An Introduction to **Statistics** using Microsoft **Excel**

BY

Dan Remenyi
George Onofrei
Joe English

Published by Academic Publishing Limited

Academic Publishing Ltd
Curtis Farm
Kidmore End
Nr Reading
RG4 9AY
UK
info@academic-publishing.org

When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.
William Thomson later Lord Kelvin, Mathematician and Physicist 1824-1907

ISBN 978-1-906638-55-9
First edition 2009

Preface

To illustrate the statistical functions and techniques within Excel we have used examples drawn for the world of research. We chose to do this because we believe this will clearly demonstrate some of the facilities within this spreadsheet. However, the statistical power of Excel will allow readers to use this software in a variety of ways, far beyond that of research.

Statistics is both a science and an art. It is a science in that it is a body of knowledge which is built on a series of well established rules of applied mathematics. Rigorous statistical analysis is required for quantitative research. There is no argument about how statistical calculations or techniques are to be performed and if the rules are not followed then incorrect answers will most probably be produced. However statistics is also an art that requires a considerable amount of judgement on the part of a practitioner. Judgements and decisions have to be made that relate to deciding how a research question should be designed and the role of data and statistics in answering it. There are issues relating to which statistical technique to use, what level of significance to work at, and how the results can be interpreted.

What is also problematic in statistics, which is why it should be regarded as an art, is that there may be disagreements among practitioners about the meaning on these judgements and what answers could be given. It is indeed common to find different researchers taking quite different positions concerning which statistical technique is best to use. Not to mention the fact that the same statistical results can be understood differently.
Excel, with its built in statistical functions is a good tool to use to become familiar with statistics. There are many statistical functions in Excel but this book only addresses the functions required to perform the tasks most often required by researchers. Advanced functions are not used. It is also important to say that Excel has limitations as a statistical package and experienced researchers may well need to use a dedicated statistics package as well as Excel.

This book is not intended to cover all the possible statistical procedures or techniques that can be performed within Excel. The intention is that it will be an introduction to statistics which will facilitate readers to acquire the knowledge to understand the basics and to progress further if he or she so wishes. Furthermore for those who wish to use Excel in more advanced ways there is a list of add-in products in the Appendix.

How to use this book

This book starts with the assumption that little is known about statistics; however it does assume that the reader has some knowledge of Excel.

This book has been written as a tutorial and as such the techniques of statistics are illustrated through many examples. The mathematical equations required for statistical concepts such as the mean and the standard deviation have not been provided as these are easily performed with functions in Excel.

It is hoped that readers will follow the examples by using the techniques and for this reason the data referred to in the book is available to the reader by downloading files from the web.

For beginners to statistics it is preferable to start reading at the beginning i.e. Part 1 and proceed slowly through the book. For those who have a knowledge of statistics the book may be started at Part 2 or at Part 3.

A glossary has been provided at the start of each Part which covers the statistical terms used in that section of the book.

Self tests, assignments and exercises are provided and worked solutions to these are available on request.

Three types of files are available for use with this book. These are obtainable at the following website http://www.academic-publishing.org/intro_excel.htm. The first set of files contains data for the worked examples in the book. The second sets of files are Excel files which can be used as templates. The third set of files contains data which is used in the exercises at the end of each part.

Acknowledgements

We would like to acknowledge the contribution of many colleagues and students over the years who have all made a contribution to our understanding of the body of knowledge we call statistics.

Table of Contents

Understanding statistical techniques

2.1 Introduction

Part 1 addressed the issues of descriptive statistics. Means, medians, standard deviations, correlation coefficients and other statistical measures have been used to describe a sample of data which we have obtained. It will have been noticed that most of the exercises in Part 1 were based on data obtained by the use of a simple questionnaire.

In Part 2 new ideas and new techniques are introduced and in so doing, the learner will move **beyond the world of descriptive statistics** into what is referred to as **inferential statistics**. The difference between descriptive and inferential statistics is that with descriptive statistics it is only possible to make statements about the sample. In inferential statistics it is possible to use data from the sample to make statements about the whole population. Specifically, if we have a suitable sample it is possible to know quite a lot about the whole population from which the sample came. It is important for researchers to always keep in mind the difference between the population (the total set of all possible elements from which a sample is drawn.) and the sample (a part of something larger such as a subset of a population. This sample is usually drawn using a sampling frame).

Before commencing the discussion of inferential statistics it is necessary to introduce learners to a few other concepts and the first issue we address is the shape of the data through data frequency tables.

2.2 Data Frequency Tables

Data frequency tables have been considered before in Part 1 but they are addressed here again because much of what follows is based on the idea that if we know what a particular data frequency table or data frequency distribution looks like, then we really know quite a lot about the variable which this table or distribution represents. Remember we discussed some of the issues about a data distribution when we addressed skewness (left or right) and kurtosis (peaked or not peaked) in Part 1. Shortly we will look at a data distribution which is not skewed and where the kurtosis is neither over or under peaked.

