University of Stirling
Computing Science and Mathematics
School of Natural Sciences

MATU9D2: PRACTICAL STATISTICS

Chapter 1 Introduction

1.1 General Background

Two important branches of applied mathematics are PROBABILITY THEORY and STATISTICS.

There has been a rapid expansion in both the theory and application of both these areas in the last century and in particular in the years since the advent of powerful computers.

Definition : STATISTICS

A collection of data originally about the state of the nation e.g. size of the population, levels of trade or unemployment. The dictionary definition is a (large) collection of numerical facts or figures. The alternative definition is 'the Science concerned with the collection, classification and interpretation of data'.

Definition : STATISTICAL INFERENCE

In many cases, we will be looking at problems where we wish to draw general conclusions on the basis of a limited amount of data.

Because they are based on a limited amount of data, the conclusions will be subject to uncertainty. Inference is the branch of statistics which attempts to quantify the uncertainty using probability and related measures.

1.1.1 Study Design

Statisticians are also concerned with the design of appropriate methods of data collection. The design of the experiment is of crucial importance if the analysis of the data is going to yield the greatest amount of correct and accurate information.

In this course we shall make a further distinction between two different types of problem which involve using two further different elements of statistics:

A Descriptive Statistics

B Statistical Inference

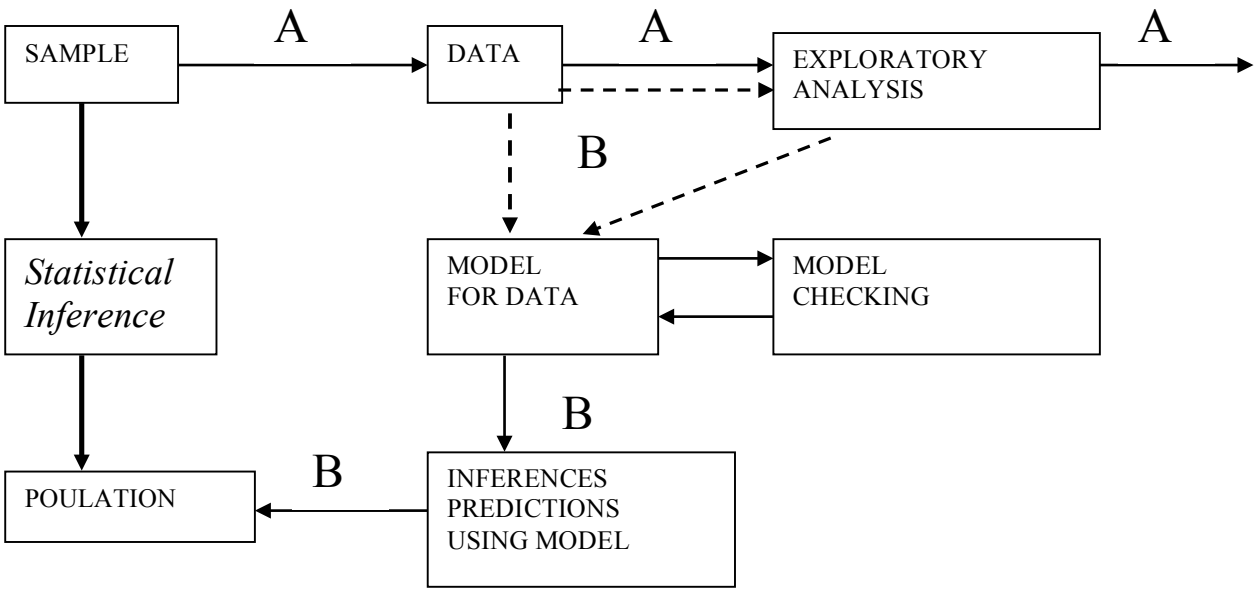
In many cases we will want to use statistical inference to extrapolate conclusions from a sample and apply them to a wider population. For example, the use of opinion polls or a survey of television viewing would usually be conducted only in a subset of the population.

This is a very common type of study i.e. it is usually not feasible to examine the entire 'population of interest' so information is collected on a 'representative sample' and we use the results from the sample to infer the results for the population.

You will hopefully see that one very important aspect of 'statistics' is the choice of the sample. We shall return to this later.

However, if we do not wish to extrapolate our conclusions to a wider population and only want to describe the data already collected, this would technically be termed using 'descriptive statistics'.

Therefore, we can clearly define two different methodologies that we will follow in practice:



Route to follow for given methodology

The appropriate model may come from theory but we should always begin analysis with Exploratory Data Analysis

We will therefore clearly identify the two streams A (Descriptive Statistics) and B (Statistical Inference).

The science of **STATISTICS** is applied in a vast array of other disciplines and a wide variety of contexts e.g. Social Sciences, Management Sciences, Biological and Medical Sciences.

1.2 Data

1. Observations or data are the raw materials with which statisticians work.

2. For most statistics to be applicable these observations must be in the form of numbers or be able to be converted into numbers.

3. We now classify the different types of data and this is a most important step in any analysis. Usually the type of analysis to be performed depends on the type of data being considered.

4. All formal techniques have assumptions of various kinds to do with the type and structure of the data. Hence in order to use a technique we must check that the assumptions are 'reasonably' valid.

1.2.1 Data Types

We say that our data consists of observations which are values of variables.

Examples of variables include

• Number of live births outside marriage	:	0, 1, 2, 3,.....
• Geographical region	:	North Scotland, West Scotland etc.
• Weekly earnings	:	From £80 upwards
• Dose of chemical	:	0 upwards
• Smoking Behaviour	:	'Non smoker', 'Ex smoker'; Occasional smoker'; ' Heavy Smoker'
• Gender	:	Male or Female

Are there any obvious differences between the variables listed above?

There are two main types of data:

CATEGORICAL and QUANTITATIVE

TYPE 1. Examples of **CATEGORICAL VARIABLES** are:

Gender, Smoking Behaviour, Geographical Region, Hair Colour, Eye Colour

However, there are two sub-types of Categorical Variable:

NOMINAL and ORDINAL

With an **ORDINAL** variable the categories the variable can take can be ordered.

For example,

Smoking : Non- Smoker, Ex-Smoker, Occasional Smoker, Heavy Smoker

Experience of Statistical Packages : None, A Little, Some, A lot

Strength of agreement with a statement : Disagree Strongly
Disagree Mildly
Neutral
Agree Mildly
Agree Strongly

However, with **NOMINAL** variables, the categories are simply labels i.e. one category cannot be ranked as 'greater than' or ' less than' other categories. For example, 'male' and 'female' are different categories but we cannot say that greater than or better than the other; with eye colour we cannot say that 'green eyes' are 'greater than 'brown eyes'. The categories are often conveniently designated by numbers but the numbers do not imply any relationship between individual numbers.

TYPE 2. Examples of **QUANTITATIVE VARIABLES** are:

Number of live births, weekly earnings, height, weight, blood pressure, expenditure

In this case we attempt to quantify something by counting or measuring.

N.B. There is a rather subtle distinction between two types of such variables into INTERVAL (scale of equal intervals, addition and subtraction possible) and RATIO (in addition, the

scale starts at true zero) variables. However, we shall ignore this distinction.

The two major sub-types of **QUANTITATIVE** data are

DISCRETE and CONTINUOUS.

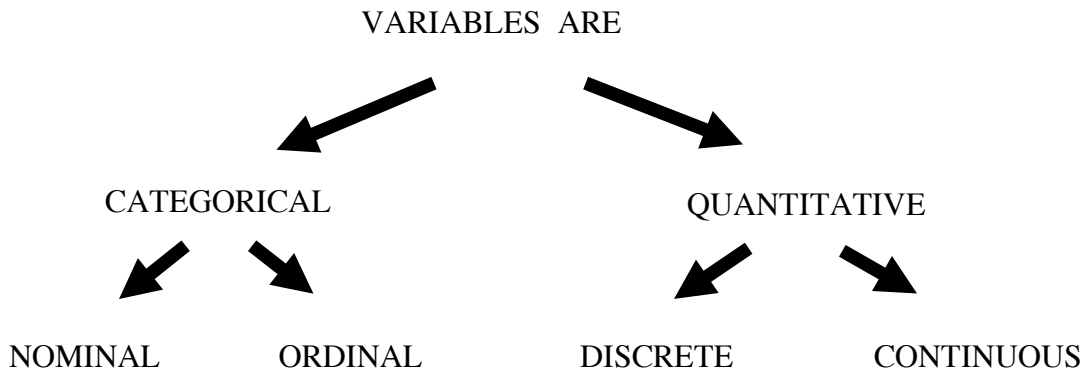
Discrete variables are any kind of count, they take values in a restricted set such as integers { 0, 1, 2... }.

For example, number of defective items.

Continuous variables are any kind of measurement, they may possibly take any real number.

For example, height, weight and age.

In summary:



In this module (as with many others) you will be using data that has been collected by other people so……

When confronted with data we initially ask many questions such as:

Why were these data collected?

Who collected these data?

What methods did they use to collect the data?

Are there any sources of bias?

Has randomisation been used?

What is the target population for which inference is desired?

Is an informal / descriptive / exploratory analysis sufficient?

Do we have all the relevant data?

Has some been concealed from us?

Has the investigator who collected the data removed any that "didn't quite fit in"?

1.3 Looking at the Data

People sometimes say 'let the data speak for themselves'. What do the following 'say to you'?

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6
1	1	1	2	87	95
2	1	1	2	85	90
3	1	1	2	84	88
4	1	1	2	31	90
5	1	1	2	83	91
6	1	1	2	82	96
7	1	1	2	79	96
8	1	1	2	76	96
9	1	1	2	74	86
10	1	1	2	60	74
11	1	1	2	64	86
12	1	1	2	23	79
13	1	1	2	35	80
14	1	1	2	17	75
15	1	1	2	16	57
16	1	1	2	36	80
17	1	1	2	39	85
18	1	1	2	38	56
19	1	1	2	45	74
20	1	1	2	17	91
21	1	1	2	41	77
22	1	1	2	52	90
23	1	1	2	50	90
24	1	1	2	35	69
25	1	1	2	16	72
26	1	1	2	35	65
27	2	2	2	39	64
28	2	2	2	33	88
29	2	2	2	55	77
30	2	2	2	43	92
31	2	2	2	29	58
32	2	2	2	66	96
33	2	2	2	80	100
34	2	2	2	72	100
35	2	2	2	88	100
36	2	2	2	37	63
37	2	2	2	66	95

EXAMPLE

This is part of a set of data looking at different teaching methods where :

Col 1 : Index Number

Col 2 : Group 1 = Boys/Comp

: 2 = Girls/Comp

: 3 = Girls/Trad

: 4 = Boys/Trad

Col 3 : Gender 1=Boys

2=Girls

Col 4 : Method

1=Traditional

2=Computerised

Col 5 : Pre Score

Col 6 : Post Score

We will look at the full set of data from this study when we look at plots for quantitative data.

A useful way of gaining insight about the data is by creating pictures and numerical summaries which help to pick out the general features and also the unusual aspects of the data

1.3.1 CATEGORICAL DATA

For categorical data, useful pictures are pie-charts and bargraphs. We may use the pie-chart and bargraph to create a pictorial representation of categorical data. Let us illustrate using an example.

EXAMPLE In a survey, a random sample of electors were asked their political preferences and social class. The results were as follows:

		Political Preference			
		Tory	Labour	Lib-Dem	Other
Social Class	A	110	11	37	2
	B	420	99	237	8
	C1	321	119	143	12
	C2	353	363	205	9
	D	183	286	134	6

In social class A, there are 160 people in the sample. In percentage terms

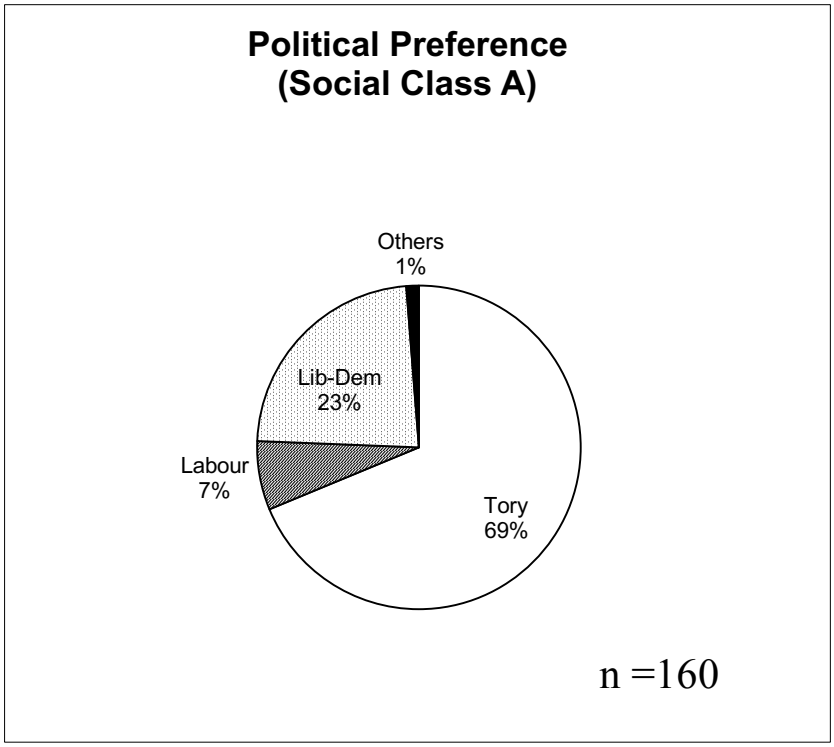
110/160 voted for the Conservatives
11/160 voted for Labour
37/160 voted for the Liberal Democrats
2/160 voted for other parties.

The percentages are 69%, 7%, 23% and 1% respectively.

1.3.2 Pie-Chart

There are 4 categories (i.e. voting groups) so we split the circle into four sectors, with each sector having an area proportional to the %vote for the corresponding party.

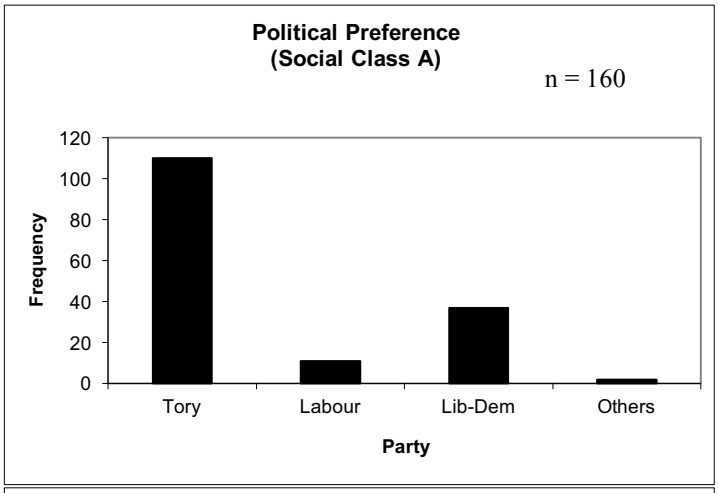
The resulting pie-chart appears below. It is easily constructed with a protractor (or even better using a computer). One percentage point is equivalent to an angle of 3.6 degrees. Hence to find the appropriate size of each sector we multiply the percentage by 3.6 to obtain the size of angle of the sector.

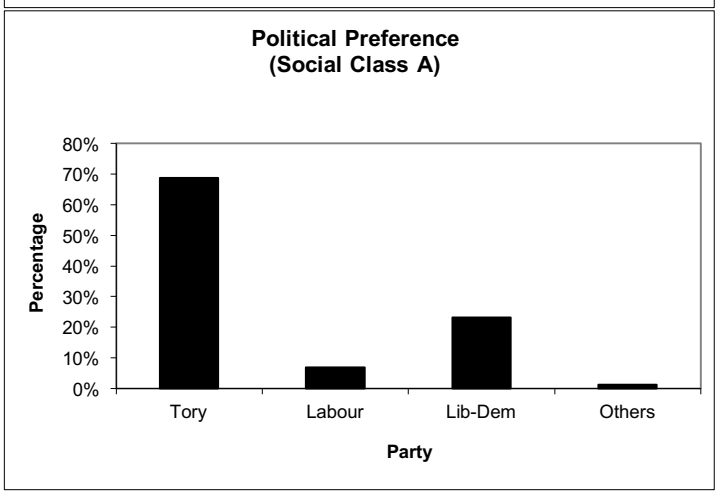


1.3.3 Bargraph

We may also represent the distribution of a categorical variable by using a bargraph. Hence we draw a bar for each category (all of the same thickness) with the height of the bar representing the frequency (i.e. count) or percentage of the sample that belong to each voting category.

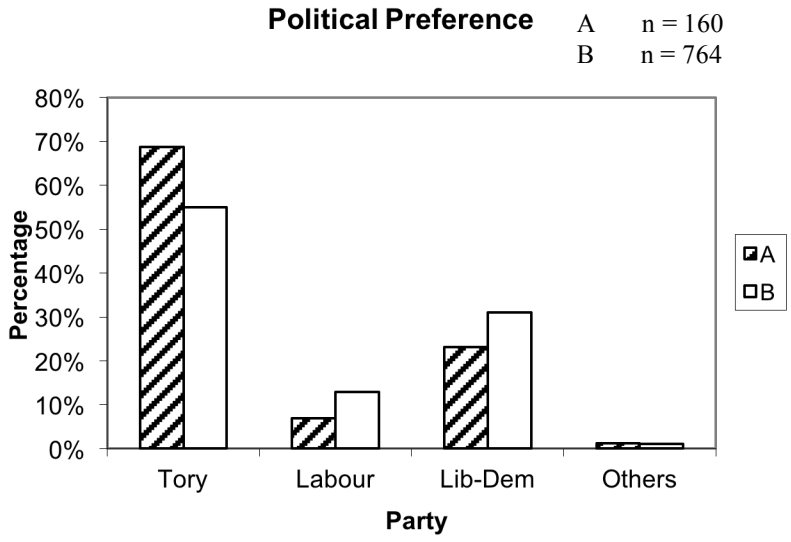
Note that it is essential to provide a suitable scale.





Similar bargraphs could be drawn for the other social class groupings. If we wished to compare the voting pattern of two or three of the social groupings, we could use a 'multiple' bargraph (side-by-side bargraph) as follows:

From such a display it is easy to compare the support obtained from classes A and B for each political party separately. The percentages are:



		Political Preference			
		Tory	Labour	Lib-Dem	Others
Social	A	69%	7%	23%	1%
Class	B	55%	13%	31%	1%

Alternatively, we could plot a side-by-side bargraph of the observed frequencies as follows:

So, we have seen how to use pie-charts and bargraphs (and percentages) to describe a categorical variable.

Let us now consider some simple pictures which are appropriate for describing the distribution of a single quantitative variable and also for making a comparison of several such distributions.

Does this second side-by-side bargraph provide a different 'picture' of the data?

1.3.5 QUANTITATIVE DATA

We consider, firstly, a simple approach called a **stem and leaf plot** or stemplot.

EXAMPLE Consider the following data about the ozone concentrations between 1985 and 1989 at various sites in the United Kingdom.

Consider the 1989 values of average annual ozone concentration (in parts per billion). They were collected at different monitoring stations throughout the U.K. Hence looking at the distribution of all the values will give an indication of the spatial variation of average annual ozone concentration across different parts of the U.K. We may also determine the typical or 'average' concentration and in addition discover areas where the reading is particularly high

The 1989 data are:

10, 13, 25, 31, 31, 32, 26, 25, 32, 26, 21, 17, 27, 28, 31, 26, 17, 34

The data have two digits: a 'tens' digit and a 'units' digit
e.g. 32 = 3 * 10 + 2

so 3 is the 10's digit and 2 is the units digit. We make the **10's** digit the **stem** digit and the **units** digit the **leaf** digit and obtain the following display:

WORKING VERSION

1	
2	
3	

It is useful to order the digits on each stem. (This is actually a quick method for ordering data that we shall need soon). This gives an **ordered** stemplot.

FINAL VERSION

Stem Unit : 10 Leaf Unit : 1

1	0	3	7	7				
2	1	5	5	5	6	6	7	8
3	1	1	1	2	2	4		

N.B. It would be useful here to plot the data on a map of the UK; it may be that there are some spatial aspects of the data that are worth considering.

We now consider data on different education systems part of which was shown on Page 7.

Now consider separate stemplots of prescore for boys and girls.

Boys Stem Unit : 10 Leaf Unit : 1 n = 57

1	5	5	6	6	6	7	7	7							
2	1	2	3	5	5	6	7	9							
3	0	0	1	3	3	4	4	4	5	5	5	6	6	8	9
4	1	3	3	5	6	6	6								
5	0	0	2	4	7	9									
6	0	4													
7	4	6	9	9											
8	1	2	3	4	5	7									
9	3														

There appear to be two peaks on stem 3 and stem 8 which suggests that the data tends to cluster into two groups - a 'low' group and a 'high' group. The low group is much larger. The scores range from 15 to 93 - a wide spread of values.

Choosing An Appropriate Scale for a Stem and Leaf Plot

In practice, if drawing a stemplot by hand, it is usually a process of trial and error to obtain a reasonable number of stems

- too few and we lose the shape of the data
- too many and we cannot obtain the overall shape of the distribution.

A rule of thumb is : use $\sqrt{n}$ stems for n observations

When there are a lot of observations on each stem, we may split each stem into more bits. For example, in this example split each stem into two halves and record leaf digits 0-4 on one of them and the digits 5-9 on the other.

We obtain Stem Unit: 10 Leaf Unit : 1

1.	5	5	6	6	6	/	/	/
2*	1	2	3					
2.	5	5	6	7	9			
3*	0	0	1	3	3	4	4	4
3.	5	5	5	6	8	9	9	
4*	1	3	3					
4.	5	6	6	6				
5*	0	0	2	4				
5.	7	9						
6*	0	4						
6.								
7*	4							
7.	6	9	9					
8*	1	2	3	4				
8.	5	7						
9*	3							

Now the two clusters are more apparent with peaks at 93-84 and 33-34 (as well as 15-17). The lower cluster appears to be asymmetric with the scores skewed towards lower values. The overall middle of scores lies the range 35-39.

