weekends to a surgical unit was undertaken to determine the proportion (percentage) of inappropriate referrals. Surgeons reviewed the notes of patients admitted over the weekend as emergencies in 4 weeks and determined which were 'inappropriate'. Of 93 admissions, 18 were considered inappropriate. The proportion (P) was 18/93 = 0.19, or 19%. But, this is only a snapshot and another audit of 4 weeks may find a different result. We want to estimate the 'true' proportion and calculate a range in which this 'true' proportion is likely to lie. For this we estimate the standard error of the proportion from just one sample based on the 93 observations (n):

SE (P) = $\sqrt{(P\ (1\text{-}P)/n)}$ = $\sqrt{0.19(1-0.19)/93}$ = 0.04

The 95% CI = P $\pm$ 1.96 x SE (P)

the upper CI = 0.19 + (1.96 x 0.04) = 0.27

the lower CI = 0.19 – (1.96 x 0.04) = 0.11

As a percentage, these calculations are:

SE (P%) = $\sqrt{(P\%\ (100\text{-}P\%)/n)}$ = $\sqrt{19(100-19)/93}$ = 4%

The 95% CI = P% $\pm$ 1.96 x SE (P%)

the upper CI = 19 + (1.96 x 4) = 27%

the lower CI = 19 – (1.96 x 4) = 11%

Hence, we are 95% confident that the 'true' proportion of inappropriate admissions lies between 0.11 and 0.27 (11% to 27%, or about 1 in 10 to 1 in 4 admissions).

Were we safe to use the interval of 1.96 taken from the Normal distribution to derive the 95% confidence interval? The rule of thumb specified above requires that both (n x P) and n(1-P) are greater than 5. In this example, (n x P) = 93 x 0.19 = 17.7 and n(1-P) = 93 x 0.81 = 75.3, so the criteria are met.

Note: we used the value of 1.96 taken from the Standard Normal distribution though, with only 93 observations we should have checked the tabulated values from the t-distribution (Appendix B). With a sample size of 93 the interval needed to define 95% of the observations is approximately 2.0 so, in reality, the difference from using the t-distribution instead of the Standard Normal distribution is minimal.

In this example the 95% CI may be considered too wide at 11% to 27%. To increase precision by obtaining a narrower confidence interval we must increase the sample size. Determining how big a sample is needed can be estimated by first deciding the size of the SE we need and, secondly, assuming a likely value for the true proportion (percentage). This again, is the basis of undertaking a power calculation and is dealt with further in the NHS Fife Study Guide 11: 'How to calculate sample size and statistical power'.

Quiz 5:
A survey of body weight was undertaken in 100 randomly selected girls aged 12 – 16 years. The mean weight was 50 kg and the SD was 10 kg.
What information is needed to calculate the 95% confidence interval of the 'true' mean weight of the population from which the sample was drawn?

Now calculate the 95% confidence interval. The results are in Appendix C.

Quiz 6:
A survey of attitudes to physical activity was undertaken in 230 adult men, aged 45 – 60 years randomly selected from a GP practice.
They were asked: "Do you think you do enough exercise to keep fit?"
30% answered Yes, 70% answered No
What information is needed to calculate the 95% confidence interval for those answering 'yes' for the population of men from which the sample was selected?

Now calculate the 95% confidence interval. The results are in Appendix C

(15) Summary

Congratulations of getting this far! The principles of statistics can be difficult to grasp (and to teach). The subject is complex but it is worth the struggle to gain a better understanding of the interpretation of data, either for your own projects or when reading the work of others. Now try the following exercises which include the questions posed in the Introduction (pages 2 and 3). The answers are in Appendix C but please resist the temptation to take a peek first! Hopefully, your answers will be correct but, if not, do not give up, re-read the relevant sections in this guide or seek further advice in one of the many excellent books on the subject. Some books are easier to read than others so find a book you are at ease with and read the equivalent section in it in case my explanations have left you confused (and for which I apologise). Finally, remember, no amount of clever statistics can salvage a badly designed study! But 'study design' is another story.

Quiz 1: True or false?