Remember the data frequency table refers to a way of presenting a summary of the data in such a manner that it facilitates the possibility of seeing patterns or relationships in the data. A data frequency table shows how many times each data point (observation or outcome) occurs in a given data set. Distributions may take different shapes and this section of

the book will consider data which is typical of that obtained from the questionnaire concerning the background of students registered at a School of Business and a School of Accounting used in Part 1. The issue or variable which will be considered here is the length (in months) of working experience which students obtained before registering for their post-experience degree.

In this example we have obtained 30 completed and usable questionnaires from students from the School of Business and another 30 completed and usable questionnaires from students from the School of Accounting. We have 60 data points in all.

A usable questionnaire is one which has been sufficiently completed that it may be included in the data set obtained. Researchers sometimes exclude questionnaires where more than a few questions have not been completed or answered by the respondent. When this happens it is usually believed that the questionnaire was poorly designed. In the same way as discussed in Part 1 a small number of missing data points or elements may be estimated. Sometimes there can be a very large number of non-respondents i.e. questionnaires not being returned at all, and this can damage the credibility of the survey.

The 30 respondents from the School of Business supplied the following number of months working experience:

Table 2.1: School of Business, number of months working experience

23	28	29	34	34	39	43	44	45	45	48	48	49	54	54
54	55	56	56	65	65	65	67	73	76	76	77	78	87	92

Respondents from the School of Accounting replied with the following data:

Table 2.2: School of Accounting, number of months working experience

10	12	12	16	19	20	22	23	23	23	26	28	29	32	33
34	34	41	43	43	44	45	45	54	56	56	56	65	67	76

These two data sets have been reproduced in this section across columns to conserve page length in this book. These data would normally be entered in a spreadsheet in one column.

It is possible to create a data frequency table by just counting the number of times each data point occurs. However, it is often the case that it

is better to group data into intervals such as 10-19, 20-29 etc. This occurs with data for quantitative variables such as age, weight, income etc, where there are likely to be many different outcomes.

The first step in producing a Frequency Table is to establish the range of the data. The technique for doing this has been described in Part 1. The range for the School of Business is 69 and the range for the School of Accounting is 66.

The second step is to decide the number of groups/intervals into which the data should be divided. A heuristic for this is that the data may be grouped into the number of intervals represented by the square root of the sample size. As the sample size is 30 five or six groups would be appropriate in this exercise[26]. It is useful to keep the width of the intervals constant except perhaps for the first and final groups. In this exercise we have used interval markers of under 25, 36, 48, 60, 72 and greater than 72. These numbers need to be entered I6 to I11.

Excel has a function which allows data frequency tables to be constructed which is called =frequency().

The =frequency() function is an array function which means that it is entered in a different way to other functions. The function needs two pieces of information:-

1. the full range of data from which the frequency distribution is required
2. the intervals which are to be used in the data frequency table. This is called the bin range.

Considering Figure 2.1, the required data distribution in intervals has been entered into the range I6 through I11. The =frequency() function can now be used in the adjacent column to calculate the frequencies. As =frequency() is an array function the range J6 through J11 is first selected and then to produce a frequency table for the responses from the School of Business, the following formula is entered in J6.

=frequency(C3:C32,I6:I11) [CTRL + Shift + Enter]

[26] There is in general agreement that the number of intervals should not be less than 5 and not more than 20.

	A	B	C	D	E	F	G	H	I	J
1			Raw data							
2			Months SoB	Months SoA						
3		1	29	10						
4		2	76	23					Data	
5		3	77	32					Distribution	Frequency
6		4	56	26					25	1
7		5	45	28					36	4
8		6	54	56					48	7
9		7	56	43					60	7
10		8	34	44					72	4
11		9	54	56					>72	7
12		10	65	20						
13		11	48	41						
14		12	76	29						
15		13	44	45						
16		14	34	56						
17		15	43	34						
18		16	48	76						
19		17	49	65						
20		18	54	45						
21		19	23	22						
22		20	28	12						
23		21	55	33						
24		22	39	43						
25		23	65	54						
26		24	67	23						
27		25	73	12						
28		26	45	16						
29		27	92	19						
30		28	78	34						
31		29	87	23						
32		30	65	67						

Figure 2.1: The data and the =Frequency() function

In the case of an array function it is necessary to hold down the **CTRL key and the Shift Key and to then press Enter**. The result is that the frequency is calculated for each interval or bin in the range and are placed in cells J6 to J11.

To calculate the frequencies of the School of Accounting, select the range K6:K11 and enter the formula

=frequency(D3:D32,I6:I11) [CTRL + Shift + Enter]

The frequency table which is produced by Excel is shown in the right hand corner of Figure 2.1 in the range I4 to J10.

The results of the =frequency() function are now reproduced in Table 2.3.

Table 2.3: School of Business Frequency Table of the number of students and the number of months working experience.