It is natural to compare these scores with the corresponding pre-scores for the girls. We do this by means of a **'back-to-back' stemplot.**

N.B. This only works for a comparison between 2 stemplots - but with more than two we can draw them all on a common scale to attain the same effect and aid comparison

Stem Unit: 10 Leaf Unit: 1

Girls n = 67										**Boys n = 57**							
								4	1.	5	5	6	6	6	7	7	7
							9	7	2*	1	2	3					
						4	3	2	2.	5	5	6	7	9			
				9	9	8	7	7	3*	0	0	1	3	3	4	4	4
			3	3	3	2	2	1	3.	5	5	5	6	8	9	9	
			9	8	8	8	7	5	4*	1	3	3					
							2	1	4.	5	6	6	6				
							6	5	5*	0	0	2	4				
					4	1	1	1	5.	7	9						
9	9	8	8	7	7	7	6	6	6	6*	0	4					
		4	3	2	2	2	1	1	1	6.							
					9	8	6	6	5	7*	4						
		4	3	3	2	0	0	0	0	7.	6	9	9				
						8	7	6	5	8*	1	2	3	4			
										8.	5	7					

1 | 9* | 3

The distribution of girls' pre-scores also clusters with two groups, with more in the 'higher' group. Note also that the lower group of girls is more symmetrical, peaking at 42-43, whereas the boys' scores were skewed towards lower values.

In general it seems as though the bulk of the girls' marks are greater than those of the boys.

Describing The 'Shape' of The Plot (or Distribution)

A stem and leaf plot is useful for gaining information about the shape of the distribution :

1. The spread of the distribution
2. The number of peaks
3. The symmetry of the distribution - symmetric or asymmetric
4. Any gaps
5. Any unusual values - **outliers** - that don't fit in with the bulk of the distribution

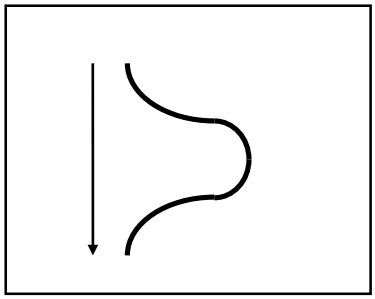
When is the Stem & Leaf Plot Symmetric?

If there is more than one peak the distribution can only be either symmetric or not symmetric. However, if the plot is **unimodal** (i.e. single peaked) then we have the following additional definitions:

Increasing values top to bottom in all the figures below

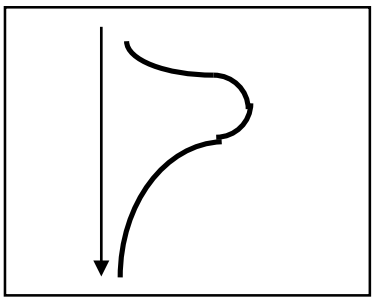
If the Stem & Leaf Plot is of this shape ⇒

then we say that it is symmetric



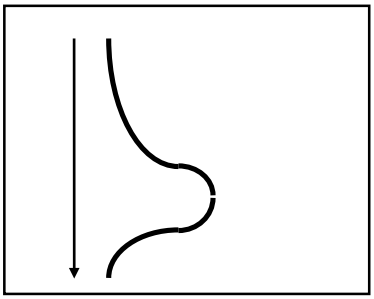
If the Stem & Leaf Plot is of this shape ⇒

then we say that it is positively skewed



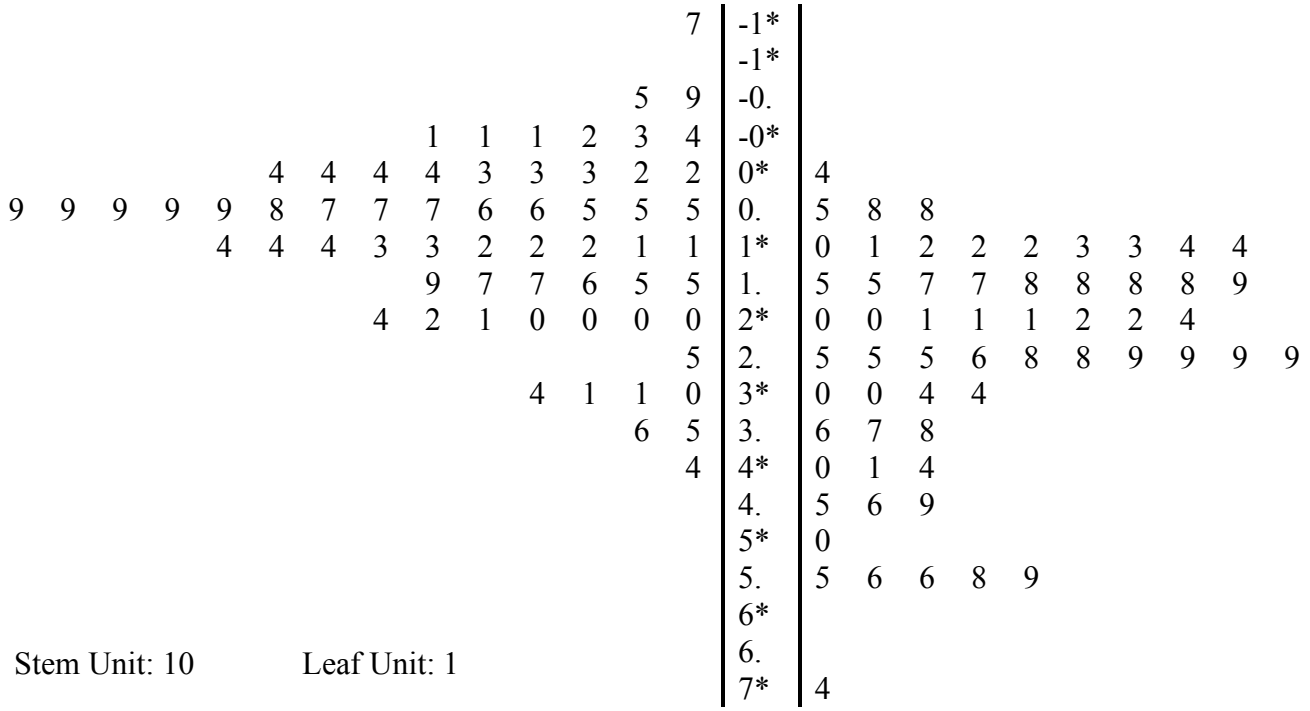
If the Stem & Leaf Plot is of this shape ⇒

then we say that it is negatively skewed



The main purpose of the educational investigation was to compare the improvement obtained with the two teaching methods. We take the improvement score to be Post-Score - Pre-Score.

Traditional **Computerised**



Generally, it appears that the values for the computerised method tend to be larger. While such plots give important information about level, spread, shape and outliers, it is useful to calculate numerical measures for these aspects of the distribution.

1.3.6 The Five Number Summary

First of all we **order** the data values from smallest to largest and denote them by

$x_{(1)}, x_{(2)}, \ldots\ldots\ldots\ldots x_{(n)}$ [This is called the order statistic]

where the number in the brackets tells us the position in the of the observation within the sample.

E.g. $x_{(1)}$ = minimum value in sample
$x_{(n)}$ = maximum value in sample

The **Five Number Summary** gives us a 'Numerical Summary' of our data. The Five Numbers involved are

1.	**Minimum**	-	The smallest number in our set of data.
2.	**Maximum**	-	The largest number in our set of data.
3.	**Median**	-	The 'middle' number of our set of data i.e. half the data is less than this value and half above.
4.	**Lower Quartile (Q_1)**	-	One quarter of the data is less than this value and three quarters greater.
5.	**Upper Quartile (Q_3)**	-	One quarter of the data is greater than this value and three quarters less.

(A) The minimum and maximum are straight-forward to find!

(B) The middle value in the sample is called the **Sample Median**. It is the value which splits the data into two equal halves.

Example 1	Consider the data	2	7	12	13	15	17	21
	Then the sample median	=	13					
Example 2	Consider the data	2	7	12	13	15	17	

Now there is not a unique middle value and we normally take the sample median to be the mid-point between the two middle values. i.e. $(12 + 13) / 2 = 12.5$

So if the sample size is odd, we take the sample median to be the middle value. If the sample size is even, we take the sample median to be the 'average' of the two middle values.

We can think of the calculations in 2 parts:

(i) Find the **POSITION** of the **Median**

The Median is $\left(\frac{n+1}{2}\right)th$ smallest observation i.e. in Position $\left(\frac{n+1}{2}\right)$

E.g. Example 1 n = 7 so Median is 4th smallest
Example 2 n = 6 so Median is 3.5th smallest

(ii) Find the **VALUE** of the **Median**

Find the value at the position given in Step (i)

e.g. Example 1 4th smallest =

Example 2 3.5th smallest =

(C) So we have 3 numbers, the data extremes and the middle. To complete the 5 number summary we consider the quartiles.

The upper quartile (Q3) is such that one-quarter of the distribution lies above it and three-quarters below. The lower quartile (Q1) is such that one-quarter of the distribution lies below it and three-quarters above.

(i) Find the **POSITION** of the **Upper Quartile**

The Upper Quartile is $\frac{3\times(n+1)}{4}th$ smallest observation i.e. in Position $\frac{3\times(n+1)}{4}th$

E.g. Example 1 n = 7 so Upper Quartile is 6th smallest
Example 2 n = 6 so Upper Quartile is 5.25th smallest

(ii) Find the **VALUE** of the **Upper Quartile**

Find the value at the position given in Step (i)

e.g. Example 1 6th smallest =
So the Upper Quartile is

Example 2 5.25th smallest =
So the Upper Quartile is

(i) Find the **POSITION** of the **Lower Quartile**

The Lower Quartile is $\left(\frac{n+1}{4}\right)th$ smallest observation i.e. in Position $\left(\frac{n+1}{4}\right)$

E.g. Example 1 n = 7 so Lower Quartile is 2nd smallest
Example 2 n = 6 so Lower Quartile is 1.75th smallest

(ii) Find the **VALUE** of the **Lower Quartile**

Find the value at the position given in Step (i)

e.g. Example 1 2nd smallest =
So the Lower Quartile is

Example 2 1.75th smallest =
So the Lower Quartile is

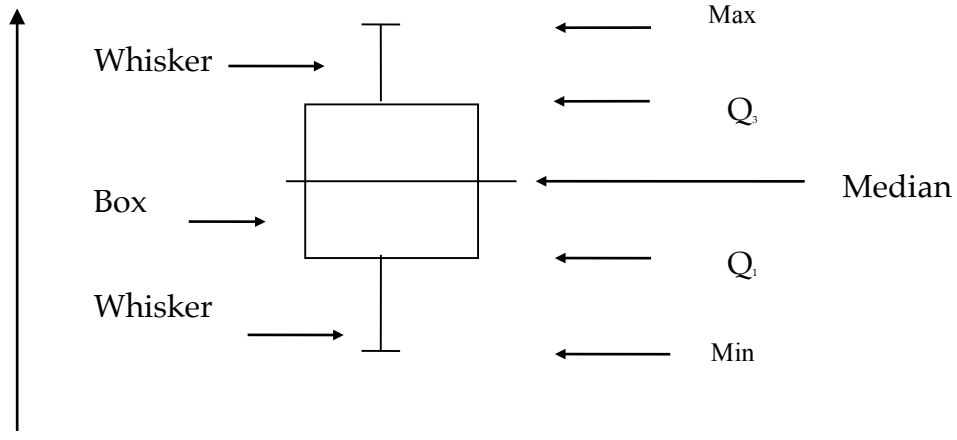
The Five Number Summary is usually written as

n

Median	
Q_1	Q_3
Minimum	Maximum

It is possible to turn the Five Number Summary into a picture by means of a **BOX AND WHISKER PLOT**. (Boxplot, for short!!!)

We mark off the five numbers on a vertical (or horizontal) scale - long lines for the median and quartiles and short lines for Min and Max. A box is then constructed about the first three lines and whiskers between Min and Q_1 and between Max and Q_3.



This plot is useful for assessing the symmetry of the data distribution and particularly for identifying outliers.

EXAMPLE Consider the data

(a)	7	10	12	15	17	19	23	27
(b)	17	27	32	41	53	62	7	

1.3.7 Application to the Educational Data

Parts of the data set were presented in raw form on Pages 7 with complete information given via the stem and leaf plots on subsequent pages.

Let us consider the Improvement Scores (i.e. Post-Score - Pre-Score). The stem and leaf plot for this data is shown below and was shown and discussed on Page 17.

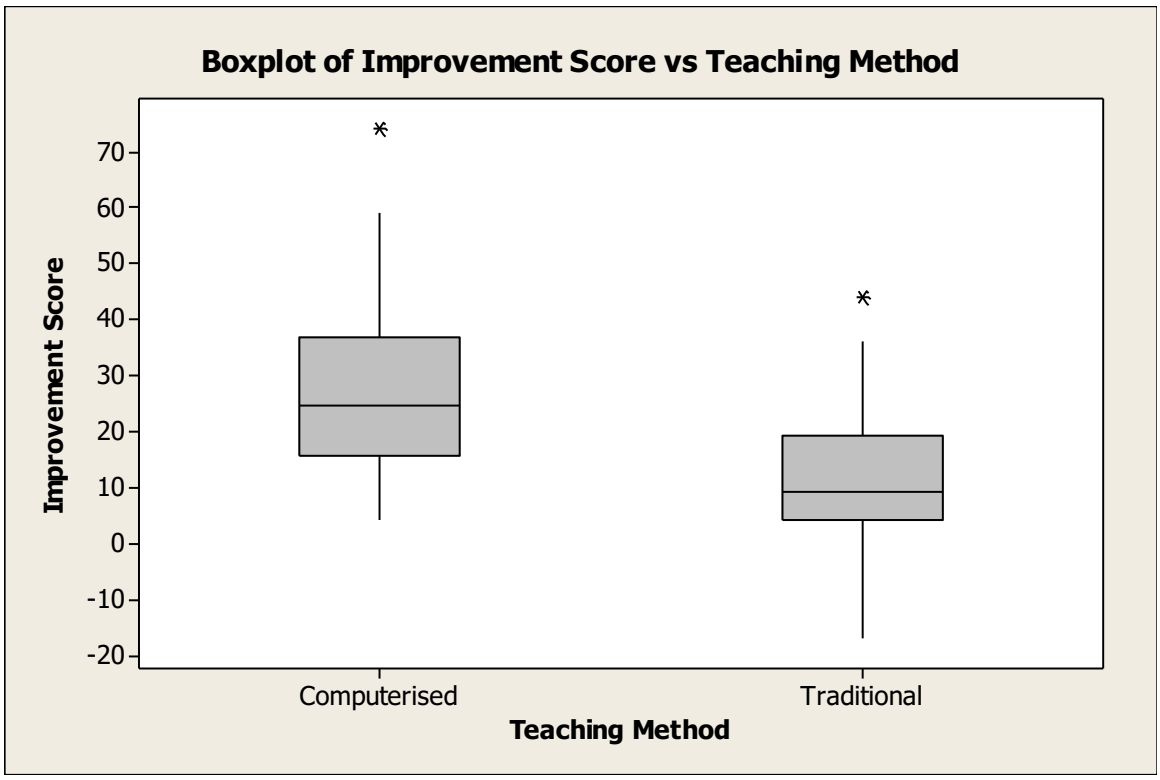
Stem Unit: 10 Leaf Unit: 1

Traditional														**Computerised**										
												7	-1*											
													-1*											
											5	9	-0.											
								1	1	1	2	3	4	-0*										
					4	4	4	4	3	3	3	2	2	0*	4									
9	9	9	9	9	8	7	7	7	6	6	5	5	5	0.	5	8	8							
				4	4	4	3	3	2	2	2	1	1	1*	0	1	2	2	2	3	3	4	4	
								9	7	7	6	5	5	1.	5	5	7	7	8	8	8	8	9	
							4	2	1	0	0	0	0	2*	0	0	1	1	1	2	2	4		
													5	2.	5	5	5	6	8	8	9	9	9	9
										4	1	1	0	3*	0	0	4	4						
												6	5	3.	6	7	8							
													4	4*	0	1	4							
														4.	5	6	9							
														5*	0									
														5.	5	6	6	8	9					
														6*										
														6.										
														7*	4									

In particular, we will focus on the comparison of the improvement scores using Traditional and Computerised systems of teaching.

Solutions

1.3.8 Box - Plots (Box and Whisker Plots)



Comparing the boxplots, we see that the distribution of the improvement scores obtained using the computerised method is shifted up, and so tend in general (on average) to be larger, compared to the 'traditional' improvement scores.

Thus it seems that (on average) the computerised method has been more effective e.g.

	Md (Computerised)	=	24.5
compared to	Md (Traditional)	=	9

The spreads of the two distributions are comparable, while the computerised distribution is more asymmetric due to the highest value (74) being rather "disconnected" from the main body of the data.

Note that $Q_3 - Q_1$ is called the Interquartile Range

and $\frac{Q_3 - Q_1}{2}$ is called the Quartile Deviation (or the Semi-Interquartile Range)

OUTLIER DETECTION

Rules : If a data value x is such that

$x < Q_1 - 1.5(Q_3 - Q_1)$

or $x > Q_3 + 1.5(Q_3 - Q_1)$

we term x a **mild outlier**

If $x < Q_1 - 3(Q_3 - Q_1)$

or $x > Q_3 + 3(Q_3 - Q_1)$

we term x an **extreme outlier**

Often if we have possible outlier, we mark them on the boxplot.

Possible outlier should be given extra attention i.e. they may be mistakes in the data collection.

Note, however, that if the underlying distribution is asymmetric then we expect there to be some outlying values.

Statistical procedures can be sensitive to the effects of such possible outliers. When they are present, we can try to transform the data or if that does not work then we can analyse the data both including and excluding the outliers in/from the sample. If our conclusions are different then a more **sophisticated analysis** will be necessary and if not, then our conclusions are not sensitive to the presence of the possible outliers and are then termed **robust.**

Example Consider another set of data:

Soil pH, a measure of the extent to which soil is acidic or basic, is one characteristic that plays an important role in the suitability of soil to support vegetation at mine reclamation sites. The following data were recorded for 26 mine soil specimens.

There are 26 observations and the 5 number summary is

$Q_3 - Q_1 = 1.06$

$1.5\,(Q_3 - Q_1) = 1.59$

$3\,(Q_3 - Q_1) = 3.18$

26

4.315	
3.59	4.65
2.62	6.79

The stem and leaf plot looks like

Stem Unit = 1
Leaf Unit = 0.01

2.	62	83	91				
3*	49						
3.	58	58	59	84	86	90	
4*	11	25	27	36	41	43	46
4.	53	58	65	75	78		
5*	21						
5.							
6*	00	49					
6.	79						

Mild outliers

Extreme outliers

Now the real question is, "What do we do with these values?" We should check them to see if any mistake was made in the readings or perhaps a different person recorded those data points. If not, then the data do appear to be asymmetric, so perhaps it is not surprising that there are some possible outliers in the data.

We now consider another useful data-picture for quantitative data - the **HISTOGRAM**.

A histogram is a very useful picture giving the shape, spread and level of a distribution and is more useful than the stem plot with large data sets.

We now construct histograms.

The basic idea is

1. Define a set of class intervals in such a way that the class intervals cover the set of possible data values and that each data value belongs to one and only one interval.

2. Allocate each observation to a class interval and count up the frequencies for each interval.

3. If the class intervals have equal width then plot a bar to represent the frequency or relative frequency of each interval; if the widths are unequal, then

   Height of bar = relative frequency / width of class interval

   Note : Relative Frequency = Class Frequency / Total Frequency

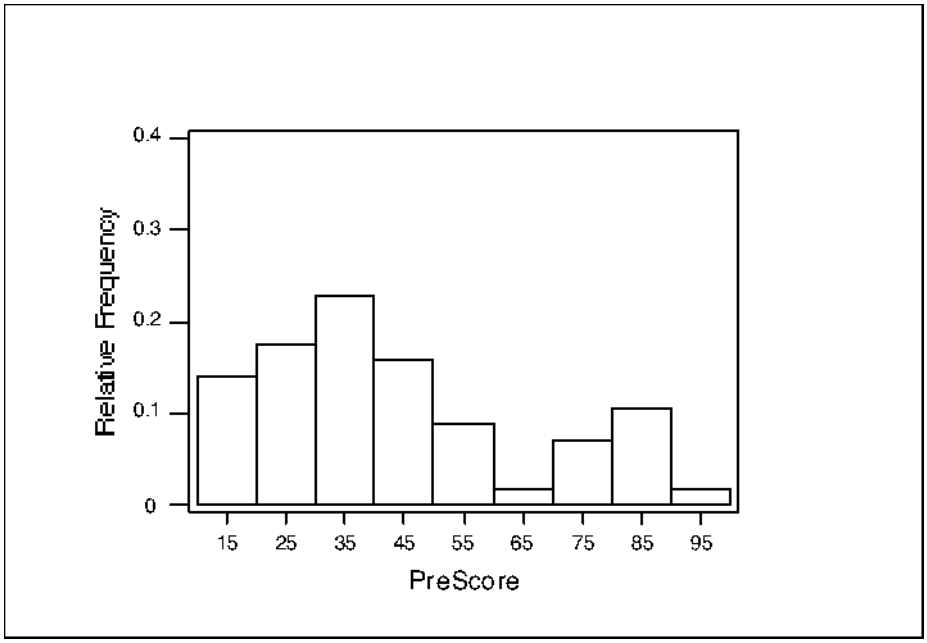
N.B. It is important to define the boundaries between adjacent class intervals in such a way that an observation cannot equal a boundary value.

Example Boys Pre-Score

We take the class intervals to be as follows and construct a Grouped Frequency Table.

Class Interval		Frequency	Relative Frequency
11 - 20	~~////~~ ///	8	
21 - 30	~~////~~ ~~////~~	10	
31 - 40	~~////~~ ~~////~~ ///	13	
41 - 50	~~////~~ ////	9	
51 - 60	~~////~~	5	
61 - 70	/	1	
71 - 80	////	4	
81 - 90	~~////~~ //	6	
91 - 100	/	1	
		---	---
		57	

The intervals are of equal width, so we plot relative frequency which is the proportion of the data values falling within a given class interval.



Choosing An Appropriate Scale for a Histogram

In practice, if drawing a histogram by hand, it is usually a process of trial and error to obtain a reasonable number of class intervals

- too few and we lose the shape of the data
- too many and we cannot obtain the overall shape of the distribution.