The Normal Distribution

1) is followed by many variables
2) is also called the Gaussian distribution
3) is followed by all measurements made in healthy people
4) is described by two parameters
5) is skew to the left

The Standard Normal Distribution
6) has mean = 1.0
7) has standard deviation = 0.0
8) has variance = 1.0
9) has the median equal to the mean

Quiz 2: The FEV_1 is a measure of lung capacity. The FEV_1 of a group of women aged 20-25 follows a normal distribution with mean of 3.0 litres, standard deviation 0.4 litres.

Which statements below are true?

1). The distribution of FEV is symmetric about the mean.
2). About 95% of the women have an FEV between 2.2 and 3.8 litres.
3). About 5% of the women have an FEV below 2.2 litres.
4). 50% of the women have an FEV below 3.0 litres.
5). The largest FEV will be 4.6 litres (mean + 4 SD)
6). A woman with an FEV below 2.2 litres is abnormal (unhealthy).
7). If the sample size was doubled the standard deviation would decrease.

(16) Further Reading

Medical Statistics at a Glance. 4th ed. Aviva Petrie & Caroline Sabin, 2019, Blackwell Publishing. *Particularly recommended with its associated Workbook*

Medical Statistics at a Glance Workbook. 1st ed. Aviva Petrie & Caroline Sabin, 2013, Blackwell Publishing.

A-Z of Medical Statistics. Pereira Maxwell F. 1998, Arnold.

Medical Statistics: An A-Z Companion. 2nd ed. Pereira Maxwell F. 2018, CRC Press, Taylor & Francis Group.

An Introduction to Medical Statistics. 4th ed. Martin Bland, 2015, Oxford Medical Publications.

The Art of Statistics. Learning from Data. David Spiegehalter, 2019, Pelican Books.

Essential Medical Statistics. 2nd ed. Betty Kirkwood & Jonathan Sterne, 2003, Blackwell Scientific Publications.

Essential Statistics for Medical Examinations. 2nd ed. Brian Faragher and Chris Marguerie, 2005, PASTEST

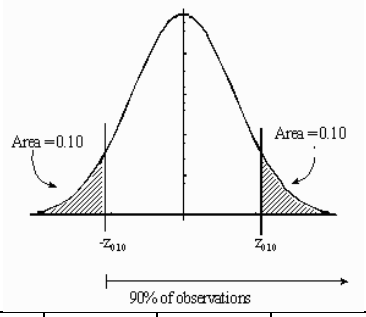
Interpreting Statistical Findings. A guide for health professional and students. Walker J, Almond P. 2010. Open University Press.

Practical Statistics for Medical Research. 2nd ed. Douglas G Altman, 2011, Chapman and Hall.

Statistical Questions in Evidence-Based Medicine. Martin Bland & Janet Peacock, 2000, Oxford Medical Publications.

Appendix A:

Areas in the tail of the Standard Normal distribution



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.432	0.429	0.424
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.344	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.308	0.305	0.301	0.298	0.294	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.254	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.214
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.194	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.168	0.166	0.163	0.161
1.0	0.159	0.156	0.154	0.151	0.149	0.147	0.144	0.142	0.140	0.138
1.1	0.136	0.134	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1.2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1.3	0.097	0.095	0.093	0.092	0.090	0.088	0.087	0.085	0.084	0.082
1.4	0.081	0.079	0.078	0.076	0.074	0.073	0.072	0.071	0.069	0.068
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1.6	0.054	0.054	0.053	0.052	0.050	0.049	0.049	0.048	0.047	0.045
1.7	0.044	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.037	0.037
1.8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.031	0.030	0.029
1.9	0.029	0.028	0.027	0.027	0.026	0.026	**0.025**	0.024	0.024	0.023
2.0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2.1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2.2	0.014	0.014	0.013	0.013	0.012	0.012	0.012	0.012	0.011	0.011
2.3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2.4	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.007	0.006
2.5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2.6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004

Appendix B: The t-distribution - Number of Standard Deviations to define intervals and associated Two-Tailed *P*-values