Months Experience	No of Students
Under 25	1
26-36	4
37-48	7
49-60	7
61-72	4
above 72	7
Total	30

Having established the frequency of the individual groups or classes, the relative frequency can be calculated as a percentage of each group relative to the total. The cumulative relative frequency is then the percentage across the range shown in Tables 2.4 and 2.5.

Table 2.4: School of Business Frequency Table with relative frequency

Months Experience	No of Students	Relative frequency
Under 25	1	0.03
26-36	4	0.13
37-48	7	0.23
49-60	7	0.23
61-72	4	0.13
above 72	7	0.13
Total	30	1.00

Table 2.5: School of Business Frequency Table with relative frequency and **cumulative relative frequency**

Months Experience	No of Students	Relative frequency	Cumulative Relative frequency
Under 25	10	0.33	0.33
26-36	7	0.23	0.57
37-48	6	0.20	0.77
49-60	4	0.13	0.90
61-72	2	0.07	0.97
Above 72	1	0.03	1.00
Total	30	1.00	

These frequency tables may be plotted as histograms. Figure 2.2 shows the results for the School of Business and. Figure 2.3 shows the results for the School of Accounting.

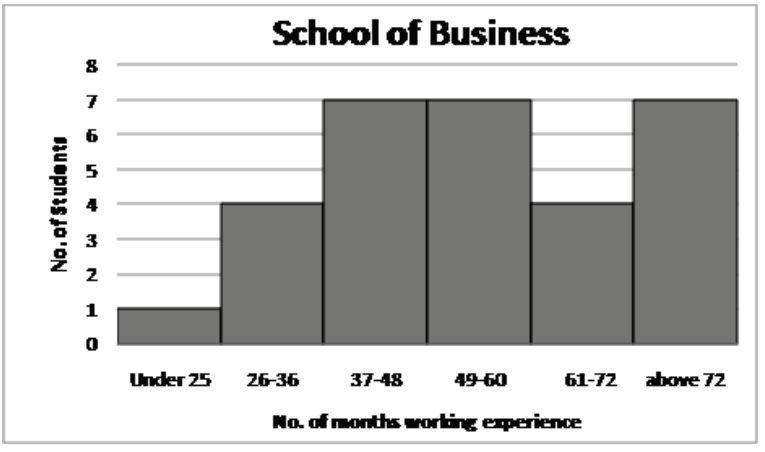


Figure 2.2: A histogram used to examine the shape of the data frequency table for the School of Business

Excel does not directly produce a histogram with the bars touching each other as shown in Figure 2.2. To obtain this effect a normal bar chart is first drawn. Then the cursor is placed on one of the bars and the right mouse button clicked. Choose Format Data Series and then Options. This will allow the Gap Width to be specified as Zero and the histogram effect is achieved.

Note that this chart using the given intervals is tri-modal. This example shows one of the difficulties with using the mode as it is not unique. It may also be the case that this data actually contains more than one distinct sub-group

In Figure 2.3 there is a distinct trend for degree candidates to register for the post-graduate degree soon after their first degree and thus without a larger number of years working experience.

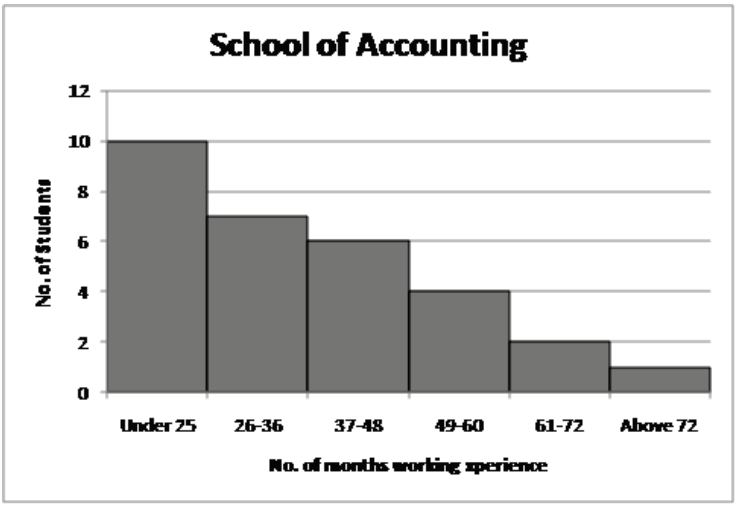


Figure 2.3: A histogram used to examine the shape of the data frequency table for the School of Accounting

No further analysis is required to recognise that the distribution of work experience had by the degree candidates in the School of Business and the School of Accounting is different. These two data sets can be plotted on the same axis using a line graph shown in Figure 2.4.