A rule of thumb is : use $\sqrt{n}$ class intervals for n observations

Note that the use of relative frequency means that the histogram display does not depend on the sample size and so histograms based on different sample sizes may be compared easily.

1.4 Numerical Summaries

We have already considered the five number summary and here we will consider some other measures of location (centre, level) and spread (variation).

N.B. We give summaries for unimodal distributions i.e. distributions having one peak. If the data appear to come to a mix of two populations it is better to describe each separately.

1.4.1 Measures of Central Tendency or Level or Location

We consider the **mean, median and mode**.

Definition The **sample mode** is that data value which occurs most frequently in the sample. If the data are grouped into class intervals, then the modal class is the class interval with the highest frequency.

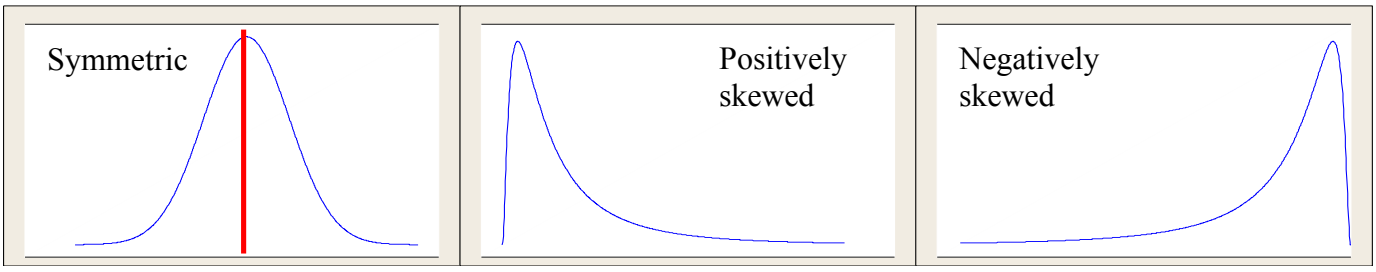
Definition The **sample mean** is the arithmetic average of the data values. If we denote the data values by $x_1, x_2, \ldots\ldots\ldots, x_n$ then

$$\text{Sample Mean} = \frac{x_1 + x_2 + \ldots\ldots\ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{Sum\ of\ data\ values}{Number\ of\ data\ values}$$

The sample mean is usually denoted by $\bar{x}$.

If the distribution is symmetric, the mean, median and mode all coincide.
If it is asymmetric then the mean and median can be quite different.
The sample mean is not resistant to the effect of outliers and so the sample median is a preferable measure in asymmetric distributions.



Mean
Median
Mode

e.g. For the data 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and 10000

Md = ...

and $\bar{x}$ = ...

1.4.2 Measures of Spread

What is spread (variability)? It happens because all observations are different from the sample mean. Hence we may measure spread using the deviations

$$x_1 - \bar{x}, \quad x_2 - \bar{x}, \ldots\ldots\ldots\ldots x_n - \bar{x}$$

of each data value from its sample mean.

How could we measure total variation? Unfortunately adding up these deviations just gives zero so we use

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \left\{ or \qquad \sum_{i=1}^{n}\left|x_i - \bar{x}\right| \right\}$$

as a measure of total variation. We define the sample variance of the data to be

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

For hand calculation, this equals

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}$$

Notice that we have omitted the subscripts from the formula i.e. it looks tidier!!

The **standard deviation** of the sample distribution is defined to be the square root of the sample variance $\sqrt{s^2}$.

There are two methods to perform the calculations:

Although it 'looks worse' Method 2 is much easier if we have a lot of data.

Especially if your calculator has the necessary functions!!

Method 1

To calculate the **variance & standard deviation - long hand** - we carry out the following:

(i) Find the mean $\bar{x}$
(ii) Find the deviation of each observation x from the mean $\bar{x}$, i.e. calculate x - $\bar{x}$.
(iii) Square this deviation to get $(x - \bar{x})^2$
(iv) Add all the squared deviations.
(v) Divide by (the number of observations - 1), i.e. divide by (n - 1):
This is the Sample Variance, s^2

(vi) Take the square root to finally obtain the standard deviation (s). (square root button on calculator is indicated by $\sqrt{}$.
This gives the Sample Standard Deviation, s

Method 2

To calculate the **variance & standard deviation using the hand calculation version** - we carry out the following:

(i) Add all the observations: to give the sum of the observations Σx
(ii) Square each observation
Add all the squared observations: to give the sum of the observations squared Σx^2
(iv) Square the sum of observations from (i) and divide by n : to give $\frac{(\Sigma x)^2}{n}$
(v) Subtract this from the sum of observations squared calculated in (ii): to give $\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n} \right)$

(v) Divide by (the number of observations - 1), ie divide by (n - 1):
This is the Sample Variance, s^2

Take the square root to finally obtain the standard deviation (s). (square root button on calculator is indicated by $\sqrt{}$.
This gives the Sample Standard Deviation, s

1.4.3 Example For the data 6, 4, 5, 2, 8, 3, 4, 2, 1, 1

Definition :
A resistant measure of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large the changes may be. i.e. a resistant measure is not sensitive to outliers.

Using Numerical Summaries

1. The standard deviation is often used as the measure of spread in preference to the variance because the units as the same as that of the observed data.

2. Another measure of spread sometimes used along with the interquartile range is the range. i.e. the difference between the maximum and minimum.

Note, however, that this measure, along with the maximum and minimum themselves, is sensitive to outliers.

3. When using numerical summaries, it is important to use the appropriate summary. i.e.

(i) Use either resistant or non-resistant measures (not a mixture of both).

(ii) If there are outliers or multiple peaks, the resistant measures of location and spread must be used.

Types of Numerical Summary

Type	Location	Spread
Nonresistant	Mean	Standard Deviation Variance
Resistant	Median	Quartiles Interquartile Range (IQR) Quartile Deviation

Appropriate Numerical Summaries

Shape	Location	Spread
Symmetric / No outliers	Mean	Standard Deviation Variance
Not symmetric or outliers	Median	Quartiles Interquartile Range (IQR) Quartile Deviation

Chapter 1: Summary

Chapter 2 Probability and Distributions

2.1 Mathematical Nature of Probability

Definition
The probability of some specific outcome as the proportion of times that outcome would occur if we repeated the experiment or observation a 'large' number of times.

Examples From our 'intuition', we would expect the following

(i) If we tossed an unbiased coin a large number of times we would intuitively expect that 50% of the times we would get heads and 50% tails i.e. probability of getting heads is 0.5.

(ii) Similarly, if we rolled an unbiased die, with faces labelled 1, 2, 3, 4, 5, and 6, a large number of times we would expect to get each face occurring 1/6 of the time.

Note that probability of any outcome is estimated by the relative frequency of the outcome when the experiment is repeated a large number of times.

In practice, when we have no 'intuition' rely on , the probability of any outcome is estimated in the same manner. In any case, our 'intuition' always has to be backed up by reality. In fact our intuition regarding tossing a coin and rolling a die is based on our past experiences.

Examples

(i) We can estimate the probability that a baby is a boy by observing what proportion of a large number of babies are boys.

(ii) We can estimate the probability that students smoke by collecting data on a large number of students.

2.2 Features of Probability

1. By definition a probability lies between 0 and 1 i.e. something that cannot happen has a probability of 0 and something that is certain to happen has a probability of 1.

So probability is similar to a proportion or a percentage: an outcome with a probability of 0.25 means that there is a one in four or a 25% chance of it happening.

2. In practice as mentioned above we have to estimate a probability because there is often no way of knowing the true value.

3. There are two rules that it is important to know

   (i) For a given event, for any two outcomes that might happen the probability of either occurring is the sum of the individual probabilities.

e.g. the probability of rolling a 2 with the unbiased die presented above is 1/6 or 0.167 and the probability of rolling either a 2 or a 3 is 2/6 or 0.33.

(ii) If we consider two or more different events which are independent of each other, then to get the probability of a combination of specific outcomes for each of the events we must multiply the individual probabilities.

Independence is an important concept. Independence means that the outcome of one event tells us nothing about the outcome of another event i.e. the probability of each possible outcome of the second event is the same regardless of the outcome of the first event.

e.g. the probability of rolling a 2 then a 3 in two rolls of an unbiased die is 1/6 x 1/6 = 1/36 = 0.028

Note that if two events are dependent, the multiplicative property does not apply. For example, if the probability of a man being more than six feet tall is 0.2, the probability that both he and his son are over six feet is not 0.2 x 0.2 = 0.04 because the heights of children tend to be related to the heights of their parents.

In fact, this idea is used in reverse in cases of uncertainty to investigate whether two events are independent.

We will discuss the concept of probability in more detail in future lectures.

2.3 Probability Distributions

In the introduction we talked about collecting data of a sample of the population of interest. We stressed that population need not be people but for example in agriculture we may be interested in the average crop yield on farms in Southern England in which case the population is 'farms in Southern England'.

We collect data on a sample and use the information to make inferences about the population. It is rare that we could ever contemplate collecting data on an entire population.

The relation between sample and population is subject to uncertainty and the ideas of probability are used to indicate the uncertainty. One important aspect in the context is the idea of a theoretical probability distribution.

In the previous chapter we spent a lot of time looking at the distribution of observed data using both graphical and numerical methods. In fact we could call the distribution of observed data the empirical distribution.

Many statistical methods use the related idea of a probability distribution which is specified mathematically.

A probability distribution is used to calculate the theoretical probability of different values occurring and is thus the theoretical equivalent of an empirical relative frequency histogram.

N.B. In a relative frequency histogram, the area of the bar above each class interval is equal to the proportion of total number of observations in that class. i.e. the total area of a relative frequency histogram is equal to one.

All probability distributions are described by one or more parameters. The mean and standard deviation are two examples of particular parameters.

2.3.1 The Normal Distribution

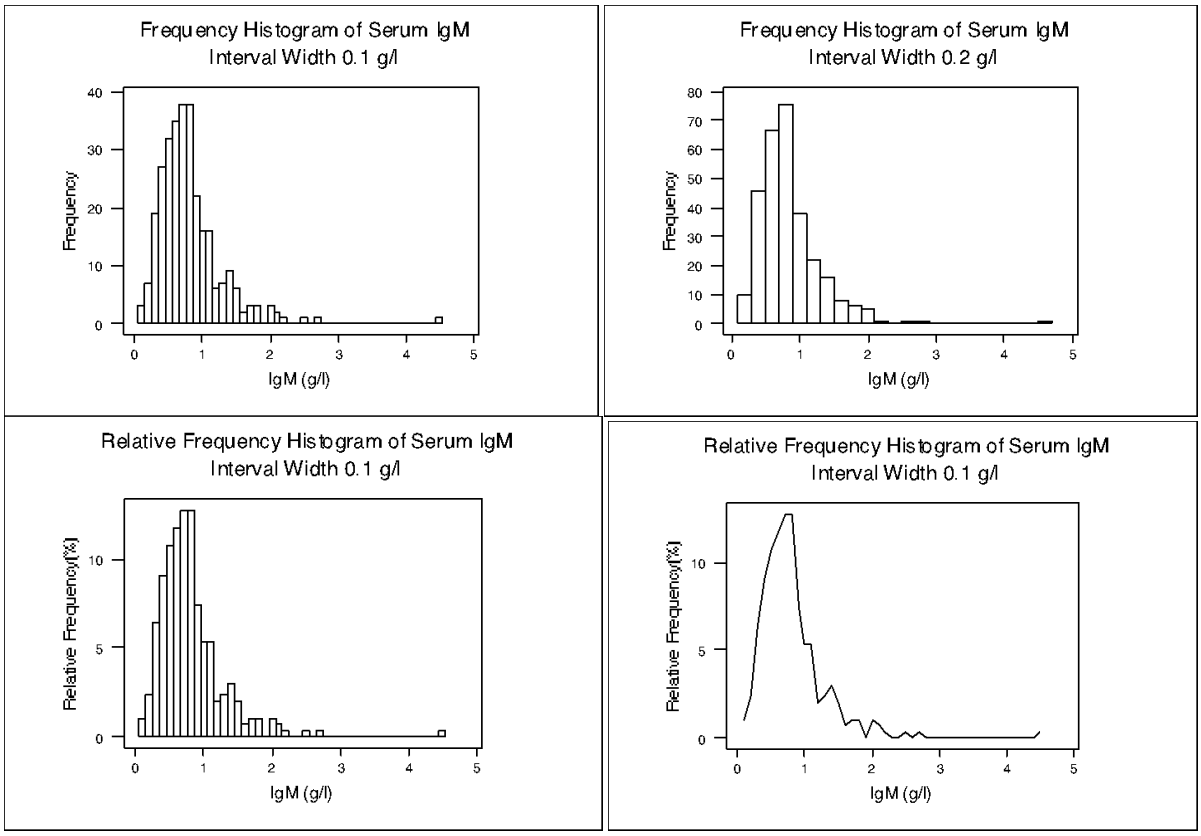
The Normal distribution is by far the most important probability distribution in statistics. It is important to understand its nature and role. However, it should be stressed that this distribution is no more normal (usual, common) than many others. It is sometimes also called the Gaussian distribution, after the mathematician Gauss.

We saw in the previous chapter that a histogram can be used to show the distribution of a continuous variable.

Example Immunoglobulin IgM in 298 healthy children

The following figure shows histogram of this data

a. Frequency Histogram (Interval width 0.1)
b. Frequency Histogram (Interval width 0.2)
c. Relative Frequency Histogram
d. Relative Frequency Polygon.

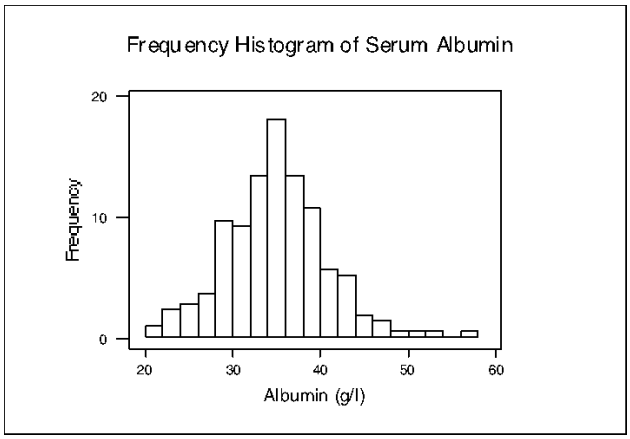


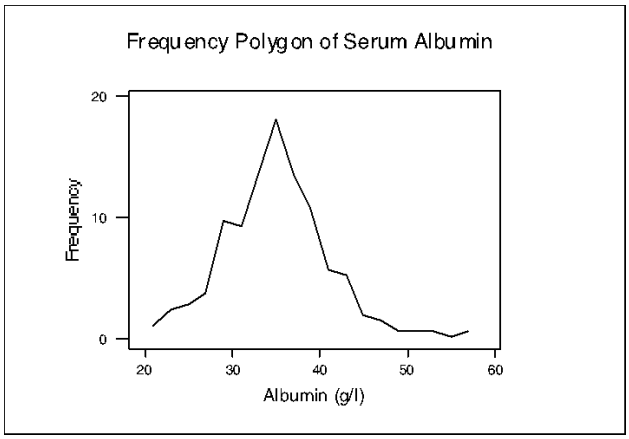
If there had been thousands of observation, and the IgM had been recorded more precisely , the IgM values could be divided into many tiny intervals, and the histogram would appear more like a smooth curve. So it is not difficult to imagine that the histogram or frequency polygon of some observed data is an approximation to some 'underlying' smooth frequency distribution.

Example Serum Albumin values in 216 patients with biliary cirrhosis

(i) Histogram (ii) Frequency Polygon

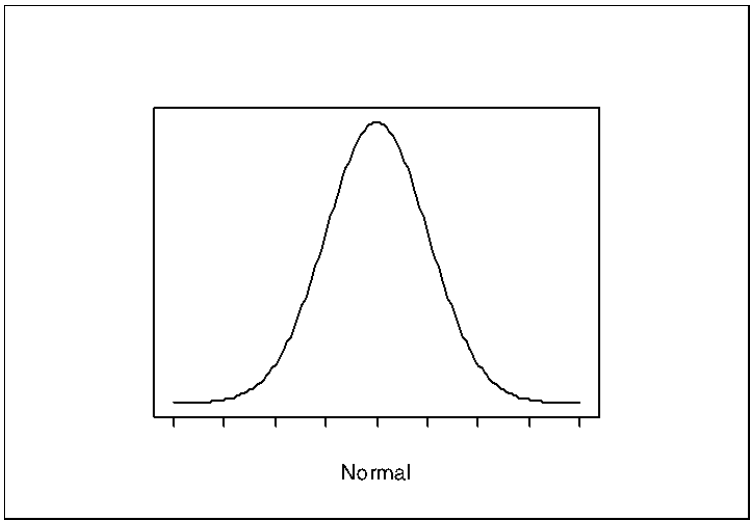
Note that the shape is easier to see using the frequency polygon.





Frequency distributions for continuous measurements such the serum albumin above often have a single peak (called unimodal). They may be symmetric (Serum Albumin) or asymmetric (Immunoglobulin).

The Normal distribution is a probability distribution which is unimodal and symmetric (its shape is shown below).



2.3.2 Important Facts about Continuous Probability Distributions

1. They usually have no upper or lower limit. In theory the Normal distribution extends from minus infinity ($-\infty$) to plus infinity ($+\infty$).

2. The height of the frequency curve, the probability density, cannot be taken as the probability of a particular value. This is because for a continuous variable there are infinitely many possible values so that the probability of any specific value is zero.

3. The height of the curve is of no practical use - its value is determined by the fact that the total area under the curve is always taken to be 1.

4. As with histograms of observed data, we use a probability distribution by considering the area corresponding to a particular restricted range of values. Because the total are is one this area corresponds to the probability of those values.

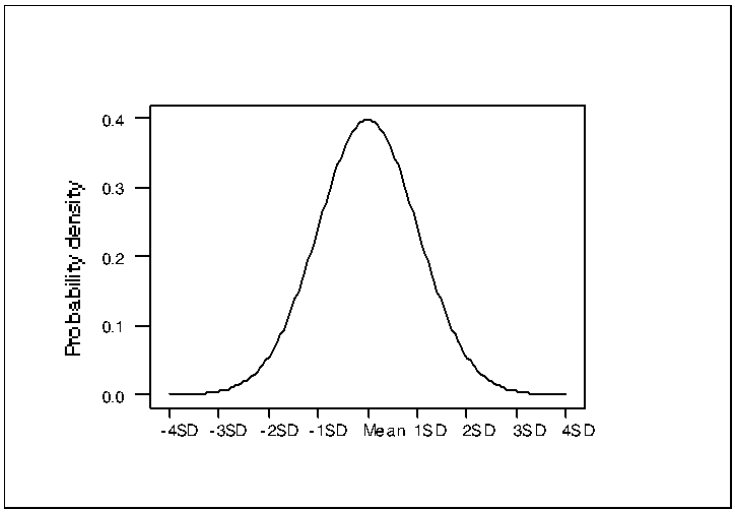
2.3.3 Using the Normal Distribution

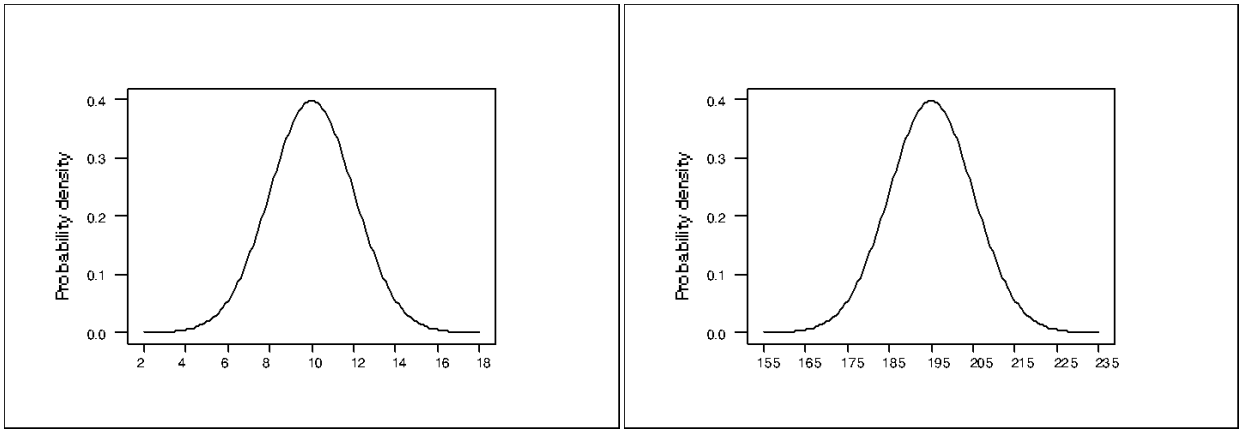
The mathematical equation of the Normal distribution is rather complicated but you will glad to hear that there is no need to know it in order to use the Normal distribution. All the necessary information is available via tables.

It is necessary to know that the Normal distribution is completely described by two parameters, the mean and the standard deviation. These are usually denoted as follows :

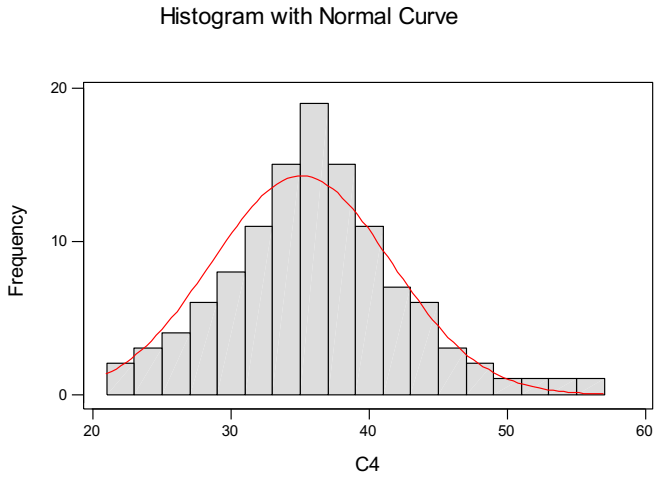
	the mean	is called	μ (mu)
and	the standard deviation	is called	σ (sigma)

The Normal distribution is related to the mean and standard deviation in the manner shown in the first figure below. That this is always the case is shown by the last two figures.