		Number of Standard Deviations to define interval and *P*-value			
Interval		90%	**95%**	99%	99.9%
***P*-value (two-tail)**		0.10	**0.05**	0.01	0.001
n: 2	**df:** 1	6.314	**12.706**	63.656	636.58
3	2	2.920	**4.303**	9.925	31.600
4	3	2.353	**3.182**	5.841	12.924
5	4	2.132	**2.776**	4.604	8.610
6	5	2.015	**2.571**	4.032	6.869
7	6	1.943	**2.447**	3.707	5.959
8	7	1.895	**2.365**	3.499	5.408
9	8	1.860	**2.306**	3.355	5.041
10	9	1.833	**2.262**	3.250	4.781
11	10	1.812	**2.228**	3.169	4.587
12	11	1.796	**2.201**	3.106	4.437
13	12	1.782	**2.179**	3.055	4.318
14	13	1.771	**2.160**	3.012	4.221
15	14	1.761	**2.145**	2.977	4.140
16	15	1.753	**2.131**	2.947	4.073
17	16	1.746	**2.120**	2.921	4.015
18	17	1.740	**2.110**	2.898	3.965
19	18	1.734	**2.101**	2.878	3.922
20	19	1.729	**2.093**	2.861	3.883
21	20	1.725	**2.086**	2.845	3.850
22	21	1.721	**2.080**	2.831	3.819
23	22	1.717	**2.074**	2.819	3.792
24	23	1.714	**2.069**	2.807	3.768
25	24	1.711	**2.064**	2.797	3.745
26	25	1.708	**2.060**	2.787	3.725
27	26	1.706	**2.056**	2.779	3.707
28	27	1.703	**2.052**	2.771	3.689
29	28	1.701	**2.048**	2.763	3.674
30	29	1.699	**2.045**	2.756	3.660
31	30	1.697	**2.042**	2.750	3.646
41	40	1.684	**2.021**	2.704	3.551
51	50	1.676	**2.009**	2.678	3.496
101	100	1.660	**1.984**	2.626	3.390
201	200	1.653	**1.972**	2.601	3.340
5001	5000	1.645	**1.960**	2.577	3.293

Derived using Microsoft Excel version 5
[source: Medical Statistics at a Glance, Petrie A, Sabin C]

Appendix C: Answers to the quizzes

Quiz 1: True or false?

	The Normal Distribution	True or false?
1)	is followed by many variables	True
2)	is also called the Gaussian distribution	True
3)	is followed by all measurements made in healthy people	False
4)	is described by two parameters	True (mean, μ and SD, σ)
5)	is skew to the left	False
	The Standard Normal Distribution	
6)	has mean = 1.0	False
7)	has standard deviation = 0.0	False
8)	has variance = 1.0	True
9)	has the median equal to the mean	True

Quiz 2: The FEV is a measure of lung capacity. The FEV of a group of women aged 20-25 follows a normal distribution with mean of 3.0 litres, standard deviation 0.4 litres. Which statements below are true?

		True or False?
1)	The distribution of FEV is symmetric about the mean	True
2)	About 95% of the women have an FEV between 2.2 and 3.8 litres.	True
3)	About 5% of the women have an FEV below 2.2 litres.	False (2½%) *
4)	50% of the women have an FEV below 3.0 litres.	True
5)	The largest FEV will be 4.6 litres (mean + 4 SD)	False **
6)	A woman with an FEV below 2.2 litres is abnormal (unhealthy).	False ***
7)	If the sample size was doubled the standard deviation would decrease.	False ****

* See Figure 3, page 12

** There is no reason why the maximum would be mean + 4 SD

*** A woman with a value this low could just be part of the normal (healthy) population as we would expect about 2½% of such women to have values in the tail of the distribution.

**** If the sample size was doubled the standard error (SE) would decrease. SE = SD/$\sqrt{n}$. Check out Table 7, page 31.

Quiz 3: True or False?

1)	Marital status is a categorical variable	True
2)	The mean is a better measure of central tendency when the distribution of data is skewed (i.e. does not conform to a 'bell shape')	False, the median is a better measure of central tendency
3)	The variability of a set of data with a skewed distribution is best described by the standard deviation	False, it's best described by the IQR
4)	The standard deviation of a set of data is derived from the individual data values	True

Quiz 4:

Which distribution has the greater variance and standard deviation?