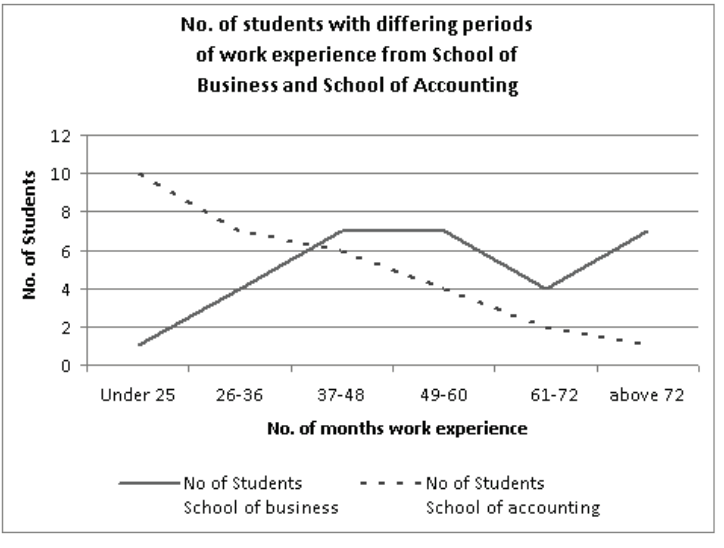


Figure 2.4: The School of Business and the School of Accounting data as traces on one graph

The line graphs as shown in Figure 2.4 are sometimes called a trace or frequency polygon as they show the line which joins the mid-points of the tops of the rectangles in the histograms.

It may be seen from both the tables and the graphs that there are differences between these two distributions. The School of Business has more mature students (more working experience) and these are spread across the work experience range with three modal points, which are the most work experienced groups. In the case of the School of Accounting there are fewer students with more working experience and the mode is the least work experienced group.

As mentioned above the data frequency table describes the pattern in the data and this is one way in which a distribution may be represented.

In addition a distribution may be described using the mean and the standard deviation. It is also possible to use a mathematical formula to describe a distribution. However, it should be noted that the data distributions described above are not probability distributions. Probability distributions are described below.

2.3 Normal Distribution

In Figures 2.2 and 2.3 above a practical data frequency distribution was created using the =frequency() function, and in this case graphed, from the sample data we obtained using a questionnaire. This way of presenting the data showed that there was a particular shape to the data obtained and this represented a pattern in the opinions which were offered. In other instances the shape of the data would represent how events took place. There are many different types of shapes which will appear when data is graphed. One of the most frequently encountered is bell shaped and thus suggests that the sample and population come from what is called a normal distribution. The normal distribution is of great importance in statistical analysis. The normal distribution is referred to as a probability distribution. The notion of probability is central to the idea of the normal distribution and we will refer to this many times in the balance of this Part of the book. The normal distribution is sometimes called a Gaussian distribution. Carl Friedrich Gauss was a famous German mathematician working in the 19^{th} century.

A probability distribution is sometimes thought of as a theoretical data frequency distribution which states how outcomes or observations are expected to behave. As with the data frequency distributions already de-

scribed above it may be represented by tables or diagrams or a mathematical formula.

The graph of a normal distribution is bell shaped which is symmetrical around the mean value in the middle of the graph, and it has a number of particular properties. Figure 2.5 shows the shape of a typical Normal distribution curve produced in Excel and the distribution table of values is shown in Figure 2.8, which will be referred to in detail later.

By convention the total area under a normal shaped curve is always 1. Thus the bell shaped curve consists of two sides each of which is 0.5 in area. The probabilities are represented by the area under the curve. By definition this means that we cannot talk about the probability of a single value occurring. We talk about the probability of values being greater than or equal to some observed or specified value. We can also talk about the probability of values being between certain limits. Because the area under the curve is equal to 1 the probabilities mentioned above may be understood as being the fraction of the population that lies above or below or within the specified values. It is important to recall that the probability of an event which will certainly occur such as the fact that we will all die is 1 and the probability of an event which is impossible such as we will all fly to the moon is 0 (zero).