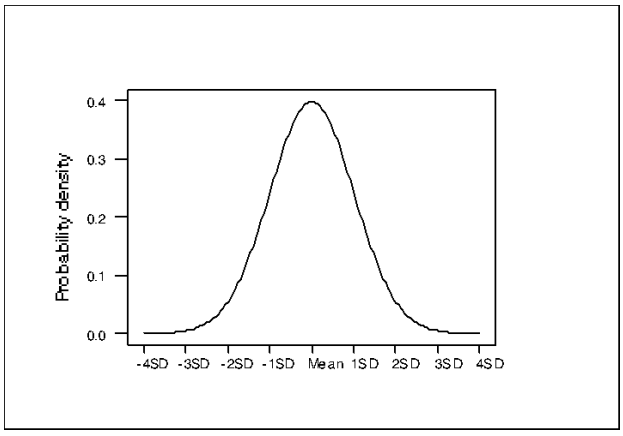


Now let us consider the serum albumin data again. The histogram of the data and the Normal distribution with the same mean and variance are very similar. In practice, we are often faced with deciding with whether data are 'Normally distributed'.



2.3.4Standard Normal Deviate

As the figure below shows, any position along the x axis can be expressed as a distance of a number of standard deviations (negative or positive) from the mean. This distance is known as a standard Normal deviate.



It is equivalent to looking at a Normal distribution with mean 0 and standard deviation of 1, a special Normal distribution known as the **standard Normal distribution**. Any normal distribution can be converted (or transformed) into a standard Normal distribution by subtracting the mean and dividing by the standard deviation.

Definition: Standard Normal Distribution
If a variable X has a normal distribution with mean μ and standard deviation σ, $N(\mu, \sigma^2)$, then the standardised variable

$$Z = \frac{X - \mu}{\sigma}$$

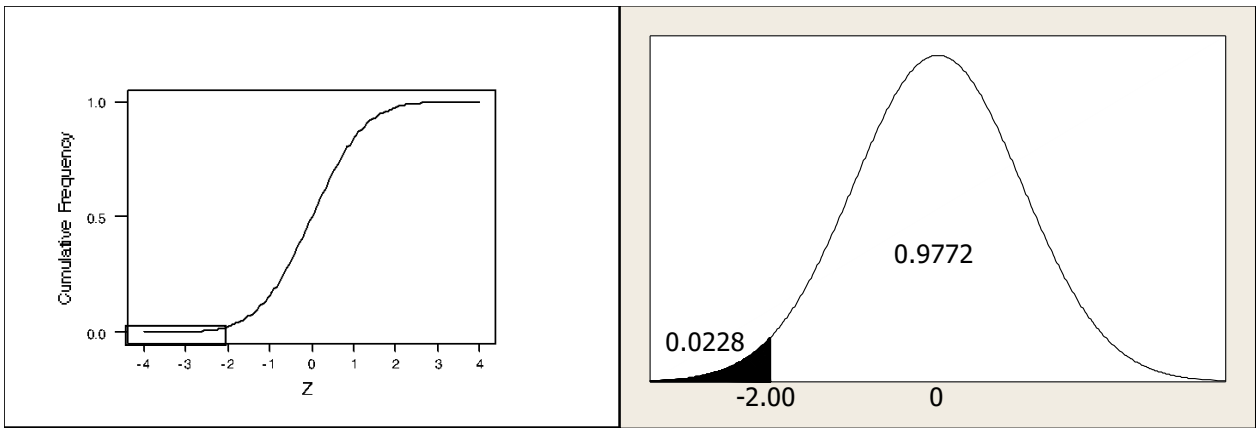
has the normal distribution N(0, 1) with mean 0 and standard deviation 1. This is called the standard normal distribution.

2.3.5 Standard Normal Table

The table at the end of these notes shows the lower tail areas of the standard Normal distribution. The lower tail means the area under the curve from $-\infty$ to the value of interest.

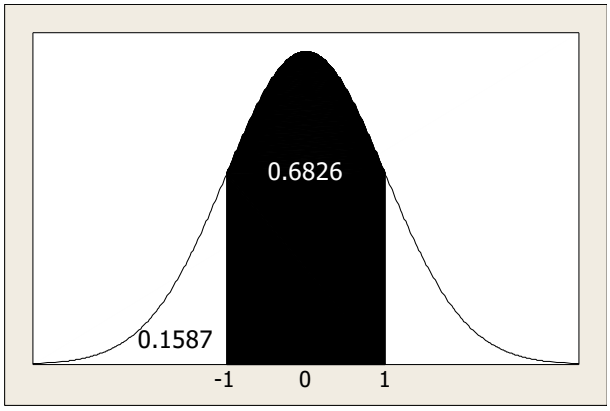
This is equivalent to the probability of a value lower than the specified value. This idea can also be expressed as the cumulative relative frequency distribution, which is shown in the following figure. The table is simply a more accurate version of the curve.

Example The area below -2 is 0.0228 (shaded area below) i.e. P(Z<-2.0) = 0.0228



Example The area above -2 is 1-0.0228 = 0.9772 (unshaded area above)
i.e. P(Z>-2.0) = 1- P(Z<-2) = 0.9772

Example The area between -1 & +1 i.e. P(-1<Z<1)



P(Z<1) = 0.8413
P(Z<-1) = 0.1587

P(-1<Z<1) = P(Z<1) - P(Z<-1)
= 0.8413 - 0.1587
= 0.6826

In other words, for data with an exactly Normal distribution there is a probability of 0.68 of being within one standard deviation of the mean.

Repeating these calculations for other numbers of standard deviations we get

Range	Probability of being within range	outside range
mean ± 1 SD	0.683	0.317
mean ± 2 SD	0.954	0.046
mean ± 3 SD	0.9973	0.0027

In each the probability of not being within the stated range is 1 minus the probability of being within the range. We see that there is a minimal chance - 0.0027 or 0.27% or about 1 in 400 - that a value from a Normal distribution will be more than three standard deviations above or below the mean. Note that this agrees with the visual impression gained from looking at previous figures in these notes. Of course, in very large samples we would expect several values to be this extreme.

The probability of being within two standard deviations of the mean is just over 0.95. In other words, about 95% of observations from a Normal distribution will be within the range mean - 2SD to mean +2SD.

As we will see later, exactly 95% of the area under the Normal distribution curve actually falls within the slightly narrower range of mean ± 1.96 SD.

We can also find the values which enclose a given percentage of the distribution - the central range.

For example, 90% of the distribution lies within the range mean ± 1.645 SD, 95% within the range mean ± 1.96SD and 99% within the range mean ± 2.57 SD.

Thus,

P(X < mean - 1.645 SD) = 0.05
and P(X > mean + 1.645 SD) = 0.05

i.e.

P(X < mean - 1.645 SD) = 0.05
and P(X < mean + 1.645 SD) = 0.95

2.3.6 Using the Normal Distribution (continued)

One way we use the Normal distribution is as follows. When a set of observations has a distribution that is similar to a Normal distribution we assume that in the population the distribution of the variable actually is Normal and carry out calculations on this basis.

Example We can calculate the probability that a patient with primary biliary cirrhosis has a serum albumin level less than 42.0 g / l if we are willing to assume that, among the population of all patients with primary biliary cirrhosis, serum albumin has a Normal distribution.

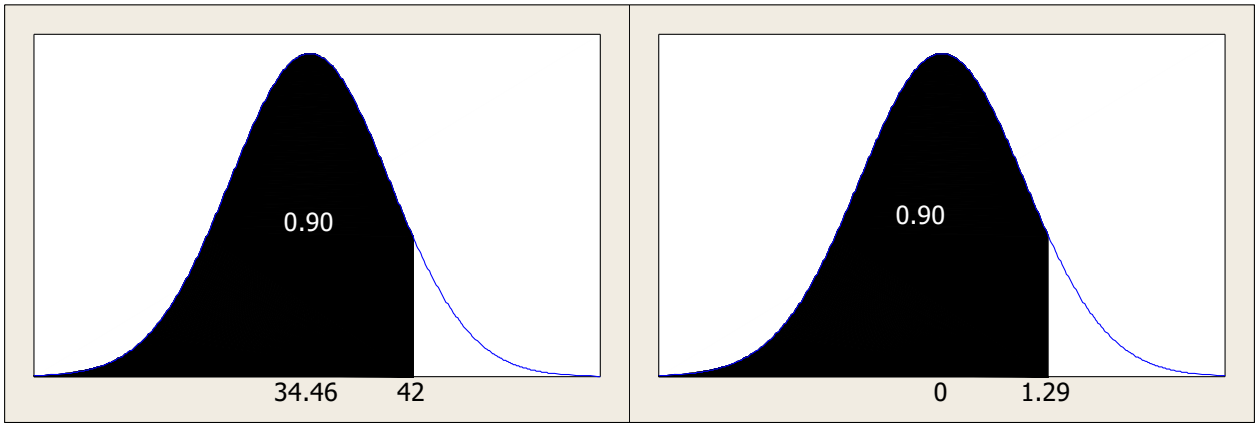
Solution

We can calculate the probability of a value being less than 42.0 on the assumption that the true distribution is Normal. The mean serum albumin level was 34.46 g/l and the standard deviation was 5.84 g/l. We first calculate how many standard deviations from the mean the value of 42 g/l is, given by

$$\frac{42 - 34.46}{5.84} = 1.29$$

From the Table at the end of these notes of the standard normal probabilities, we find that the probability of being less than 1.29 is 0.90. Therefore, the probability of being less than 1.29 is 0.90, so the probability of a value than 42 g/l is 0.90 or 90%.

P(X<42) P(Z<1.29)



Example From above, 90% of the distribution lies within the range mean $\pm$ 1.645 SD, 95% within the range mean $\pm$ 1.96SD and 99% within the range mean $\pm$ 2.57 SD.

For the serum albumin data (mean 34.46 g/l and the standard deviation was 5.84 g/l) we get the following ranges

Central Range	Serum Albumin (g/l)
90%	24.85 to 44.07
95%	23.01 to 45.91
99%	19.39 to 49.53

We can thus use the Normal distribution to estimate the centiles of the distribution of the variable in the population. we could have calculated the observed centiles of the sample data and used these values as estimates of the population centiles but when the data are near Normal the use of the Normal distribution is more reliable especially in the tails of the distribution. It is also easier requiring just two values and a table of the Normal distribution rather than the whole set of raw data.

We noted previously that many statistical methods incorporate important assumptions about the distribution of data. (These methods are called parametric methods). In most cases the distribution involved is the Normal distribution which is one of the reasons why it is the most important distribution in statistics.

Although many measurements do have a reasonably Normal distribution , such as human height , many do not such as human weight or serum cholesterol. There are various ways in which data may deviate from Normality, notably by being asymmetric or skewed. The IgM data plotted in previous lectures illustrated positive skewness.

It should not be assumed that a set of observations is approximately Normal - this must be established.

Examples

1. Find the total area under the normal curve right of the mean.

2. Give the z-value corresponding to the 80th percentile for the Normal curve.

3. Suppose that X is a normally distributed random variable. Find the probability that X will fall within:

   (a) 2.58 standard deviations of its mean
   (b) 2.33 standard deviations of its mean
   (c) 1.28 standard deviations of its mean

4. Suppose that X is a Normally distributed random variable with mean equal to 60 and standard deviation equal to 5. How many standard deviations above or below the mean are the following values of X?

   (a) 53.6

(b) 63
(c) 47
(d) 36
(e) 69.5

5. A large set of test grades was approximately normally distributed with mean equal to 712 and standard deviation equal to 120. If your grade on the test was 920, how did your grade rank relative to the other grades? i.e. calculate the approximate percentile score of your grade.

6. The mean and standard deviation of the scores on a national achievement test were 655 and 146, respectively. If you were informed that you scored at the 90th percentile, what was your test score ?

2.4 Binomial Distribution, Probability and Inference

2.4.1 Binomial Experiments

A poll was conducted, in January 2005, to investigate the opinions of Scottish adults on their attitudes the new smoking legislation. The sampled population consisted of all adults living in the Scotland. Each adult represented an element in the population and each possessed or did not possess a particular characteristic - they either agreed or did not agree with a statement posed in the survey.

A sample of n = 1000 adults was selected from the population and the number x of people in the sample favouring a statement was recorded. For example, x = 590 or 59% of all people in the sample supported new legislation.

The objective of this sampling was to use the sample data, the number x in the sample who respond 'yes' to a particular question, to estimate the proportion (or percentage) of all Scots who possess the same opinion.

The above opinion poll is typical of a whole class of similar statistical problems. All involve sampling n elements from a population where each element can elicit one, and only one, of two responses - yes or no, red or white, agree or disagree, and so on. Sampling that satisfies these conditions is called a **binomial experiment** and the random variable x is called a binomial random variable.

2.4.2 Characteristics of a Binomial Experiment

1. The experiment consists of n identical trials.

2. Each trial can result in one and only one of two possible outcomes. we will denote one outcome by the symbol S (for success) and the other by the symbol F (for failure).

3. The probability of success S on a single trial is equal to p and remains the same from trial to trial. The probability of a failure F on a single trial is equal to q = (1 - p).
Note that p + q = 1.

4. The trial outcomes are independent.

5. The binomial random variable x is the number of successes in n trials.

To determine whether a particular discrete random variable is a binomial random variable, check the experiment that gave rise to the variable to see whether it satisfies the five characteristics of a binomial experiment.

Example Toss a coin n times and observe the number x of heads. Explain why x is or is not a binomial random variable.

Solution

1. The experiment consists of n = 10 identical tosses of a coin. Each toss is a trial.

2. Each toss results in one of two possible outcomes : a head, denoted as a success S, or a tail, which we will denote as a failure F.

3. The probability of a head on a single trial is 1/2 (assuming that the coin is balanced).

4. The trials are independent.

5. The variable x is the number of heads (successes) in n = 10 tosses of the coin.

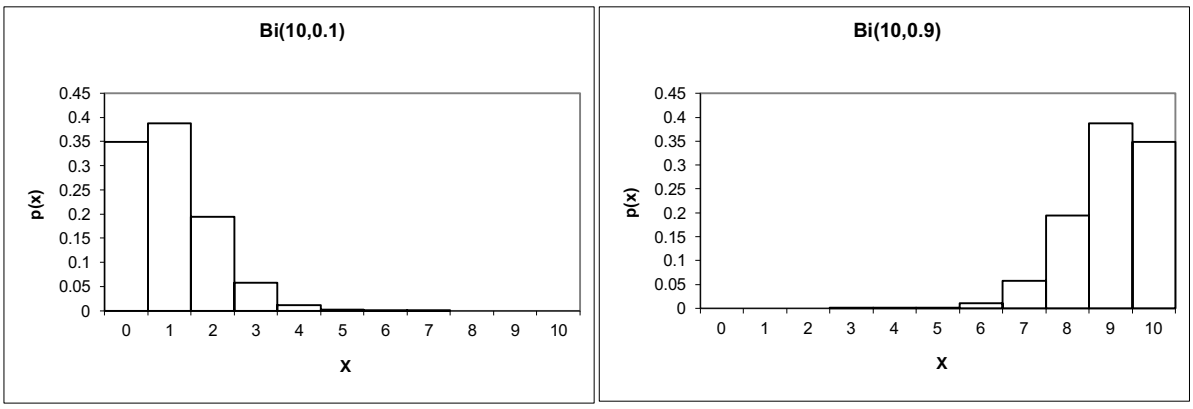
Since the experiment satisfies the five characteristics of a binomial experiment, x is a binomial random variable.

Example Consider whether the opinion poll example fulfils the five characteristics of a binomial experiment.

2.4.3 The Binomial Probability Distribution

The binomial probability distribution p(x) depends on the number of trials and the probability p of success. We do not need to know the formula for p(x) because the probabilities have been tabulated for small samples and they can be approximated for large samples using an approximation (see later).

The following figures show the binomial distributions for n = 10 and p = 0.1, 0.9 and 0.5.



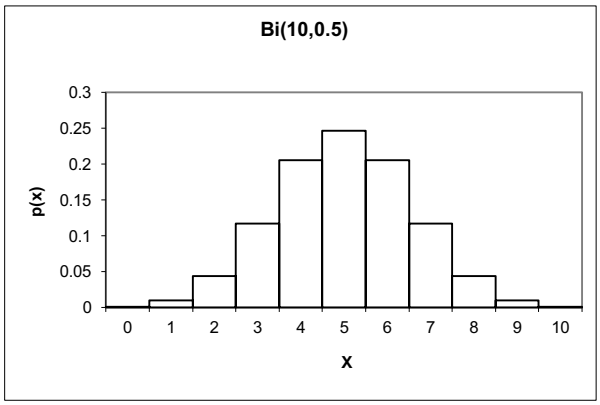
The Mean and Standard Deviation of a Binomial Probability Distribution

The formulae for calculating the mean μ and the standard deviation σ for a particular binomial distribution (i.e. for given values of n and p) are :

Mean : $\mu = np$

Standard Deviation : $\sigma = \sqrt{np(1-p)}$

For example, the binomial distribution for n = 10 and p = 0.5 (Figure c above) has



$\mu = np = 10 \times 0.5 = 5$

$\sigma = \sqrt{np(1-p)} = \sqrt{10x0.5x0.5} = \sqrt{2.5} = 1.58$

Note that when p =0.5, the binomial probability distribution is symmetric about its mean.

When p = 0.1 (Figure a),

$\mu = np = 10 \times 0.1 = 1$

$\sigma = \sqrt{np(1-p)} = \sqrt{10x0.1x0.9} = \sqrt{0.9} = 0.95$

and the distribution is skewed to the right (positively skewed).

When p = 0.9 (Figure b),

$\mu = np = 10 \times 0.9 = 9$

$\sigma = \sqrt{np(1-p)} = \sqrt{10x0.9x0.1} = \sqrt{0.9} = 0.95$

and the distribution is skewed to the left (negatively skewed).

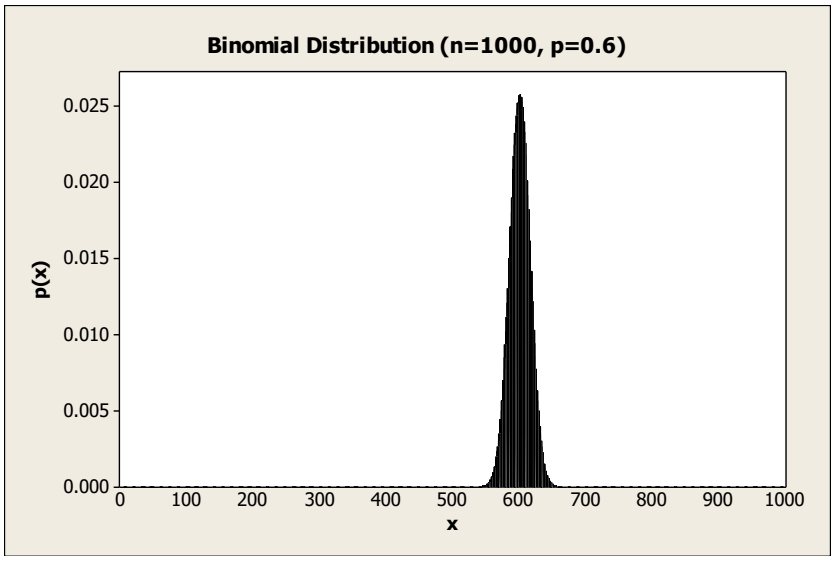
When the sample size n is large enough so that the interval $\mu - 3\sigma$ to $\mu + 3\sigma$ falls in the interval x = 0 to x = n, the binomial probability distribution will be nearly symmetric about its mean.

If $\mu - 3\sigma$ is less than 0, the distribution will tend to be skewed to the right. If $\mu + 3\sigma$ is larger than n, the distribution will tend to be skewed to the left.

Example

A political analyst claims that 60% of the eligible voters in UK favour a local income tax. Suppose that a random sample of 1000 voters is selected from the population of all voters in the country.

Solution



2.4.4 An Example of Statistical Inference Using the Binomial Probability Distribution

An automobile manufacturer claims that fewer than 5% of its car buyers complain of lack of service from its dealers. A survey of 500 buyers of the manufacturer's new cars found that 39 new car buyers were dissatisfied with their dealer's service. Is this sample outcome inconsistent with the manufacturer's claim?

N.B. This problem is asking us to make an inference about a binomial parameter p. It is our first attempt to employ statistical inference.

A. Suppose that the manufacturer's claim is true; i.e. only 5% or less of its car buyer's complain of lack of service. Describe the probability distribution of the number x of buyers in a sample of 500 who complain about the service.

B. Why will answering the question 'Is this sample outcome inconsistent with the manufacturer's claim?' result in an inference about a population parameter?

C. If the manufacturer's claim is true, is it likely that the number x of dissatisfied customers is as large as 39?

D. What can we conclude about the manufacturer's claim?

Solution

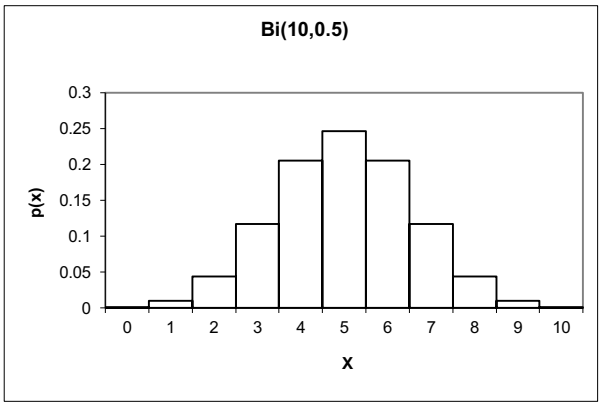
2.4.5 Binomial Probability Tables

The table at the end of these notes gives values of the binomial probability distribution p(x) for values of n = 5, 10, 15, 20, and 25 and p = 0.01, 0.05, 0.1, 0.2,.....0.9, 0.95, 0.99.