Distribution 1: mean = 80, variance = 86.6, SD = 9.3

Distribution 2: mean = 80, variance = 573.8, SD = 23.95

Quiz 5:

The information needed: n = 100, mean = 50 kg, SD = 10 kg (and the number of standard deviations to define the 95% limits)

Standard Error of the mean, SE ($\bar{x}$) = SD of sample / $\sqrt{n}$ (*see* equations 5 and 6)

SE=10 / $\sqrt{100}$ = 1 kg,

To calculate the 95% confidence interval we need the number of standard deviations that encompasses 95% of the observations, as defined from the t-distribution (see Appendix B, where n=101). This value is 1.984, or 2 for convenience.

Hence 95% CI= 50 $\pm$ 2 x SE, so the lower limit is 50 – 2 and the upper limit is 50 + 2, and the 95% CI is then cited as 48 – 52 kg

Quiz 6:

The information needed: n=230, % answering 'Yes' = 30% (and the number of standard errors to define the 95% limits)

Standard Error, SE (P%) = $\sqrt{}$ (P% (100 – P%) / n) = $\sqrt{}$ (30 (100 – 30) / 230)
= $\sqrt{}$ (30 x 70 / 230) = 3

Hence 95% CI= 30 $\pm$ 2 x SE, so the lower limit is 30 – 6 and the upper limit is 30 + 6, and the 95% CI is then cited as 24 – 36 %

Were we safe to use the approximate value of 2 taken from the Normal distribution to derive the 95% confidence interval? The rule of thumb requires that both (n x P) and n(1-P) are greater than 5, where P is the proportion. In this example, P=0.3, so
(n x P) = 230 x 0.3 = 69 and n(1-P) = 230 x 0.7 = 161, so the criteria are met.

Glossary Sources: adapted from A-Z of Medical Statistics. Pereira Maxwell, and Medical Statistics at a Glance. 4th ed. Aviva Petrie & Caroline Sabin (*see Further reading*).

Bonferroni correction	A procedure for adjusting the P-value in a statistical analysis involving multiple significance testing. When testing, for example, 20 different measures between two groups it is likely that at least one measure will differ statistically at the 5% level by chance alone and may not represent a true difference between those groups.
Chi-squared test	A significance test for comparing two or more proportions from independent groups. The observed proportion in each group is compared with the expected proportion based on a null hypothesis.
Confidence interval, CI	A range of values in which the true mean for a population is likely to lie. It usually has a proportion assigned to it (for example 95%) to give it an element of precision.
Continuous variable	A numerical variable which can theoretically take any value within a given range (for example, height, weight, blood pressure).
Correlation coefficient (Pearson's)	A measure of the linear association (a straight line in a scatter plot) between quantitative or ordinal variables.
Data cleaning	The process of trying to find errors in the data set.
Database	A systematised collection of data that can be accessed and manipulated by a stats package such as SPSS.
Degrees of freedom	A concept used with statistical tests that refers to the number of sample values that are free to vary. In a sample, all but one value is free to vary, and the degrees of freedom is then N-1. For example, consider a set of four values with the mean of 5 and a sum of 20. If you are asked to 'invent' the individual four values then you are only 'free' to invent three of them as the fourth must ensure the sum adds to 20 (note, it can be a negative number).
Effect size	A standardised estimate of the treatment effect calculated by dividing the estimated difference between two groups by the standard deviation of the measurements (means or proportions). In the context of power calculations the effect size is the same as the standardised difference (*see below*).
Frequency distribution	A display of data values from the lowest to the highest, along with a count of the number of times each value occurred.
Heteroscedasticity	Unequal variances between two or more subgroups