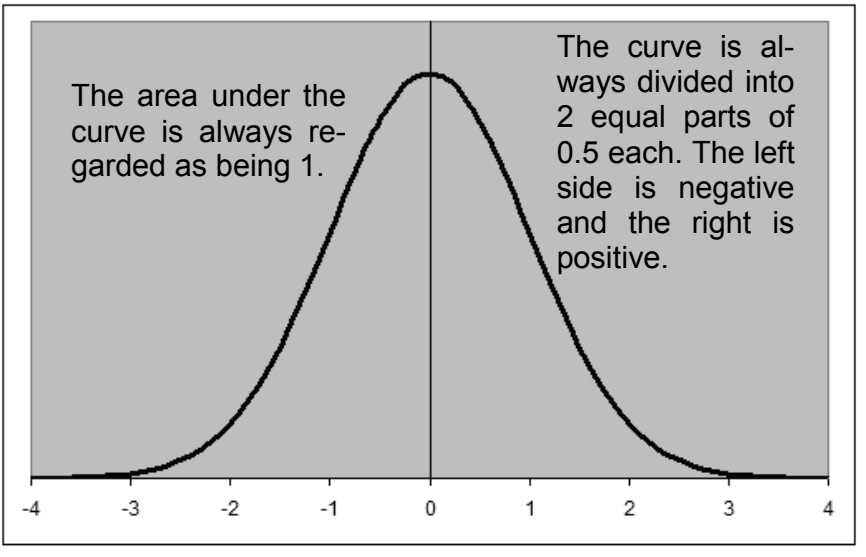


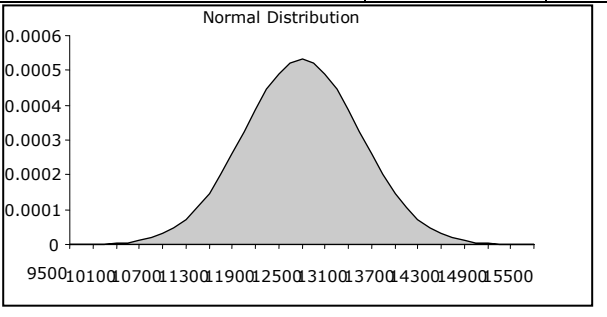
Figure 2.5: The bell shape of a standardised normal distribution curve or graph

The Normal distribution is a continuous probability distribution which describes the frequency of the outcomes of the random variable. The actual shape of the normal distribution curve is a function of the mean value and standard deviation[27] of the distribution. The mean and the standard deviation of the distribution are referred to as the parameters of the distribution.

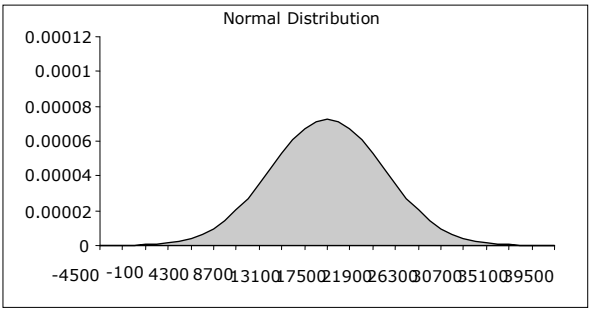
All the following distributions are normally distributed.

Different means and different standard deviations lead to different bell shaped curves as may be seen in Figure 2.6.

Distribution	*Mean*	*Std dev*
Distribution 1 The number of work permits issued each year in the country	12,500	750



Distribution 2 The number of children vaccinated against flu each year in the country	17,500	5000



[27] The mean and the standard deviation were discussed in Part 1 where it was shown how to calculate them using Excel functions.

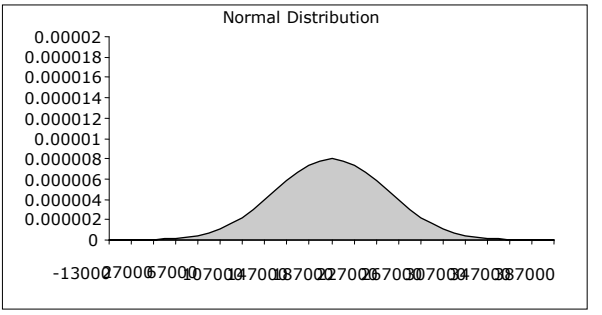
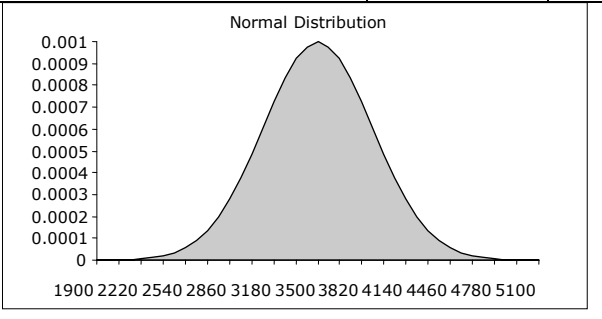
Distribution	*Mean*	*Std dev*
Distribution 3 The number of microchips manufactured in China each day	187,000	50,000
		
Distribution 4 The number of burgers sold in the fast-food restaurants in Dublin airport each month.	3,500	400
		

Figure 2.6: Different data sets which display the characteristics of a normal distribution.

The theoretical distributions in Figure 2.6 have different means and different standard deviations, but it is also possible for there to be many distributions with the same mean, but with **different standard deviations** as is shown in Figure 2.7. Note the smaller the standard deviation the "tighter" the graph, and the larger the standard deviation the "flatter" the graph. Tight graphs, i.e. ones with small standard deviations, suggest that there will be less variability in the sample and flat graphs i.e. ones with large standard deviations, suggest that there will be a high degree of variability. This higher standard deviation can mean more risk than a lower degree of variability.