The values of p appear across the top of the table. The probabilities for a binomial probability distribution with p = 0.4 appear in the column under 0.4. The values of n and x are shown at the left side of the table. To find the value of p(x) for n=5, x = 3 and p =0.4, we read the tabulated number in the row corresponding to n = 5 & x = 3 and the column for p = 0.4. The value is p(3) = 0.2304.

Example
Find the probability that x is less than or equal to 3 for a binomial random variable with n = 10 and p = 0.5.

Solution
Using the table, to find the probability that x is less than or equal to 3, we need to find the sum of p(x) from x = 0 to x = 3. See the figure below.



$P(x \leq 3)$	=	p(0)	+	p(1)	+	p(2)	+	p(3)
	=	0.0010	+	0.0098	+	0.0439	+	0.1172
	=	0.1719						

Example

A manufacturer of cameras claims that no more than 1% of its cameras are defective.

a. Suppose that the manufacturer's claim is true and that the proportion defective in the population is p = 0.01. If you receive a shipment of n = 25 cameras, what is the probability that more than one will be defective?

b. If the number x of defective cameras in a sample is 2, what would you conclude about the manufacturer's claim?

Solution

It is convenient to use tables to calculate the probabilities but we should also be able to calculate them for ourselves:

2.4.6 Binomial Expansion

1. Consider an event which has just two possible outcomes: Success or failure. The probability of a success is p and of a failure is q. What are the respective probabilities of 0, 1, 2 successes if the event occurs twice?

N.B. Independent events

(a) Probability of no successes =

(b) Probability of one success can happen in two ways:

either a success followed by a failure or a failure followed by a success =

(c) Probability of two successes is =

These probabilities are identical with the terms formed by the expansion of $(q + p)^2$

i.e.

2. Extending the analysis to the case where the event occurs three times, what are the probabilities of 0, 1, 2 and 3 successes?

(a) Probability of no successes =

(b) Probability of one success =

(c) Probability of two successes =

(d) Probability of three successes =

These probabilities are successive terms in the expansion $(q + p)^3$

i.e.

3. It can be shown, quite generally, that if the event were repeated n times, the respective probabilities of 0, 1, 2, ... n successes would be given by successive terms in the expansion of $(q + p)^n$

Once the expansion of the expression has been carried out, the respective values of q and p can be inserted, and the probability of any number of successes in any number of events can be calculated.

The expression (q + p) is called a binomial and hence the distribution of probabilities formed by the expansion of $(q + p)^n$ is called the binomial probability distribution.

2.4.7 Pascal's Triangle

The numerical coefficient of the terms of $(p + q)^n$ are

$$1, \quad n, \quad \frac{n(n-1)}{2!}, \quad \frac{n(n-1)(n-2)}{3!}, \quad etc$$

These coefficients may be obtained from an arrangement of numbers known as Pascal's Triangle, part of which is shown below :

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1 etc.

Table : Normal Distribution

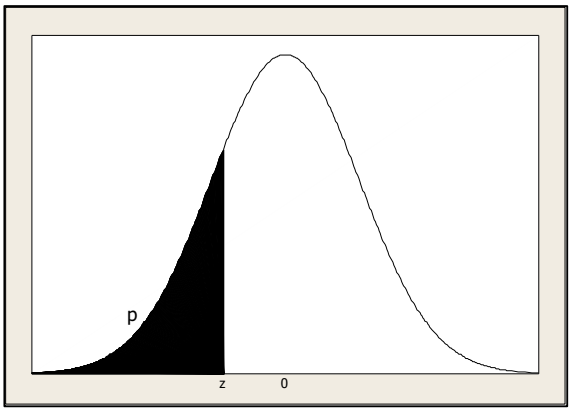


Table of Standard Normal Probabilities [Table entry for z is the probability lying below it]

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table : Normal Distribution

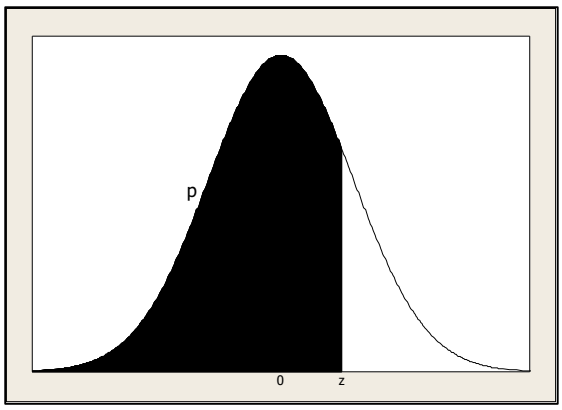


Table of Standard Normal Probabilities [Table entry for z is the probability lying below it]

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Binomial Tables

		p												
n	**x**	**0.01**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**0.95**	**0.99**
5	**0**	0.9510	0.7738	0.5905	0.3277	0.1681	0.0778	0.0313	0.0102	0.0024	0.0003	0.0000	0.0000	0.0000
	1	0.0480	0.2036	0.3281	0.4096	0.3602	0.2592	0.1563	0.0768	0.0284	0.0064	0.0005	0.0000	0.0000
	2	0.0010	0.0214	0.0729	0.2048	0.3087	0.3456	0.3125	0.2304	0.1323	0.0512	0.0081	0.0011	0.0000
	3	0.0000	0.0011	0.0081	0.0512	0.1323	0.2304	0.3125	0.3456	0.3087	0.2048	0.0729	0.0214	0.0010
	4	0.0000	0.0000	0.0005	0.0064	0.0284	0.0768	0.1563	0.2592	0.3602	0.4096	0.3281	0.2036	0.0480
	5	0.0000	0.0000	0.0000	0.0003	0.0024	0.0102	0.0313	0.0778	0.1681	0.3277	0.5905	0.7738	0.9510
10	**0**	0.9044	0.5987	0.3487	0.1074	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0914	0.3151	0.3874	0.2684	0.1211	0.0403	0.0098	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.0042	0.0746	0.1937	0.3020	0.2335	0.1209	0.0439	0.0106	0.0014	0.0001	0.0000	0.0000	0.0000
	3	0.0001	0.0105	0.0574	0.2013	0.2668	0.2150	0.1172	0.0425	0.0090	0.0008	0.0000	0.0000	0.0000
	4	0.0000	0.0010	0.0112	0.0881	0.2001	0.2508	0.2051	0.1115	0.0368	0.0055	0.0001	0.0000	0.0000
	5	0.0000	0.0001	0.0015	0.0264	0.1029	0.2007	0.2461	0.2007	0.1029	0.0264	0.0015	0.0001	0.0000
	6	0.0000	0.0000	0.0001	0.0055	0.0368	0.1115	0.2051	0.2508	0.2001	0.0881	0.0112	0.0010	0.0000
	7	0.0000	0.0000	0.0000	0.0008	0.0090	0.0425	0.1172	0.2150	0.2668	0.2013	0.0574	0.0105	0.0001
	8	0.0000	0.0000	0.0000	0.0001	0.0014	0.0106	0.0439	0.1209	0.2335	0.3020	0.1937	0.0746	0.0042
	9	0.0000	0.0000	0.0000	0.0000	0.0001	0.0016	0.0098	0.0403	0.1211	0.2684	0.3874	0.3151	0.0914
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0010	0.0060	0.0282	0.1074	0.3487	0.5987	0.9044
15	**0**	0.8601	0.4633	0.2059	0.0352	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1303	0.3658	0.3432	0.1319	0.0305	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0092	0.1348	0.2669	0.2309	0.0916	0.0219	0.0032	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0004	0.0307	0.1285	0.2501	0.1700	0.0634	0.0139	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0049	0.0428	0.1876	0.2186	0.1268	0.0417	0.0074	0.0006	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0006	0.0105	0.1032	0.2061	0.1859	0.0916	0.0245	0.0030	0.0001	0.0000	0.0000	0.0000
	6	0.0000	0.0000	0.0019	0.0430	0.1472	0.2066	0.1527	0.0612	0.0116	0.0007	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0003	0.0138	0.0811	0.1771	0.1964	0.1181	0.0348	0.0035	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0035	0.0348	0.1181	0.1964	0.1771	0.0811	0.0138	0.0003	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0007	0.0116	0.0612	0.1527	0.2066	0.1472	0.0430	0.0019	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0001	0.0030	0.0245	0.0916	0.1859	0.2061	0.1032	0.0105	0.0006	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0006	0.0074	0.0417	0.1268	0.2186	0.1876	0.0428	0.0049	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0001	0.0016	0.0139	0.0634	0.1700	0.2501	0.1285	0.0307	0.0004
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0032	0.0219	0.0916	0.2309	0.2669	0.1348	0.0092
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0047	0.0305	0.1319	0.3432	0.3658	0.1303
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0047	0.0352	0.2059	0.4633	0.8601

Binomial Tables (continued)

		p											
n	**x**	**0.01**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**0.95**	**0.99**
20	**0**	0.8179	0.3585	0.1216	0.0115	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1652	0.3774	0.2702	0.0576	0.0068	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0159	0.1887	0.2852	0.1369	0.0278	0.0031	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0010	0.0596	0.1901	0.2054	0.0716	0.0123	0.0011	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0133	0.0898	0.2182	0.1304	0.0350	0.0046	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0022	0.0319	0.1746	0.1789	0.0746	0.0148	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0003	0.0089	0.1091	0.1916	0.1244	0.0370	0.0049	0.0002	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0020	0.0545	0.1643	0.1659	0.0739	0.0146	0.0010	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0004	0.0222	0.1144	0.1797	0.1201	0.0355	0.0039	0.0001	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0001	0.0074	0.0654	0.1597	0.1602	0.0710	0.0120	0.0005	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0020	0.0308	0.1171	0.1762	0.1171	0.0308	0.0020	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0005	0.0120	0.0710	0.1602	0.1597	0.0654	0.0074	0.0001	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0001	0.0039	0.0355	0.1201	0.1797	0.1144	0.0222	0.0004	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0010	0.0146	0.0739	0.1659	0.1643	0.0545	0.0020	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0002	0.0049	0.0370	0.1244	0.1916	0.1091	0.0089	0.0003	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0013	0.0148	0.0746	0.1789	0.1746	0.0319	0.0022	0.0000
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0046	0.0350	0.1304	0.2182	0.0898	0.0133	0.0000
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0011	0.0123	0.0716	0.2054	0.1901	0.0596	0.0010
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0031	0.0278	0.1369	0.2852	0.1887	0.0159
	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0068	0.0576	0.2702	0.3774	0.1652
	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008	0.0115	0.1216	0.3585	0.8179
25	**0**	0.7778	0.2774	0.0718	0.0038	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1964	0.3650	0.1994	0.0236	0.0014	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0238	0.2305	0.2659	0.0708	0.0074	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0018	0.0930	0.2265	0.1358	0.0243	0.0019	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0001	0.0269	0.1384	0.1867	0.0572	0.0071	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0060	0.0646	0.1960	0.1030	0.0199	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0010	0.0239	0.1633	0.1472	0.0442	0.0053	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0001	0.0072	0.1108	0.1712	0.0800	0.0143	0.0009	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0018	0.0623	0.1651	0.1200	0.0322	0.0031	0.0001	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0004	0.0294	0.1336	0.1511	0.0609	0.0088	0.0004	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0001	0.0118	0.0916	0.1612	0.0974	0.0212	0.0013	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0040	0.0536	0.1465	0.1328	0.0434	0.0042	0.0001	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0012	0.0268	0.1140	0.1550	0.0760	0.0115	0.0003	0.0000	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0003	0.0115	0.0760	0.1550	0.1140	0.0268	0.0012	0.0000	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0001	0.0042	0.0434	0.1328	0.1465	0.0536	0.0040	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0013	0.0212	0.0974	0.1612	0.0916	0.0118	0.0001	0.0000	0.0000
	16	0.0000	0.0000	0.0000	0.0000	0.0004	0.0088	0.0609	0.1511	0.1336	0.0294	0.0004	0.0000	0.0000
	17	0.0000	0.0000	0.0000	0.0000	0.0001	0.0031	0.0322	0.1200	0.1651	0.0623	0.0018	0.0000	0.0000
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0009	0.0143	0.0800	0.1712	0.1108	0.0072	0.0001	0.0000
	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0053	0.0442	0.1472	0.1633	0.0239	0.0010	0.0000
	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0016	0.0199	0.1030	0.1960	0.0646	0.0060	0.0000
	21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0071	0.0572	0.1867	0.1384	0.0269	0.0001
	22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0019	0.0243	0.1358	0.2265	0.0930	0.0018
	23	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0074	0.0708	0.2659	0.2305	0.0238
	24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0014	0.0236	0.1994	0.3650	0.1964
	25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0038	0.0718	0.2774	0.7778

Chapter 3 Introduction to Hypothesis Testing and Confidence Intervals

In the previous section we posed the first of many questions that come under the heading of Statistical Inference. We have also discussed how we might summarise the data we collect

Example The mean birth rate in 2004.
The proportion of the sample expressing the wish to vote Conservative.
The mean rainfall in Stirling during 2006.

Each of the above summaries is called a statistic and we usually collect data from a sample of the population in order to draw conclusions about the corresponding parameter in the wider population (statistical inference).

Definitions

Parameter A parameter is a number describing the population.

Statistic A statistic is a number that can be computed from the data without making use of any unknown parameters.

Example Voting Preference

Let us take the opinion poll as an example and think about it in more detail. We want to estimate the proportion θ of the population who will vote Conservative. The poll reports that of the 1100 randomly selected adults questioned, 654 say that they will vote Conservative.

The sample proportion is $\hat{\theta}$ = 654 / 1100 = 0.59 is a statistic that we can use to estimate the unknown parameter θ.

How can $\hat{\theta}$ based on a sample of 1100 out of the entire voting population (even if all precautions are taken to ensure that it is a random sample) be an accurate estimate of θ? Another sample taken at the same time would ask different people and therefore probably produce a different value of $\hat{\theta}$. This is called sampling variability i.e. the value of a statistic varies in repeated sampling.

In fact, it can be shown that if we examined the distributions of several of the commonly used statistics we would find that the distributions are approximately normal.

i.e. Another reason why the normal distribution is important in statistics.

Crucial to the above statement is that the samples are randomly selected. When randomisation is used in a design for producing data, statistics computed from the data have a definite pattern of behaviour over many repetitions, even though the result of a single repetition is uncertain. The regular pattern of outcomes described by the sampling distribution is the basis for understanding how trustworthy a statistic is as an estimator of a parameter.

Definition

Sampling Distribution	The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same sample size from the same population.
Bias	A systematic error in the estimation process that causes the statistic to regularly miss the true value of the parameter in the same direction.
Unbiased Estimator	A statistic used to estimate a parameter is an unbiased estimator of the parameter if the mean of its sampling distribution is equal to the true value of the parameter.
Variability of a Statistic	The variability of a statistic from a probability sampling design is described by the spread of its sampling distribution. The spread is determined by the sampling design and the size of the sample.

In can be shown that the mean of the sampling distribution of the sample proportion $\hat{\theta}$ from a simple random sample is always equal to the population proportion θ. We can measure bias numerically as the difference between the mean of the sampling distribution and the true value of the parameter.

Unbiasedness means that the mean of the distribution reflects the truth about the population. An unbiased statistic computed from an individual sample will sometimes fall above and sometimes below the true value but there is systematic tendency to over- or under-estimate the parameter.

We also must take the variability of the statistic into account using the analogy of an archer, we require the estimate not only to be on target (unbiased) and close together (low variability).

Randomisation produces a sampling distribution and eliminates bias. The variability is controlled by the size of the sample i.e. larger samples produce less variability.

We now know that when making inferences about a population parameter we also must include some idea of the variability of the estimate.

There are two common types of formal statistical inference:

Hypothesis Tests and Confidence Intervals

The appropriate method will depend on the question being asked.

3.1.1 Hypothesis Tests / Tests of Significance

This method is appropriate when the question is of form:

Is there sufficient evidence provided by the data in favour of some claim about the population?

Example The average age of students registering for First Year Course at Stirling is 19.43 years

3.1.2 The Null and Alternative Hypotheses

The first step in hypothesis testing is to state a claim against which we will try to find evidence. This is called the **Null Hypothesis.**

Null Hypothesis The statement being tested in a hypothesis test is called a null hypothesis. The test is designed to assess the strength of the evidence against the null hypothesis.
The null hypothesis is usually a statement of 'no effect' or 'no difference'.

Usually the phrase 'the null hypothesis' is written as H_0. The **null hypothesis** is a statement about a population, in terms of some parameter or parameters.

Example If μ is the mean age of students registering for First Year Courses at Stirling our null hypothesis could be

$$H_0 \quad : \quad \mu = 19.43$$

This implies that the mean age of first year students is 19.43.

The hypothesis which we hope or suspect is true is called the **alternative hypothesis** and is abbreviated to H_1.

Example $$H_1 \quad : \quad \mu \neq 19.43$$

This says that we suspect the mean age is not equal to 19.43.

N.B. The hypotheses always refer to the population not to a particular outcome.

3.1.3 One- and Two-Sided Alternatives

The alternative hypothesis can take one of two forms depending on the question asked. For example, in the Age of First Year Students example, we could be asking

1) Is the mean age different from 19.43?

2) Is the mean age less than 19.43?
3) Is the mean age greater than 19.43?

N.B. 2) and 3) are the same except for the direction of the difference.

The hypotheses for the above are written as:

1) $H_o : \mu = 19.43$
$H_1 : \mu \neq 19.43$

2) $H_o : \mu = 19.43$
$H_1 : \mu < 19.43$

3) $H_o : \mu = 19.43$
$H_1 : \mu > 19.43$

Note that the null hypothesis is the essentially the same in each case. The alternative hypothesis is taking two different forms i.e.

1) Two-Sided Alternative

2) and 3) One- Sided Alternative

Exactly which form the alternative hypothesis takes **must** be decided upon at the planning stage of any research.

Examples What hypotheses should be used to answer the following?

1. Do more than 30% of pet owners own dogs?

2. Is the average height of a 6-year old girl the same as the average height of a 6-year old boy?

3. Do more 20-year-old males than 20 year old females go to the gym?

4. Is the average jail sentence for murder is more than fifteen years?

5. It is thought that a new drug has been invented which can make people lower cholesterol. 20 people are chosen at random and their cholesterol levels are measured. They are then given this new drug and their cholesterol levels are measured again 9 months later. We wish to know if the drug does alter people's cholesterol levels

3.1.4 What are p-values?

We have stated that we looking for evidence against a hypothesis so we require an objective rule for deciding when we have sufficient evidence. A hypothesis test assesses the evidence against the null hypothesis in terms of probability.

i.e. If the observed outcome is unlikely under the supposition that the null hypothesis is true, but is more probable if the alternative is true, that outcome is evidence against H_0 in favour of H_1. The less probable the outcome is, the stronger the evidence that H_0 is false.

In general, a hypothesis test finds the probability of getting an outcome as extreme or more extreme than the actually observed outcome. 'Extreme' means ' far from what we would expect if H_o were true'.

p-value The probability, computed assuming that H_0 is true, that the test statistics would take a value as extreme or more extreme than that actually observed is called the p-value of the test. The smaller the p-value, the stronger the evidence against H_0 provided by the data.

The final step often taken is to compare this p-value to a fixed value regarded as decisive (i.e. our rule defining when we have sufficient evidence to reject the null hypothesis).

This cut-off value is called the significance level and is denoted by α. Usually the value chosen is $\alpha = 0.05$. This means that the data give evidence against H_o that would happen no more than 5% of the time when H_0 is true.

Statistical Significance If the p-value is as small or smaller than α, we say that the data are statistically significant at level α.

By convention, usually regard $p < 0.05$ as 'statistically significant'

A Recipe for Hypothesis Tests

1. State the null hypothesis and the alternative hypothesis.
   The test is designed to assess the strength of evidence against H_o.

2. Specify the significance level α.
   This states how much evidence we regard as decisive. (Usually 0.05)

3. Calculate the value of the test statistic on which the test will be based.
   This is a statistic that measures how well the data conform to H_o.

4. Find the p-value for the observed data. This is the probability, assuming that H_0 is true, that the test statistic will weigh against H_o at least as strongly as it does for the observed data.
   If the p-value is less than or equal to α, the test result is statistically significant.

5. Draw our conclusion.

The tests discussed next are known as **Z tests** because we are using the assumption that the test statistic is Normally distributed - this is possible because we have stated that the **standard deviation is known**.

They are appropriate for one sample problems when we are testing questions about the mean of quantitative variables

Example One Sample Z test (One sided)

Is the mean age of a first year Stirling University student greater than 19.43 years?

We question a random sample of 100 first year student and find that the observed mean is 19.57. We assume that the ages come from a Normal distribution with known standard deviation 0.97.

Note that this is a one-sided test because we only reject Ho if we find evidence that the mean age is greater than 19.43 years.

The p-value for testing

$$H_o : \mu = 19.43$$
$$H_1 : \mu > 19.43$$

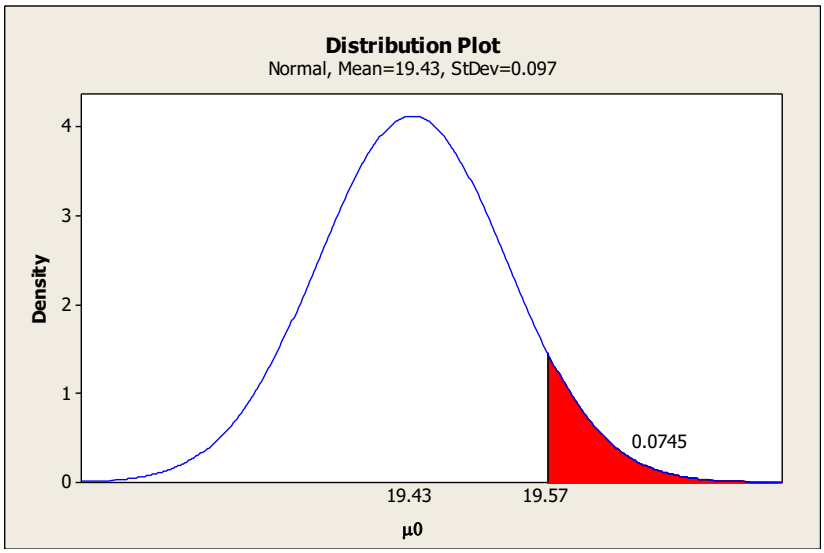
is therefore P($\bar{x}$ > 19.57) calculated assuming that H_0 is true.