Histogram	A graphic display of data frequency using rectangular bars with heights equal to the frequency count.
Homoscedasticity	Equality of variances within two or more subgroups
Hypothesis	A statement of the relationship between 2 or more study variables. *See Null Hypothesis*
Logarithm	The logarithm of a number is the exponent (power) to which another fixed value, the base, must be raised to produce that number. For example, the logarithm of 1000 to base 10 is 3 because 10 to the power of 3 (10^3 = 10 x 10 x 10) is 1000
Margin of error	A term used by pollsters to estimate the error from a survey of opinions. In this account it is a range of values equivalent to twice the standard error on either side of the estimated population mean. It is equivalent to the 95% confidence interval.
Mean	The average value or measure of central tendency. The mean is obtained by dividing the sum of values by the total number of values.
Median	Middle value when data are ordered. The value that splits the sample in two equal sized parts.
Mode	The value that occurs most frequently.
Non-parametric	Refers to data and tests of significance which makes no assumptions about the distribution of the data. Data that are skewed in distribution (to the right or left) are described as non-parametric.
Normal (Gaussian) distribution	A continuous probability distribution that is bell-shaped and symmetrical; its parameters are the mean and variance.
Null Hypothesis, H_O	The statement that assumes there is no difference between two populations being compared, or no relationship or association between two variables in a population. An experiment may be undertaken to see if H_O can be rejected in favour of an alternative hypothesis, H_A.
Outlier	Values in a set of observations which are much higher, or lower, than the 'average' and lie well away from the rest of the data (in the tail of the distribution).
Parameter	A measurable characteristic of a population (e.g. average and standard deviation of blood pressure for a group of individuals).
Parametric	Refers to data in which the distribution is bell-shaped (Normal or Gaussian). Statistical tests that rely on data being distributed this way are called parametric tests.
Power	The probability of rejecting the null hypothesis when it is false.

Power calculation	Refers to a way of calculating the number of subjects needed for the results of a study to be considered statistically significant.
Protocol	A full written description of all aspects of a study – the 'recipe'.
P-value	*See Significance Level*
Ratio	A quantitative variable that has a true zero. An example is body weight.
Regression coefficient	The slope of the line of best fit in a plot between two variables. It represents the increase in an outcome variable from a unit increase in the predictor variable. For example, in a plot of total lung capacity against height in women the regression coefficient is 6.60 litres/metre which means that for every increase in one metre in height the lung capacity increases by 6.60 litres.
Significance level (P-Value)	In the context of significance tests, the P-value represents the probability that a given difference (or a difference more extreme) is observed in a study sample (between means, proportions etc) when in reality such a difference does <u>not</u> exist in the population from which the sample was drawn. In effect it's the probability of getting a wrong answer by deciding that two populations differ in some way when in fact they do not. In statistical parlance, it is the probability of rejecting a null hypothesis of no difference between two populations when in fact the null hypothesis is true.
Spreadsheet	A computer program (e.g. Excel) that allows easy entry and manipulation of figures, equations and text. It displays multiple cells that together make up a grid consisting of rows and columns, each cell containing either text or numeric values or a formula that defines how the contents of that cell is to be calculated. Spreadsheets are frequently used for financial information because of their ability to re-calculate the entire sheet automatically after a change to a single cell is made.
Standard deviation, SD	A measure of variability of data. The standard deviation is the average of the deviation of individual values from the mean measured in the same units as the mean.
Standard error (of the mean), SE	A measure of precision of the sample mean. The difference between the true mean of a population and the estimated mean taken from a sample is referred to as sampling error. Estimates of a population *mean* value will vary from sample to sample. The distribution of mean values derived from multiple samples is called the sampling distribution. The SE is the 'standard deviation' of this distribution.
Standard score (also, z-score)	Refers to how many standard deviations away from the mean a particular score is located.

Standardised difference	A ratio equal to what is considered the clinically important treatment difference divided by the standard deviation of the measure in question.
T-test	A statistical test used to determine if the means of 2 groups are significantly different.
Type I error (alpha error)	The probability of making the wrong choice by rejecting a null hypothesis when it is true. In other words, a type I error occurs when it is concluded that a difference between groups is not due to chance when in fact it is (reject a true null hypothesis).Also relates to the significance level (P-value).
Type II error (beta error)	The probability of making the wrong choice by accepting a null hypothesis when it is false. In other words, a type II error occurs when it is concluded that differences between groups were due to chance when in fact they were due to the effects of the independent variable (accept a false null hypothesis).This probability becomes smaller with increasing sample size.
Variable	Any quantity that varies (e.g. blood pressure).
Variance	A measure of variability of data equal to the square of the standard deviation.
Z-score	A standard score, expressed in terms of standard deviations from the mean.