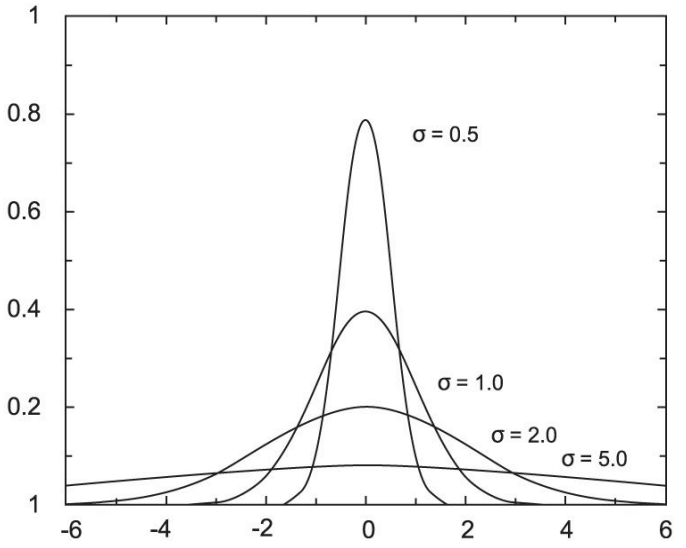


Figure 2.7: Four data distributions with the same mean but with 4 different standard deviations.

In Figure 2.7 the σ is used to denote the standard deviation.

One of the more important attributes of the normal distribution curve is the specific way in which the outcomes of the variable are spread about its mean. If a distribution is normal then by definition 68% of the population outcomes/values lies within *plus or minus one standard deviation* from the mean and 95% of the population outcomes/values lies within *plus or minus two standard deviations* (actual value is 1.96 but this is usually expressed or rounded to 2) from the mean. And finally 99% of population outcomes/values lies within *plus or minus three standard deviations* (rounded up from 2.58 to the nearest integer) from the mean. This is illustrated in Figure 2.8.
Note, in theory the tails of the graph in Figure 2.8 do not touch the x-axis. They are said to extend to infinity. As a result it is possible to occasionally find very large or very low values which are actually data points within the distribution[28].

[28] Recall the issue of outliers which was discussed in Part 1 of the book. Some outliers will be perfectly respectable members of a data set whose values are to be found deep in the tails of the distribution.

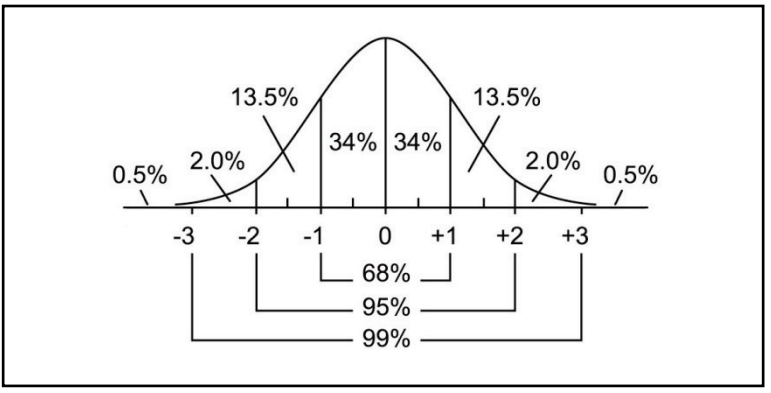


Figure 2.8: The percentage of the data within 1, 2 and 3 standard deviations[29] under a normal distribution.

Remembering the rule of how data is distributed in a normal distribution[30] and using these characteristics we can interpret research results in a meaningful manner. For a normal distribution where the mean is 25 and the standard deviation is 4, we can say that approximately 99% of the population values lie between 25 + (3 * 4) which equals 37 and 25 – (3 * 4) which equals 13. This means that we can say that we expect with 99% confidence that an observation drawn from this distribution will take a value of between 13 and 37. Note that it is assumed that the observation has been *randomly taken from the population* and it is important to remember that the two tails of the normal curve *theoretically stretch to infinity.* So it is possible to obtain a very small value and a very large value, which although highly unlikely, are valid observations or values in the normal distribution[31].

An application of this is to use the normal distribution to establish the relative position of an observed value to others in a population. For example, suppose an individual drawn at random from the above example (mean = 25 and standard deviation = 4) exhibits a value of 33 then it may be said that the individual is two standard deviations above the mean and is therefore in the top 2.5% of the population.

It should be noted in Figure 2.8 that there are three divisions of the curve on either side of the mean i.e. 3 standard deviations. If the data

[29] It is important to recall that although 1, 2 and 3 standard deviations are frequently used to describe the distribution of data under the normal curve these numbers are only approximations. The actual numbers are 0.84, 1.96 and 2.5.

[30] This is sometimes referred to as the 68-95-99 Rule.

[31] It is possible that invalid observations occur i.e. errors and if this is the case they need to be adjusted or omitted from the data.

distribution is taken as a whole there are six possible divisions between what would be, for practical purposes, the maximum and the minimum values of the distribution. Therefore if the range of a data set i.e. the difference between the maximum and the minimum values, are taken and divided by 6 then this resultant number may be used as a heuristic for the standard deviation. Clearly the actual calculation of the standard deviation is better and with a spreadsheet this is easy to do. But sometimes this heuristic is useful.