When H_0 is true, $\bar{x}$ has a normal distribution (see below for theory) with

$$\mu_{\bar{x}} = \mu = 19.43$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.97}{10} = 0.097$$

The p-value is found by the Normal probability calculation:

This can be illustrated as follows:



Example One Sample Z test (Two sided)

Is the mean age of a first year Stirling University student equal to 19.43 years?

We question a random sample of 100 first year student and find that the observed mean is 19.57.
We assume that the ages come from a Normal distribution with known standard deviation 0.97.

The P-value for testing

$$H_o : \mu = 19.43$$
$$H_1 : \mu \neq 19.43$$

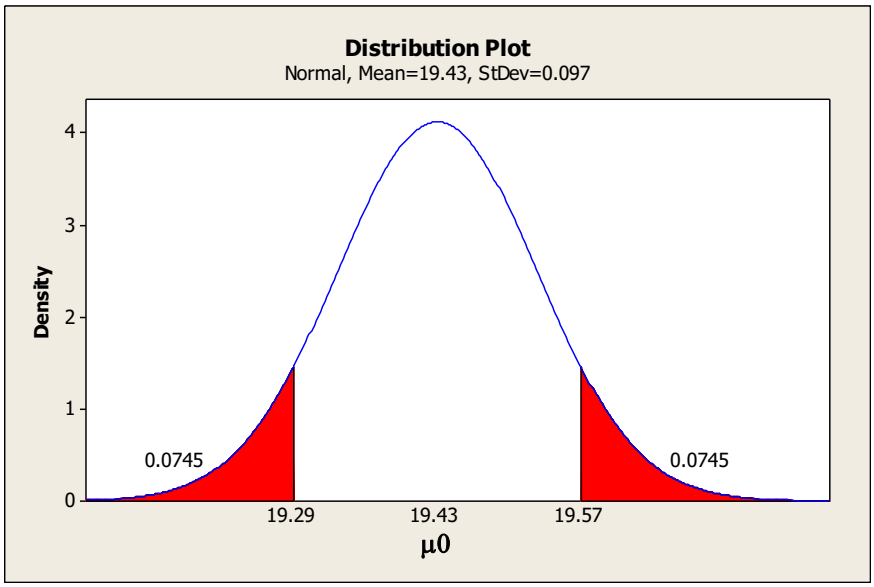
is therefore ' P($\bar{x} \neq 19.57$) '

calculated assuming that H_0 is true. When H_0 is true, $\bar{x}$ has a normal distribution with

$$\mu_{\bar{x}} = \mu = 19.43$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.97}{10} = 0.097$$

The p-value is found by the Normal probability calculation:

This can be illustrated as follows:



Before presenting the concept of confidence intervals in detail, let us consider two general examples of hypothesis tests performed using the Recipe for Hypothesis Tests.

Thus introducing the idea of **Critical Values** and **Rejection Region**.

(1) One Sided Alternative with $\alpha = 0.01$

This implies that there is only a 1 in 100 chance of the finding evidence to reject H_0 if H_0 is true.

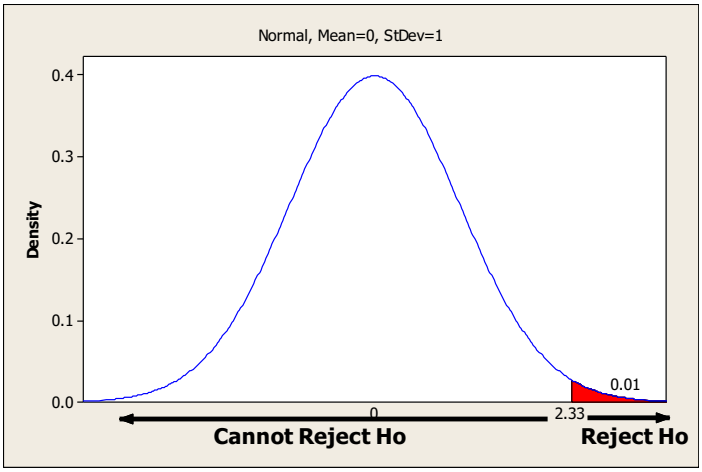
Example Is there an increase in heart rate after exercise?
Let μ = the population mean change in heart rate,
let σ = the population standard deviation and is known.
Sample size = n.

Hypotheses H_0: $\mu = 0$
H_1: $\mu > 0$

Significance Level $\alpha = 0.01$

Test Statistic $Z = \dfrac{\bar{x} - 0}{\sigma/\sqrt{n}}$ ~ N(0,1) under Ho

Rejection Region $P = P(Z > z)$
$= P(Z > 2.33)$ (Standard Normal Tables)
i.e. the critical value is 2.33.



Conclusion If our Observed Test Statistic, z, is greater than 2.33 we have sufficient evidence to reject H_0 in favour of H_1 at the 1% significance level. If our Observed Test Statistic, z, is less than 2.33 we do not have enough evidence to reject H_0.

Verify If we wanted a test at the 5% significance level i.e. $\alpha = 0.05$ the cut-off value (or critical value) would be 1.645.

(2) Two Sided Alternative with $\alpha = 0.01$

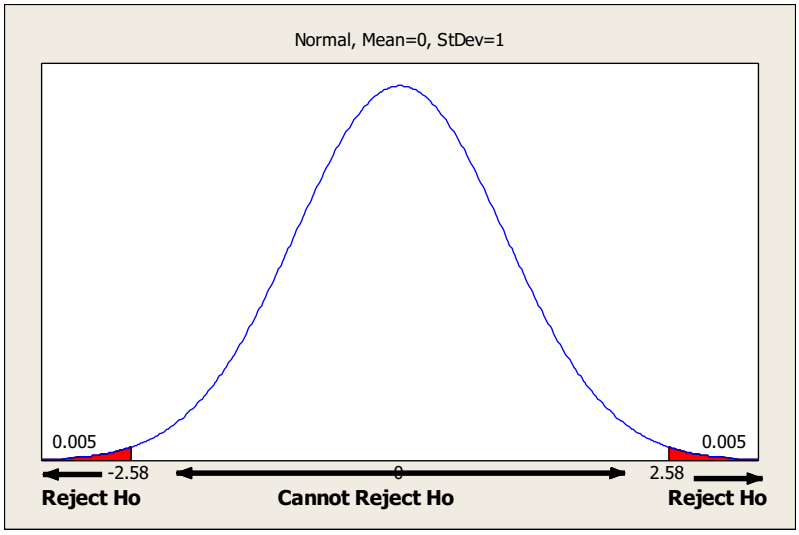
Example Is there a difference in the average birthweight of babies in UK and USA?
Let μ = the difference in the population mean birth weights in the UK and USA, let σ = the 'population standard deviation and is known.'
Sample size = n.

Hypotheses H_0: $\mu = 0$
H_1: $\mu \neq 0$

Significance Level $\alpha = 0.01$

Test Statistic $Z = \dfrac{\bar{x} - 0}{\sigma/\sqrt{n}} \sim N(0,1)$ under Ho

Rejection Region $P = P(Z > z)$
$= P(Z > 2.58)$ (Standard Normal Tables)
i.e. the critical value is 2.58



Conclusion If our Observed Test Statistic, z, is greater than 2.58 or less than -2.58 we have sufficient evidence to reject H_0 in favour of H_1 at the 1% significance level. If our Observed Test Statistic, z, is between -2.58 and 2.58 we do not have enough evidence to reject H_0.

Verify If we wanted a test at the 5% significance level i.e. $\alpha = 0.05$ the cut-off value (or critical value) would be 1.96.

Interjection : Additional Information

We discussed in the previous lecture that when we are performing hypothesis tests we are, in fact, investigating properties of a sampling distribution of a statistic. We found that as with distributions of any random variable as well as requiring knowledge of the centre of the distribution we also need a measure of the spread.

In the above examples, we have used the measure of the spread of the sampling distribution of $\bar{x}$.

3.1.5 The Distribution of a Sample Mean

Let the sample mean $\bar{x}$ be an estimate of the population mean μ. Its distribution is determined by the design to produce the data, the sample size n and the population distribution.

Let our data consist of observations of n independent random variables $X_1, X_2, \ldots\ldots X_n$ where X_i is a measure on a single experimental unit drawn at random from the population and therefore has the same distribution as the population.

The sample mean is $$\bar{x} = \frac{1}{n}(X_1 + X_2 + \ldots\ldots + X_n)$$

If the population has mean μ then μ_i is the mean of each observation X_i. Therefore

$$\mu_{\bar{x}} = \frac{1}{n}\left(\mu_{X_1} + \mu_{X_2} + \ldots\ldots + \mu_{X_n}\right)$$
$$= \frac{1}{n} n\mu = \mu$$

i.e.the mean of $\bar{x}$ is the same as the population mean.

This means that the sample mean $\bar{x}$ is an unbiased estimator of the unknown population mean μ.

Since the observations are independent, the variance of $\bar{x}$ is

$$\sigma^2_{\bar{x}} = \left(\frac{1}{n}\right)^2\left(\sigma^2_{X_1} + \sigma^2_{X_2} + \ldots\ldots + \sigma^2_{X_n}\right)$$
$$= \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

i.e. the mean of several variables is less variable than a single observation. Note that the standard deviation of $\bar{x}$ decreases in proportion to the square root of the sample size.

Finally, we stated that the distribution of the sample mean depends on the population distribution. In particular, if the population distribution is normal, then so is the distribution of the sample mean.

i.e. If the population has a N(μ, σ^2) distribution, then the sample mean of n independent observations has the $N\left(\mu, \frac{\sigma^2}{n}\right)$ distribution

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

3.1.6Confidence Intervals

Confidence Intervals are closely related to the Two-Sided Hypothesis Test.

When we looked at the Normal distribution we found that

- there is probability of 0.95 that a observation from a Normal distribution will lie with approximately two standard deviations of the mean.
- alternatively the mean is within 2 standard deviations of an observation in 95% of the time.

We can now think of this in terms of $\bar{x}$ and μ.

i.e. there is probability of 0.95 that the sample mean, $\bar{x}$, from a Normal distribution will lie with approximately two standard deviations of the population mean μ.

i.e. approximately 95% of all samples will capture the true population mean μ, in the interval from

$$\bar{x} - 2sd \quad to \quad \bar{x} + 2sd$$

This is the general form of a confidence interval i.e.

	estimate	±	margin of error
or	estimate	±	(the variability of the estimate)

and the confidence level shows how confident we are that the method will catch the true population value.

Any confidence interval has two aspects

1. An interval calculated from the data
2. A confidence level giving the probability that the method produces an interval that covers the parameter.

Definition Confidence Interval
A level C confidence interval for a parameter θ is an interval computed from sample data by a method that has a probability C of producing an interval containing the true value of θ.

3.1.7 Critical Value

The number z* with probability p lying to its right under the standard normal curve is called the upper p critical value of the standard normal distribution.

It is useful to remember the following:

Confidence Level	z*
90%	1.645
95%	1.960
99%	2.576

Using the above we can now develop a confidence interval for a one sample mean.

A 95% Confidence Interval for μ, the population mean, (sample size of n and known standard deviation σ); known as a Z interval

$$\bar{x} \pm 1.96\sqrt{\frac{\sigma^2}{n}}$$

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Note

1. We have seen that the hypothesis test and confidence interval for the population mean use the same information but present it in different ways.

   This is the case whichever population parameter we are interested in.

2. The hypothesis test presents a yes / no answer i.e. it is either statistically significant or not.

3. The confidence interval presents a range of values within which have, for example, 95% confidence that the true population parameter will lie.

4. Hypothesis tests are used to test a claim about the population parameter while confidence intervals provide an estimate of the population parameter.

Chapter 4 Categorical Data

4.1 Using the χ^2 test with Categorical Data

We will return to consideration of the χ^2 test. One of the commonest applications of the χ^2 test is in the analysis of so called contingency tables. This is best illustrated with an example.

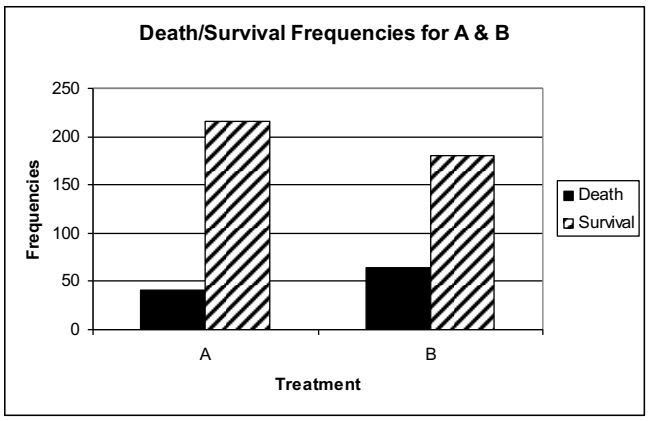
In a clinical trial to assess the value of a new method of treatment (A) in comparison with the old method (B), patients were divided at random into two groups. Of 257 patients treated by method A, 41 died; of 244 patients treated by method B, 64 died.

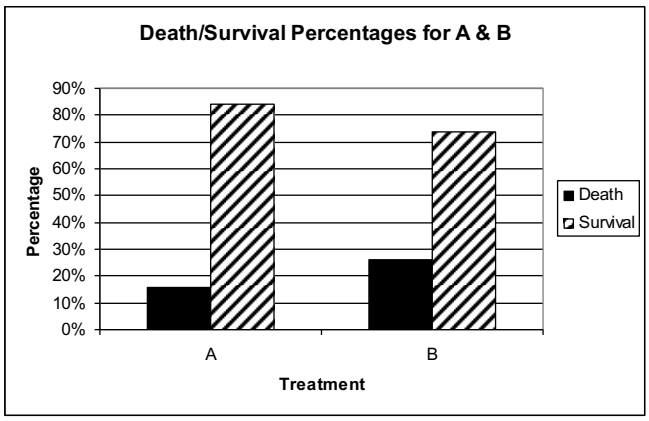
The data are displayed in the table below:

Treatment	Death	Survival	Total
A	41	216	257
B	64	180	244
Total	105	396	501

This is often called a **fourfold table** or **2x2 contingency table**. The total frequency, 501 in this example, is shown at the bottom right of the table. This total frequency or grand total is split into two different dichotomies represented by the two horizontal rows of the table and the two vertical columns.

In this example, the rows represent the two treatments and the columns represent the two outcomes of treatment. There are thus 2x2 = 4 combinations of the row and column categories, and the corresponding frequencies occupy the four inner cells in the body of the table. The total frequencies for the two row categories and those for the two column categories are shown at the right and at the foot, and are called the marginal totals.





Hypotheses: H_0: Variables are not associated / independent i.e. equal proportions
H_1: Variables are related i.e. unequal proportions.

We are concerned here with possible differences between the fatality rates for the two treatments. Given marginal totals in the table above we can calculate what numbers would have had to be observed in the body of the table to make the fatality rates for A and B equal.

In the top left cell, for example, this expected number is

$$\frac{105 \times 257}{501} = 53.862$$

since the overall fatality rate is 105/501 and there are 257 individuals with A. Similar expected numbers can be obtained for each of the four inner cells, and are shown in the table below, where the observed and expected number are distinguished by the letters O and E. The expected numbers are not integers and have been rounded off to three decimals. Clearly one could not observe 53.862 individuals in a particular cell. These expected numbers should be thought of as expectations or mean values over a large number of possible tables with the same marginal totals as those observed, when the null hypothesis is true.

Note that the values of E sum, over both rows and columns, to the observed marginal totals. It follows that the discrepancies, measured by the differences O-E, add to zero along the rows and columns; in other words, the four discrepancies are numerically the same (12.862 in this example), two being positive and two being negative.

Observed Frequencies (O)

Treatment	Death	Survival	Total
A	41	216	257
B	64	180	244
Total	105	396	501

Row Totals

Column Totals

Overall Total

Expected Frequencies (E)

Treatment	Death	Survival	Total
A			257
B			244
Total	105	396	501

In a rough sense, the greater the discrepancies, the more evidence we have against the null hypothesis it would therefore seem reasonable to base a significance test on these discrepancies. It also seems reasonable to take account of the absolute size of frequencies: a discrepancy of 5 is much more important if E=5 than if E=100.

It turns out to be appropriate to calculate the following index (the Test Statistic):

$$X^2 = \sum \frac{(O-E)^2}{E}$$

the summation being over the four inner cells of the table. The contributions to χ^2 from the four cells are shown in the previous table. The total (Observed Test Statistic) is

$$X^2 =$$

We need to say something here about the number of degrees of freedom in the contingency table. The number of degrees of freedom is the number of cells in the contingency table that can be set independently, given the marginal totals. In a 2x2 contingency table, only one cell is independent, therefore we have only one degree of freedom. In general, in a contingency table with r rows and c columns, there are (r-1) x (c-1) degrees of freedom.

Returning now to our 2x2 table we see that, on the null hypothesis, χ^2 follows the $\chi^2(1)$ distribution, the approximation improving as the expected numbers get larger. Reference to the χ^2 table shows that the observed value of 7.978 is beyond the 0.05 point of the $\chi^2(1)$ distribution, and the difference between the two fatality rates is therefore significant at the 5 per cent level.

i.e. the Rejection Region at the 5% significance level is calculated as

$P(\chi^2(1) > c) = 0.05$ which is $P(\chi^2(1) > 3.84)$

so Reject H_0 in favour of H_1 at 5% significance level if the Observed Test Statistic χ^2 is greater than 3.84

IMPORTANT

It is important to remember that the X^2 index can only be calculated from a 2x2 table in which the entries are frequencies. A common error is to use it for a table in which the entries are mean values of a certain variable; this practice is completely erroneous.

The method can be easily generalised to a contingency table with any number of rows and columns. However, the assumption that X^2 approximates to χ^2 is not valid if the cell frequencies are too small.

A useful rule is as follows:

If df =1, then no cell can have an expected frequency of less than 5.

If df > 1, then no more than 20% of the cells can have an expected frequency of less than 5, and no cell an expected frequency of less than 1.

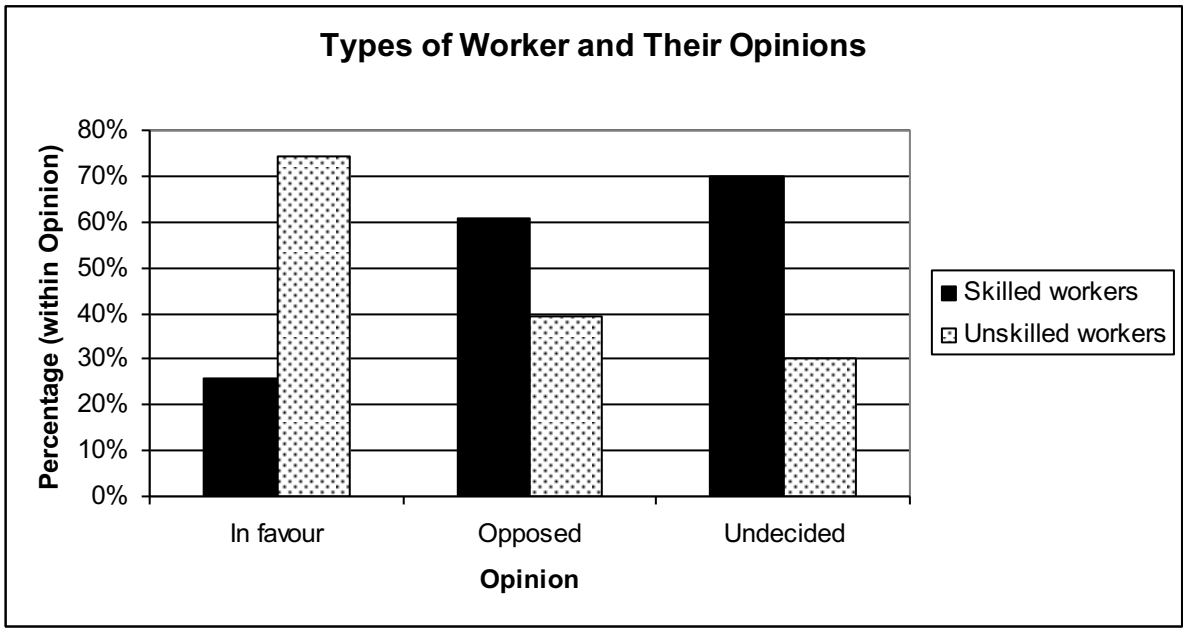
Example

In a survey concerned with changes in working procedures the following table was produced:

	Opinion on changes in working procedures		
	In favour	Opposed	Undecided
Skilled workers	19	31	35
Unskilled workers	55	20	15

Calculate the proportions of each Opinion group that are Skilled Workers. Does this suggest a relationship between opinion and type of worker?

Test the hypotheses that the opinion on working procedures is independent of whether workers are classified as skilled or unskilled.



In general, two methods of improvement to the χ^2 test are widely used; the application of a continuity correction and the calculation of exact probabilities.

4.1.1 Continuity Correction for Fourfold Tables

This method was described by F Yates and is often called the Yates Correction. If the number of degrees of freedom = 1, a small correction factor is applied to account for the fact that the distribution is continuous, whereas the observed frequencies are discrete.

Thus, for df=1, X^2 is calculated from

$$X^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

The continuity corrected version of the alternative formula is

$$X^2 = \frac{(|ad - bc| - N/2)^2 N}{r_1 r_2 s_1 s_2}$$

4.1.2 The Exact Test for Fourfold Tables

Even with the continuity correction there will be some doubt about the adequacy of the χ^2 approximation when the frequencies are particularly small. An exact test was suggested almost simultaneously in the mid-1930's by Fisher, Irwin and Yates. It consists in calculating the exact probabilities in the distribution described in the Previous subsection. The probability of a table with frequencies:

a	b	r_1
c	d	r_2
s_1	s_2	N

is given by the formula

$$\frac{r_1! r_2! s_1! s_2!}{N! a! b! c! d!}$$

Given any observed table, the probabilities of all tables with the same marginal totals can be calculated, and the p value for the significance test calculated by summation.

4.1.4 Goodness of Fit Test: Testing Whether a Distribution is Normal

A useful use of the χ^2 test is to check whether the data have a Normal Distribution or not.