2.4 The standard normal curve

Any variable a researcher is working with will have its own mean and standard deviation. All these means and standard deviations could be different and if every time we wanted to study a situation we had to work (i.e. make our calculations) with different means and standard deviations we would be faced with arithmetical challenges. However we are able to minimise this arithmetic work by ***standardising*** any variable to have a mean of zero and a variance or standard deviation equal to 1.

Returning to Figure 2.8 the normal distribution shown here is called the ***standard normal curve*** because the mean is zero and the standard deviation is 1. It is unusual for any normal distribution to have a mean of zero and a standard deviation of 1. However, any normal distribution variable can be standardised or transformed so that it may be considered to have a mean 0 and standard deviation 1. When this is done the resulting standardised variable is called the Z variable and the probability associated with the Z variable is shown in the Standard Normal Distribution Tables in Figure 2.10 (on page 86).

Now to explore the idea of probability.

Example 1

Recall that having processed the results of the questionnaire in Part 1 we found the response to Question 1 to have a normal distribution[32] of mean 3.47 and a standard deviation of 2.05.

We find an additional questionnaire (apparently a late submission) with a response rating of 8 to question 1.

The quality of the lectures is excellent
Strongly Disagree *Strongly Agree*
1. ☐ *2.* ☐ *3.* ☐ *4.* ☐ *5.* ☐ *6.* ☐ *7.* ☐ *8.* ☒ *9.* ☐

[32] In order to be able to use this type of logic it is necessary to make an assumption about how the data was distributed.

We wish to establish the probability of a score of 8 or more coming from such a distribution (i.e. the distribution we found when we first analysed the data we obtained in Part 1). If traditional tables are to be used like those in Figure 2.10 then the first step is to standardize the score 8 in terms of the number of standard deviations it is from the mean. This is done by using the formula[33]:

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

Where, X is any given data point (we sometimes call this number the X-score) in the sample, in this case 8, μ is the mean of the population, in this case 3.47 and σ is the standard deviation of the population in this case 2.05.

Therefore

$$Z = \frac{8 - 3.47}{2.05}$$

= 2.21 Standard Deviation units above the mean.

The 2.21 is sometimes called the ***standardised value*** of 8 in terms of a data distribution where the mean is 3.47 and the standard deviation is 2.05.

The location of the Z = 2.21 which is the number of standard deviations the new data point, 8, is from the mean, may be viewed in Figure 2.9.

2.5 Normal curve and probability

In Figure 2.9 the area under the curve to the left of the Z-score (2.21) represents the probability that a value could be 8 or less, i.e. there is a high probability here. The area under the curve to the right of the Z-score (2.21) represents the probability that a value could be 8 or more, i.e. there is a low probability here. Note that we cannot use this technique to say what is the probability of a score being 8 as for a continuous variable any single X-score has a probability of zero.

[33] The Greek symbols μ (this character is pronounced mu) and σ (this character is pronounced sigma) are used by statisticians to represent the mean of the population and the standard deviation of the population. This will be discussed further later in this book.