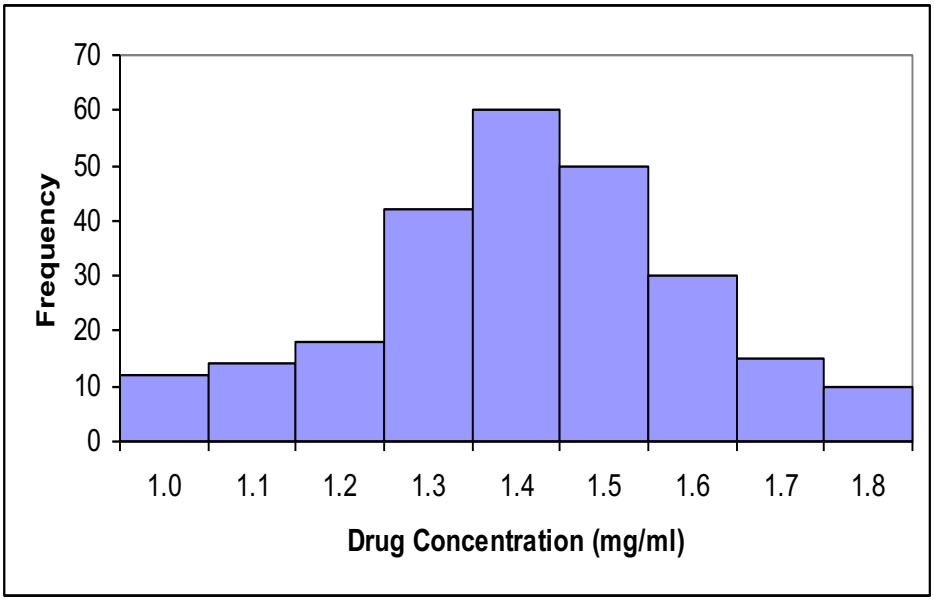
Example Drug levels are measured in a random group of 251 subjects receiving a particular drug. The data are as follows:

mg/ml	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
Frequency = No of Subjects	12	14	18	42	60	50	30	15	10

Question Do these data follow a Normal Distribution?

Solution

A. Informal



B. Formal

Hypotheses H_0: Data follow a Normal Distribution

H_1: Data follow some other Distribution

Significance Level 0.05

Test Statistic $X^2 = \sum_{i=1}^{C} \frac{(O_i - E_i)^2}{E_i}$ *where* $C = No\ of\ Classes$

Observed Test Statistic

(i) Calculate the Expected Values

$\bar{x} = 1.4$ $s = 0.2$ $n = 251$

X	Z	boundary = b	P(lb<Z<ub)	Expected	Observed
1.0	-2.0				12
1.1	-1.5				14
1.2	-1.0				18
1.3	-0.5				42
1.4	0				60
1.5	0.5				50
1.6	1.0				30
1.7	1.5				15
1.8	2.0				10

Method

Col 1 -> Col 2 $Z = \frac{X - \bar{x}}{s}$

Col 2 -> Col 3 Choose Midpoints between Z values, b = boundary

Col 3 -> Col 4 P(lower b < Z < upper b) = P(Z < upper b) - P(Z< lower b) (from N(0,1) tables)

Col 4 -> Col 5 n x (Probability from Col 4)

Col 5 Observed Frequency in Interval

(ii) Calculate the Observed Test Statistic

$$X^2 = \frac{(12-7)^2}{7} + \frac{(12-16.5)^2}{16.5} + \ldots\ldots + \frac{(10-7)^2}{7} = 13.6$$

(iii) Rejection Region Significance Level 0.05; One Sided

$\chi^2(6, 0.05) = 12.59$

where df = no of classes - 1 - no of parameters estimated
= 9 - 1 - 2 = 6

So Reject H_0 in favour of H_1 at 5% significance level only if $X^2 > 12.59$.

$X^2 = 13.6$ so reject H_0 and conclude that the data does not follow a Normal Distribution.

4.2 Inferences about a Population Proportion

4.2.1 Situations, Definitions and Objectives

Suppose we were to select a random sample of n elements from a large population and that we want to estimate or test hypotheses about the proportion θ of elements that possess a particular characteristic.

If x of n elements in the sample possess the characteristic, then the best estimate of θ is the population proportion, x/n.

i.e. $\hat{\theta} = \frac{x}{n}$

Example We want to estimate the proportion of the Electorate of the UK who will vote for the Conservative Party. A random sample of 1100 voters were questioned on April 1st. Of this sample 198 said that they intended to vote Conservative. So

4.2.2 The Sampling Distribution of the Sample Proportion

We will use the sample proportion $\hat{\theta}$ to estimate and to test hypotheses about a multinomial or binomial population parameter.

How far from the population proportion θ is the sample proportion likely to be? e.g. is it likely that the sample proportion voting for the Tories (0.198) above will deviate from the population proportion by 0.2?

To solve this, we need to know the properties of the sampling distribution of $\hat{\theta}$.

We know that the number of x of successes in a binomial experiment is approximately normally distributed when the number n of trials is large.

So the sample proportion $\hat{\theta}$ = x / n is also normally distributed when n is large. The mean of the sampling distribution of $\hat{\theta}$ is θ and the standard error is SE($\hat{\theta}$) = $\sqrt{\theta(1-\theta)/n}$

The Sampling Distribution of $\hat{\theta}$ is shown in Figure 1 below and the mean and standard error are also given below.

$$\textit{Mean of } \hat{\theta} \;:\; \theta$$

$$\textit{Standard Error of } \hat{\theta} \;:\; \sqrt{\frac{\theta(1-\theta)}{n}}$$

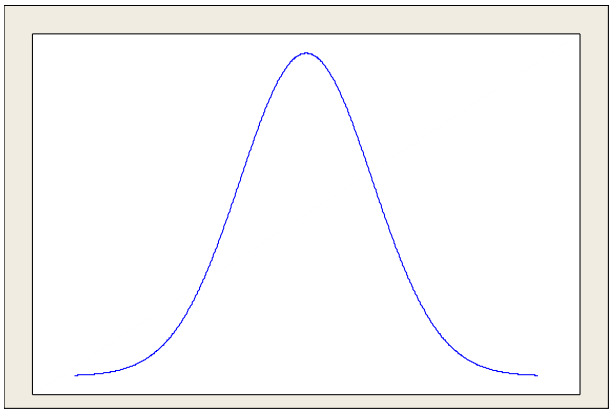


Figure 1 The Sampling Distribution of Sample Proportion $\hat{\theta}$

As with any Normal Distribution, the probability that a sample proportion will fall within 1.96 standard errors of θ is 0.95 and the probability that it will deviate from θ more than 1.96 SE($\hat{\theta}$) is 0.05.

To find the probability that a sample proportion will deviate from θ by some specified amount, we need to express the deviation in units of SE($\hat{\theta}$) i.e. in units of the standard Normal variable, calculate Z score

$$Z = \frac{\hat{\theta} - \theta}{SE\left(\hat{\theta}\right)}$$

Example Find the approximate probability that a sample proportion, based on a random sample of 1486 people, will deviate from the population proportion by 0.02 or more (i.e. minimum error of 2%)

In this example, the investigator found that 1139 of the 1486 people read a particular newspaper.

Solution

4.2.3A Confidence Interval and Test for a Population Proportion θ

The Z statistic is used to find a Confidence Interval and a Test of a hypothesis for a population proportion θ. A 100(1-α) % confidence interval for θ, based on the standard normal Z statistic

$$Z = \frac{\hat{\theta} - \theta}{SE\left(\hat{\theta}\right)}$$

is given below.

A 100(1-a) % confidence interval for a population proportion θ

$$\hat{\theta} \pm N(0,1;\alpha/2)\sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}}$$

In particular, a 95% Confidence Interval for a population proportion θ is

$$\hat{\theta} \pm 1.96 \sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}}$$

Example The Newspaper Survey continued.

Provide a 95% confidence interval for the proportion θ of people who read the given paper.

Solution

4.2.4A Test Concerning a Population Proportion θ

Null Hypothesis - $H_0 : \theta = \theta_0$

Alternative Hypothesis - $H_1 : \theta \neq \theta_0$; $H_1 : \theta < \theta_0$; $H_1 : \theta > \theta_0$

Significance Level - usually 0.05

Test Statistic - $Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\theta(1-\theta)}{n}}}$

Rejection Region - Given a 0.05 significance level, reject H_o in favour of H_1 when Z is less than -1.96 or greater than 1.96. Otherwise not sufficient evidence to reject H_0 at 5% level.

Assumptions

1. The sampling distribution satisfies the requirements of a binomial experiment.

2. The sampling distribution of x and therefore of $\hat{\theta}$ is normally distributed.
   i.e. n is large enough so that the interval $n\theta \pm 3\sqrt{n\theta(1-\theta)}$ falls within the interval from θ to n.

Example An article in the Wall Street Journal stated that the National Union of Hospital and Health Care Workers won 56 of 80 union representation elections compared with a success rate of 55% for all health care unions during the same period.

Do these data provide sufficient evidence to indicate that the probability θ of an election success for the National Union of hospital Workers was higher than 0.55, the probability for all health unions as a whole?

1. Is n large enough for the test to be valid?

2. State the hypotheses

3. State the Test Statistic

4. Calculate the Observed Test Statistic

5. Calculate the Rejection Region

6. Draw conclusions.

Solution

4.3 Comparing Two Proportions

The problem of comparing proportions from two different populations is identical to the problem of comparing two population means except that the population consists of categorical data.

We select independent random samples from two binomial populations, n_1 elements from population 1 and n_2 elements from population 2. The objective of the sampling is to estimate the difference between the proportions θ_1 and θ_2 of the elements in the two populations that fall into a particular category or to test hypotheses concerning the difference between θ_1 and θ_2

4.3.1 The Sampling Distribution of the Difference between Two Sample Proportions

Since for large samples, both $\hat{\theta}_1$ and $\hat{\theta}_2$ are approximately normally distributed, it follows that the difference $\hat{\theta}_1 - \hat{\theta}_2$ in the sample proportions will be approximately normally distributed with mean and standard error shown below.

Mean and Standard Error of Sampling Distribution of $\hat{\theta}_1 - \hat{\theta}_2$

Mean : $\theta_1 - \theta_2$

Standard Error : $SE(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}}$

4.3.2 A Confidence Interval and Test for the Difference Between Two Population Proportions

The Confidence Interval and the test of an hypothesis for $\theta_1 - \theta_2$ are based on the standard normal Z statistic

$$Z = \frac{\left(\hat{\theta}_1 - \hat{\theta}_2\right) - \left(\theta_1 - \theta_2\right)}{SE(\hat{\theta}_1 - \hat{\theta}_2)}$$

and are given below.

A 100(1 - α)% Confidence Interval for $\theta_1 - \theta_2$

$$\hat{\theta}_1 - \hat{\theta}_2 \pm N(0,1;\alpha/2)\, SE(\hat{\theta}_1 - \hat{\theta}_2)$$

i.e. A 95% Confidence Interval for $\theta_1 - \theta_2$

$$\hat{\theta}_1 - \hat{\theta}_2 \pm 1.96 \times \sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}}$$

Assumptions

1. The samples are independently and randomly drawn from the two populations.

2. The sample sizes are large enough so that the numbers of successes are approximately normally distributed.

3. The values of θ_1 and θ_2 in the formula for the standard error can be approximated by their sample estimates.

Example The following data was collected in the USA on the use of lethal force by the police.

Case Type	1970 - 1972	1973	1974 - 1978
Justified	72	30	173
Not Justified	11	9	59
Unable to Determine	18	7	53
Accidental	10	5	12
Total	111	51	297

Find a 95% confidence interval for the difference in the probability of the unjustified use of lethal force between the time periods 1970-1972 and 1974 - 1978.

Solution

Table : Chisquare Distribution

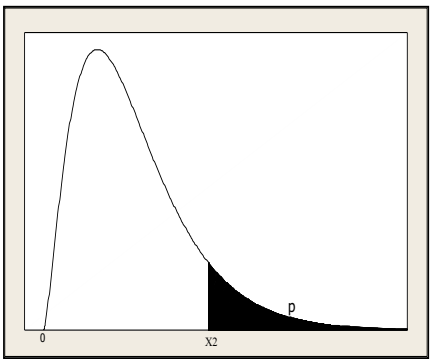


Table o Critical Values [Table entry for the point with probability p lying above it]

	Upper Tail Probability p																
df	0.995	0.99	0.975	0.95	0.9	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	0.00	0.00	0.00	0.00	0.02	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	0.01	0.02	0.05	0.10	0.21	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	0.07	0.11	0.22	0.35	0.58	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	0.21	0.30	0.48	0.71	1.06	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	0.41	0.55	0.83	1.15	1.61	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	0.68	0.87	1.24	1.64	2.20	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	0.99	1.24	1.69	2.17	2.83	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	1.34	1.65	2.18	2.73	3.49	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	1.73	2.09	2.70	3.33	4.17	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	2.16	2.56	3.25	3.94	4.87	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	2.60	3.05	3.82	4.57	5.58	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.73	26.76	28.73	31.26	33.14
12	3.07	3.57	4.40	5.23	6.30	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	3.57	4.11	5.01	5.89	7.04	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	4.07	4.66	5.63	6.57	7.79	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	4.60	5.23	6.26	7.26	8.55	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	5.14	5.81	6.91	7.96	9.31	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	5.70	6.41	7.56	8.67	10.09	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	6.26	7.01	8.23	9.39	10.86	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	6.84	7.63	8.91	10.12	11.65	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	7.43	8.26	9.59	10.85	12.44	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	8.03	8.90	10.28	11.59	13.24	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.77	46.80	49.01
22	8.64	9.54	10.98	12.34	14.04	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	9.26	10.20	11.69	13.09	14.85	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	9.89	10.86	12.40	13.85	15.66	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	10.52	11.52	13.12	14.61	16.47	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	11.16	12.20	13.84	15.38	17.29	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	11.81	12.88	14.57	16.15	18.11	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.65	52.22	55.48	57.86
28	12.46	13.56	15.31	16.93	18.94	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	13.12	14.26	16.05	17.71	19.77	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	13.79	14.95	16.79	18.49	20.60	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	20.71	22.16	24.43	26.51	29.05	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.10
50	27.99	29.71	32.36	34.76	37.69	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	35.53	37.48	40.48	43.19	46.46	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	51.17	53.54	57.15	60.39	64.28	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	67.33	70.06	74.22	77.93	82.36	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

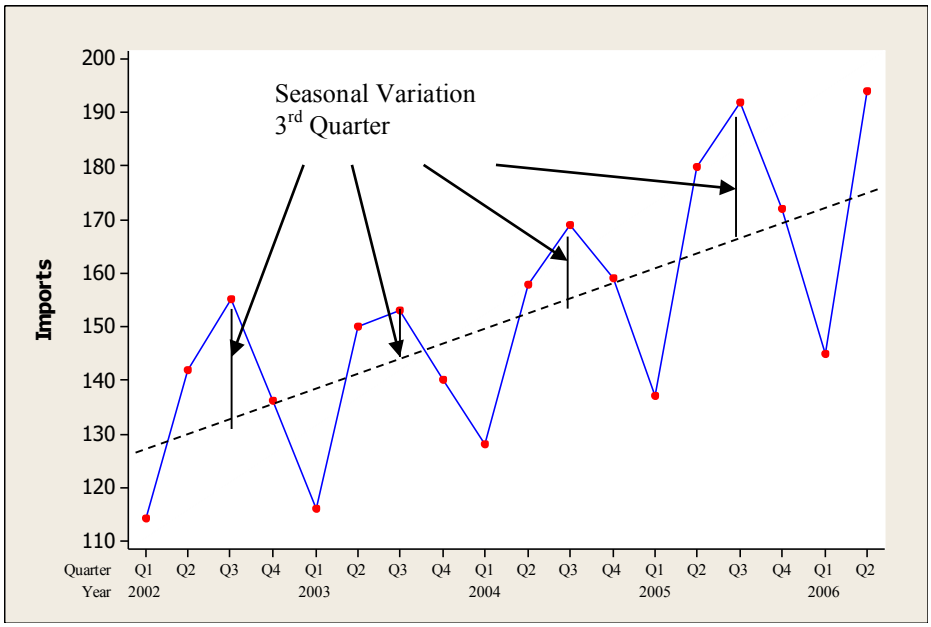
Chapter 5 Time Series

An example of a typical type of time series is:

Year	Quarter	Imports (£000's)
2002	1	114
2002	2	142
2002	3	155
2002	4	136
2003	1	116
2003	2	150
2003	3	153
2003	4	140
2004	1	128
2004	2	158
2004	3	169
2004	4	159
2005	1	137
2005	2	180
2005	3	192
2005	4	172
2006	1	145
2006	2	194

The series of data that tells us how data has been behaving in the past is the time series. It gives us the value of the variable of interest at various points in time - each day, each week, each month, each season, each year over a period of time. However, when you look at a typical time series (see the example above), the data fluctuates so much that it seems unlikely that it can help us very much. The first task we must therefore undertake is to assess what factors might cause this fluctuation.

Let us plot the data from the example above. We also plot what is called the 'trend' line.



If we examine this graph carefully we see the following:

1.

2.

3.

4.

Thus data such as this which fluctuates quite markedly over time may be responding to any or all of the above set of for influences.

5.1.1 Calculating The Trend

In the previous example the trend line was simply a straight-line drawn through the data to show that the data moves in this way. However, we require an objective technique to calculate the trend figures i.e. we cannot just guess them!

The following are examples of possible time series trend patterns that may be underlying the fluctuations.

There are several methods of calculating trend but we will only introduce one in this course - the method of moving averages. A moving average is simply an arithmetic mean. We select a group of numbers at the start of the series and average them to obtain our first trend figure.

In the example below: 5 numbers - 5point moving average or because time is measured in years in this example 5 year moving average.

e.g. First trend figure is (1.3 + 1.1 + 1.2 + 1.4 + 2.1) / 5 = 1.42
This is placed opposite the centre of the group of five numbers i.e. opposite year 3.

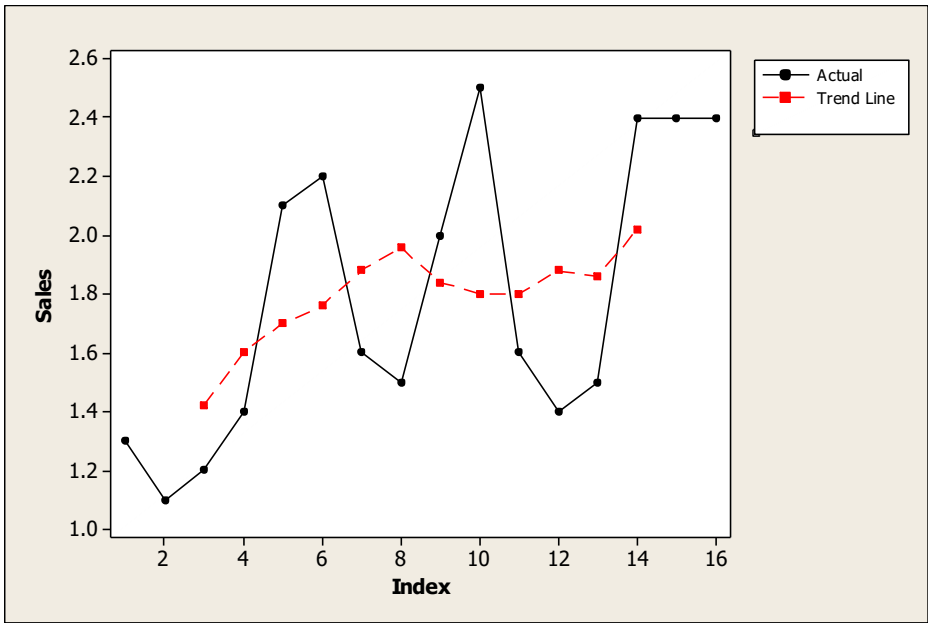
Second trend figure is calculated by dropping the first number (1.3) and including the next number in the series (2.2).

e.g. Second trend figure is (1.1 + 1.2 + 1.4 + 2.1 + 2.2) / 5 = 1.6
This is placed opposite the centre of the group of five numbers i.e. opposite year 4.

We continue in this way until we have used all the data e.g. in this example the last trend figure is opposite year 14 (this figure averages the sales on years 12,13, 14, 15 and 16). i.e. there are no trend figures for the first 2 years or the last 2 years.

Year	Sales(£000's)	5 Year Total	5 Year Moving Average
1	1.3		
2	1.1		
3	1.2		
4	1.4		
5	2.1		
6	2.2		
7	1.6		
8	1.5		
9	2.0		
10	2.5		
11	1.6		
12	1.4		
13	1.5		
14	2.4		
15	2.4		
16	2.4		

The following shows the original data and the 'trend' line.



The trend although smoother still fluctuates. There is no clear way in which the trend is moving.

We choose a five-year moving average rather than a three or seven year one because.....

Moving averages are calculated in the same way simply varying the number of points as appropriate. As with the first example, much data is collected either monthly or quarterly. We saw that often peaks and troughs occur in the same quarter each year.

5.2 Calculating The Trend for a Quarterly Series

Calculating a trend for a quarterly series does not involve any new ideas but it does raise a practical problem. When looking at quarterly data the correct moving average to use is the four quarterly or 4 point moving average. Notice, using the rules above, when we calculate the 4 quarterly moving average and place it opposite the mid-point of the group, the first figure is placed between the second and third figures. Thus we cannot relate our figures of trend to any particular quarter i.e. we cannot make any comparison between the observed figures and the trend.

To overcome this problem, we use a technique called centring and calculate a Centred Moving Average. This involves a further averaging of the 4 quarter moving average i.e.

First Centred Moving Average Figure = (First 4 Quarter Moving Average Figure
+ Second 4 Quarter Moving Average Figure) / 2

This is placed opposite quarter 3

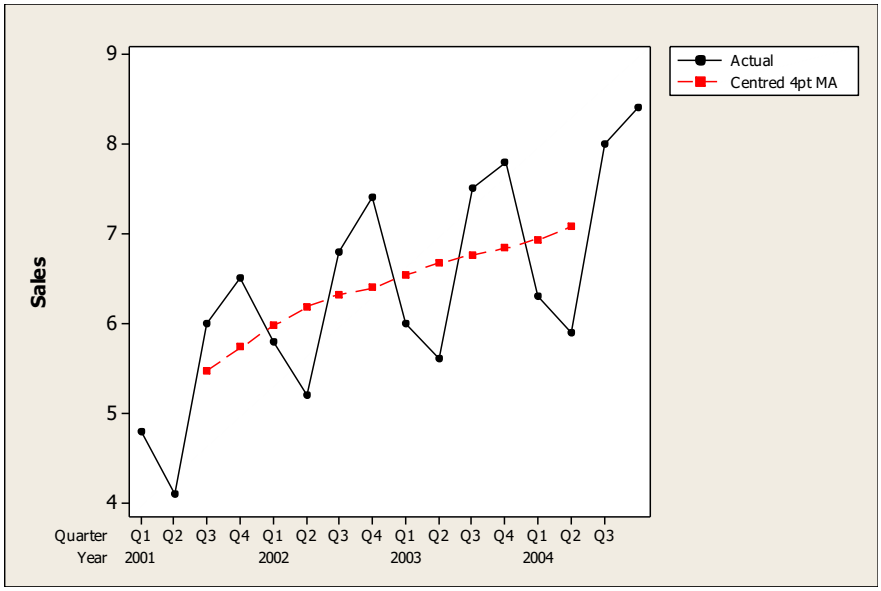
Second Centred Moving Average Figure = (Second 4 Quarter Moving Average Figure
+ Third 4 Quarter Moving Average Figure) / 2

This is placed opposite year 4.

We continue in the same manner for all the data. This is illustrated below. Note that there are no trend figures for the first two quarters or the last two quarters.

Year	Quarter	Sales (£000's)	4 Quarter Moving Average	Centred 4 point Moving Average
1	1	4.8		
	2	4.1		
	3	6.0		
	4	6.5		
2	1	5.8		
	2	5.2		
	3	6.8		
	4	7.4		
3	1	6.0		
	2	5.6		
	3	7.5		
	4	7.8		
4	1	6.3		
	2	5.9		
	3	8.0		
	4	8.4		

The following plot shows both the original data and the 'trend' line.



We have looked at one of the possible methods of detecting any underlying trend within time series data. We now want to remove any seasonal elements within a time series. This is called deseasonalisation.

5.3 Calculating Seasonal Variation

We have defined seasonal variation as an upswing and downswing in the value of the data. If we want to measure the magnitude of these fluctuations, we must have a reference point from which to measure.

It seems logical that we should measure the magnitude as the deviation from our calculated trend figure. So seasonal variation is now defined as a swing around the trend line.

The variation in any particular quarter is therefore:

Deviation = Original Data - Calculated Trend

Performing this calculation on the data above, we get

Year	Quarter	Sales (£1000's)	Centred Moving Average	Deviation from trend i.e. TREND
1	1	4.8		
	2	4.1		
	3	6.0	5.475	
	4	6.5	5.738	
2	1	5.8	5.975	
	2	5.2	6.188	
	3	6.8	6.325	
	4	7.4	6.400	
3	1	6.0	6.538	
	2	5.6	6.675	
	3	7.5	6.763	
	4	7.8	6.838	
4	1	6.3	6.938	
	2	5.9	7.075	
	3	8.0		
	4	8.4		

The deviations in the last column show that for a given quarter seasonal and other influences have caused the sales to vary from the trend by that amount. The deviations have been caused by both seasonal and random influences.

We cannot separate these two influences. However, we assume that seasonal influences always operate in the same direction, random influences can either raise or lower figures. If, we take a long enough series and take the average deviation for any particular quarter, the random influences will tend to offset one another and we will be left with purely seasonal variation. We will call this average seasonal variation or quarterly seasonal variation.

The easiest way to set about calculating the quarterly seasonal variation is to set the data in a table as follows:

Year	Quarter				
	1	2	3	4	
1					
2					
3					
4					
Total					
Average					
Adjustment					
Seasonal Variation					

NOTES

1. The total of the average deviations should be zero. However, the total is -0.042, and since we know that it should be zero an adjustment is required. We do not know where the differences come from so we adjust each quarterly average by -0.0105. In fact, because the total was below zero we add 0.0105 to each quarterly value (i.e. subtract -0.0105).

2. Finally we list the quarterly seasonal variations to the same number of decimal places as the original data.

i.e. the Quarterly Seasonal Variations are

Q1	-0.4
Q2	-1.1
Q3	0.6
Q4	0.9

Thus we can conclude from these results that for the first and second quarter of the year sales fall below the trend and in the third and fourth quarter sales are above the trend (i.e. in the fourth quarter sales are £900 (0.9 x £1000) above the trend).

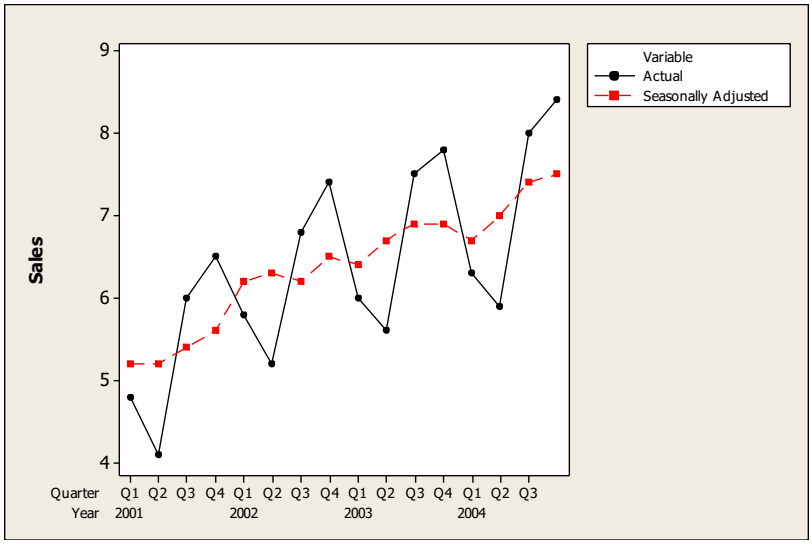
5.3.1 The Deseasonalised Time Series

While it is useful to companies, government bodies and individual to know the seasonal variation, this knowledge alone does not aid comparisons or assessments over time. i.e. the constant fluctuations still tend to hide the underlying behaviour. Since in most cases we will require to make comparisons over time it is useful to calculate deseasonalised or seasonally adjusted data i.e. data with seasonal variation eliminated.

This is simple to do i.e. if seasonal variation is +10 then we are saying that the data is 10 points above the trend because of seasonal influences and therefore we should reduce our figures by 10 to eliminate these influences.

Year	Quarter	Sales (£1000's)	Quarterly Seasonal Variation	Seasonally Adjusted Sales
1	1	4.8	-0.4	
1	2	4.1	-1.1	
1	3	6.0	0.6	
1	4	6.5	0.9	
2	1	5.8	-0.4	
2	2	5.2	-1.1	
2	3	6.8	0.6	
2	4	7.4	0.9	
3	1	6.0	-0.4	
3	2	5.6	-1.1	
3	3	7.5	0.6	
3	4	7.8	0.9	
4	1	6.3	-0.4	
4	2	5.9	-1.1	
4	3	8.0	0.6	
4	4	8.4	0.9	

Plotting both the Actual data and the Seasonally Adjusted gives:



Example Use the Time Series Analysis to examine the Import Data presented at the start of this Chapter

Step 1 Calculate the Moving Averages & the Deviation from Trend = Original - MA

Year	Quarter	Imports	4Pt MA	Centred 4Pt MA	Deviation from Trend
2002	1	114			
2002	2	142			
2002	3	155			
2002	4	136			
2003	1	116			
2003	2	150			
2003	3	153			
2003	4	140			
2004	1	128			
2004	2	158			
2004	3	169			
2004	4	159			
2005	1	137			
2005	2	180			
2005	3	192			
2005	4	172			
2006	1	145			
2006	2	194			

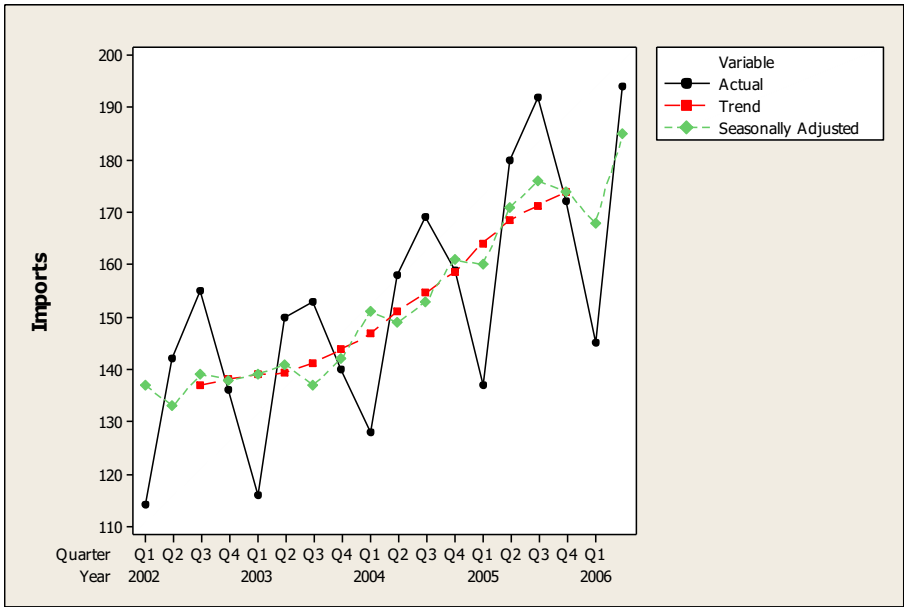
Step 2 Calculate the Quarterly Seasonal Variation

Year	Quarter				
	1	2	3	4	
2002 2003 2004 2005					Sums
Total					
Average					
Adjust Average- Adjust					=Sum/4
Quarterly Seasonal Variation					

Step 3 Calculate the Seasonally Adjusted Series

N.B. Deseasonalised / Seasonally Adjusted Data = Actual - Quarterly Seasonal Variation

Year	Quarter	Imports	Quarterly Seasonal Variation	Deseasonalised Series
2002	1	114		
	2	142		
	3	155		
	4	136		
2003	1	116		
	2	150		
	3	153		
	4	140		
2004	1	128		
	2	158		
	3	169		
	4	159		
2005	1	137		
	2	180		
	3	192		
	4	172		
2006	1	145		
	2	194		



Comments about the Analysis:

5.3.2 Why Residuals are Important

As mentioned before a graph of both the trend line and deseasonalised series tells us about the residual influences. But why are they important?

Time Series Analysis is often used in the financial and economy sector and the aim is frequently to make some type of forecast for future planning. However, any forecast has to be as accurate as possible or any planning based on it is a waste of time. We have stressed before that by their nature residual influences cannot be included in any calculations so when they do have a marked influence the forecast will be upset. In the example above we found that the influences of residuals was small so in this case it is likely that any forecast will not be upset.

However, in other cases, we may find the residual influences have a greater effect. It will be useful if we can obtain a quantitative measure of the residual variation.

We how discussed previously that

Original data = Trend + Seasonal + Residual

Therefore,

Residual = Original - Trend - Seasonal

Performing these calculations for the most recent example, we get

Year	Quarter	Imports	Trend	SV	Residual	
					Absolute	% of Imports
2002	1	114		-23		
2002	2	142		9		
2002	3	155	137.000	16		
2002	4	136	138.250	-2		
2003	1	116	139.000	-23		
2003	2	150	139.250	9		
2003	3	153	141.250	16		
2003	4	140	143.750	-2		
2004	1	128	146.750	-23		
2004	2	158	151.125	9		
2004	3	169	154.625	16		
2004	4	159	158.500	-2		
2005	1	137	164.125	-23		
2005	2	180	168.625	9		
2005	3	192	171.250	16		
2005	4	172	174.000	-2		
2006	1	145		-23		
2006	2	194		9		

Comment:

5.3.3 What type of Model should we use?

In our previous analysis, we assumed:

Actual Data = Trend + Seasonal Variation + Residuals + Cyclical

which simplified to

Actual Data = Trend + Seasonal Variation

This is called an ADDITIVE model i.e. it assumes there is a constant absolute difference between the actual data and the trend and we call this difference the seasonal variation.

However, this is often not a reasonable assumption i.e. it does not seem reasonable to adjust every (for example) fourth quarter by adding or subtracting a constant amount. It would seem much better to adjust every corresponding quarter by a constant percentage. This is called a MULTIPLICATIVE model: (in the simplified form)

Actual Data = Trend x Seasonal Variation

In this case the seasonal variation is calculated using the ratio of actual / trend values. We will call this the seasonal indices and will calculate the quarterly seasonal indices as the average of the seasonal indices for each quarter.

Example Use the Time Series Analysis to examine the Import Data presented at the start of this Chapter: ASSUMING A MULTIPLICATIVE MODEL

Step 1 Calculate the Moving Averages & the Seasonal Variation = Original / Trend

Year	Quarter	Imports	4Pt MA	Centred 4Pt MA	$\frac{Actual}{Trend}$
2002	1	114			
2002	2	142	136.75		
2002	3	155	137.25	137.000	
2002	4	136	139.25	138.250	
2003	1	116	138.75	139.000	
2003	2	150	139.75	139.250	
2003	3	153	142.75	141.250	
2003	4	140	144.75	143.750	
2004	1	128	148.75	146.750	
2004	2	158	153.50	151.125	
2004	3	169	155.75	154.625	
2004	4	159	161.25	158.500	
2005	1	137	167.00	164.125	
2005	2	180	170.25	168.625	
2005	3	192	172.25	171.250	
2005	4	172	175.75	174.000	
2006	1	145			
2006	2	194			

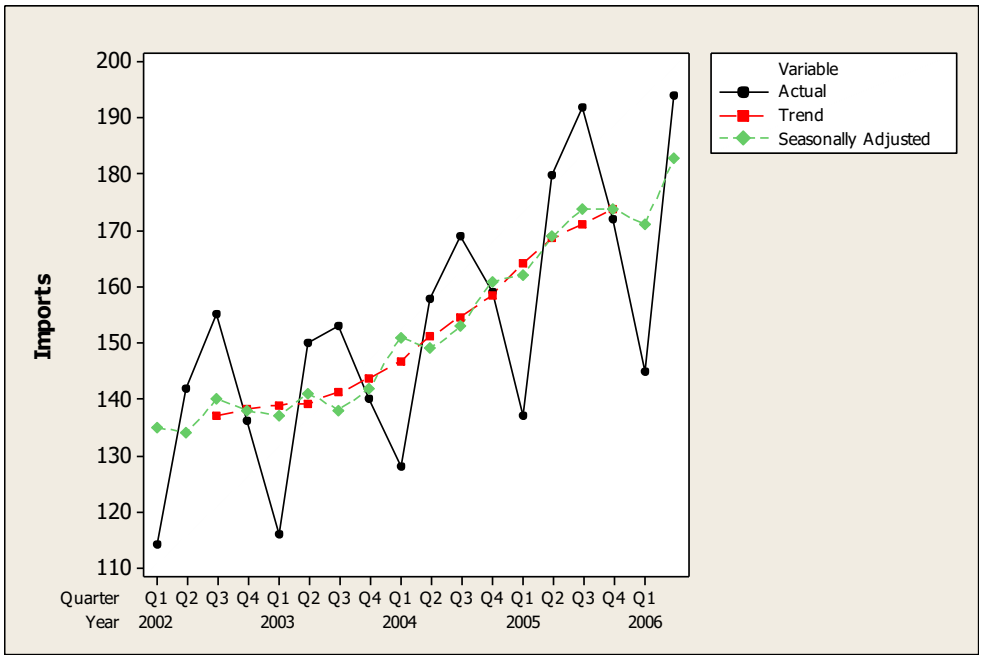
Step 2 Calculate the Quarterly Seasonal Indices

Year	Quarter				
	1	2	3	4	
2002 2003 2004 2005					Sums
Total					
Average					
Adjust Average / Adjust					= Sum/4 =4.000
Quarterly Seasonal Indices					4

Step 3 Calculate the Seasonally Adjusted Series

N.B. Deseasonalised / Seasonally Adjusted Data = Actual / Quarterly Seasonal Index

Year	Quarter	Imports	Quarterly Seasonal Variation	Deseasonalised Series
2002	1	114		
	2	142		
	3	155		
	4	136		
2003	1	116		
	2	150		
	3	153		
	4	140		
2004	1	128		
	2	158		
	3	169		
	4	159		
2005	1	137		
	2	180		
	3	192		
	4	172		
2006	1	145		
	2	194		


Imports
200
190
180
170
160
150
140
130
120
110
Quarter Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1
Year 2002 2003 2004 2005 2006
Variable
Actual
Trend
Seasonally Adjusted

Residuals: Multiplicative Model R = A/(TS)

Year	Quarter	Imports	Trend	Quarterly Index	Residual Absolute
2002	1	114			
2002	2	142			
2002	3	155	137.000	1.106	1.023
2002	4	136	138.250	0.986	0.998
2003	1	116	139.000	0.846	0.986
2003	2	150	139.250	1.062	1.014
2003	3	153	141.250	1.106	0.979
2003	4	140	143.750	0.986	0.988
2004	1	128	146.750	0.846	1.031
2004	2	158	151.125	1.062	0.984
2004	3	169	154.625	1.106	0.988
2004	4	159	158.500	0.986	1.017
2005	1	137	164.125	0.846	0.987
2005	2	180	168.625	1.062	1.005
2005	3	192	171.250	1.106	1.014
2005	4	172	174.000	0.986	1.003
2006	1	145			
2006	2	194			

With this Multiplicative Model, R is assumed to be 1. Therefore, we want the average R to be approximately 1.

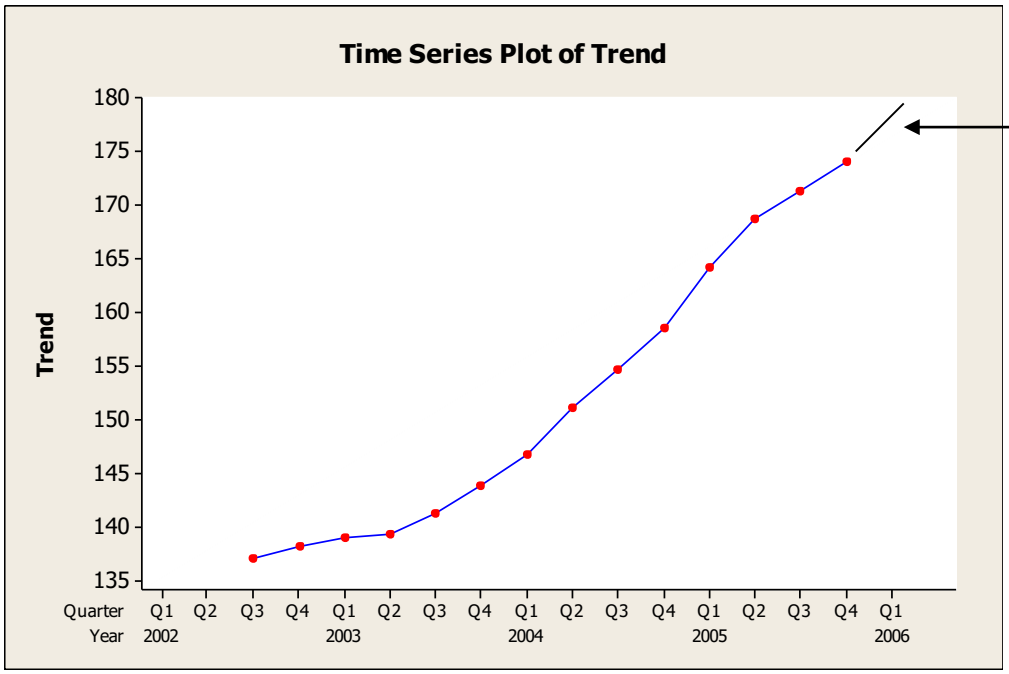
A long run of values above or below 1 will provide evidence of a cyclical effect.

(With the Additive Model, R is assumed to be 0. Therefore, we want the average R, in that case, to be zero).

5.4 Making a Forecast from a Time Series

Note that the phrase 'forecast is otherwise accurate' is very important. It is obvious that when we are forecasting the future we are unlikely to be absolutely accurate but we can achieve a reasonable degree of accuracy.

Any forecast will be based on projecting the trend i.e. any forecast relies on our knowledge of the trend. The trend will not suddenly change direction. Using the previous example, we can see from the graph below that the trend is beginning to rise more steeply.



We, of course, do not know what will happen over the next few quarters e.g. it may continue to rise reach a maximum and fall or it may slowly begin to fall or it may start to rise even more steeply. However, it will not happen quickly so we can forecast over the next few quarter with a good degree of confidence. The longer the period we want to forecast for the more potential for inaccuracy.

In the plot above we have drawn the trend and made a projection, extending over the next two quarters. Note that this is fairly subjective. Our new trend values are

2006	Quarter 1	176.50
	Quarter 2	178.75

Let us use these figures to predict the Import figures for Quarters 3 and 4.

The calculations are:

REQUIRED Predicted imports in Quarter 3
Predicted imports in Quarter 4

FROM GRAPH Trend value for Quarter 1 = 176.50
Trend value for Quarter 2 = 178.75

Note

1. Column 5 = Trend = Centred 4Pt Moving Average
= Average (T-1, T_0) 4Pt Moving Average (Column 4)

2. Column 4 = 4Pt Moving Average
5.4.1.1.1.1.1. = Average (T-2, T-1, T_0 & T+1) Original Data (Column 3)

Year	Quarter	Imports	4Pt Moving Average	Trend = Centred 4PT MA	
2002	1	114			
	2	142	136.75		
	3	155	137.25	137.00	
	4	136	139.25	138.25	
2003	1	116	138.75	139.00	
	2	150	139.75	139.25	
	3	153	142.75	141.25	
	4	140	144.75	143.75	
2004	1	128	148.75	146.75	
	2	158	153.50	151.13	
	3	169	155.75	154.63	
	4	159	161.25	158.50	
2005	1	137	167.00	164.13	
	2	180	170.25	168.63	
	3	192	172.25	171.25	
	4	172	175.75	174.00	
2006	1	145			from graph
	2	194			from graph
	3				
	4				