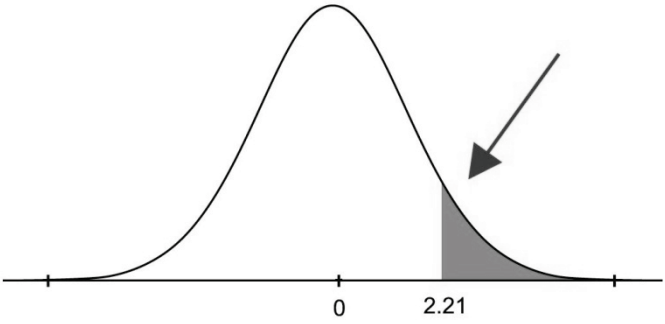


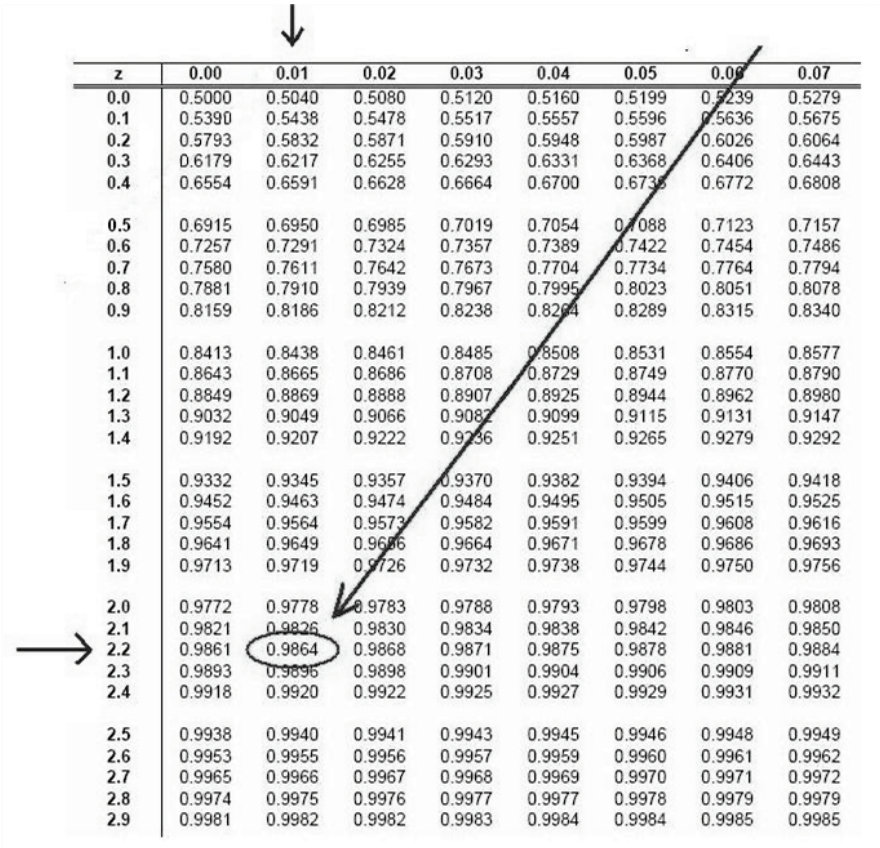
Figure 2.9: Using the normal curve to show the area of the curve representing the probability of a Z-score greater than 2.21

The Z-table which will be used here to find the corresponding probability is in Figure 2.10 below.

To use the Z-tables look down the first column in the table on the extreme left until 2.2 is found. The column next to the right shows the value of the Z-score for 2.20. If you move one column further to the right the value is that required for a Z-score of 2.21. Note if a Z-score of 2.22 had been required it would have been necessary to move another step to the right.

The table value for 2.21 is actually 0.9864 which represents the probability for a Z-score being less than 2.21 standard deviations above the mean. Therefore the probability of getting a value of 8 or more is equal to one minus 0.9864 which is 0.0136 or 1.36%[34]. Note that we subtract the 0.9864 from 1 because the total area under the curve is 1. A probability of 1.35% is very low and it would be wise to question the validity of the late arriving questionnaire with a score of 8.

[34] The area in the table is the area under the Standard Normal Curve to the left of the computed value of Z. In this case we are interested in the area under the curve to the right of the calculated value of Z.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279
0.1	0.5390	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985

Figure 2.10: A set of published Standard Normal Distribution Tables (Z-tables) showing probabilities for less than a given Z-score

The table shown in Figure 2.10 gives the probability of obtaining a Z-score from the specified value (in this case 8) back to the extreme left end (tail) of the curve. As in this example we wanted to calculate the probability of 8 or more occurring, it was necessary to find the area of the curve to the right of the z-score.

Subtracting the table value from one (unity) in order to determine the probability can sometimes present difficulties to the newcomer to this technique. In Figure 2.11 the normal curve is shown as comprising the section having a probability of 0.9864 and the section having a probability of 0.014. The area under the curve to the left of the Z-score is what is tabulated. Thus when we are interested in the area to the right of the Z-score we need to subtract the tabulated area from 1. Figure 2.11 shows the two parts of the area under the curve.

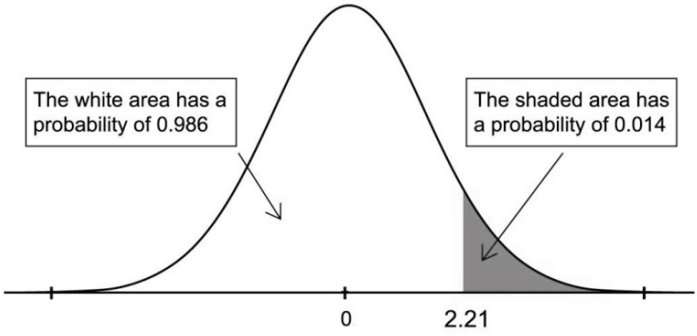


Figure 2.11: The sum of the two areas (blank and shaded) is equal to 1

Example 2

Another example to illustrate the above technique considers the probability of a value of 7 or more being returned for the question.

The quality of the lectures is excellent								
Strongly Disagree								*Strongly Agree*
1. ☐	*2.* ☐	*3.* ☐	*4.* ☐	*5.* ☐	*6.* ☐	*7.* ☒	*8.* ☐	*9.* ☐

The first step is to standardise the X-score using the same formula as before.

$$Z = \frac{7 - 3.47}{2.05}$$

= 1.72 Standard Deviation units

The location of the Z-score (1.72) is shown in Figure 2.12 below and the required area to the right has been shaded. Remember that the area to the left of the Z-score is the probability of a score of 7 or less.

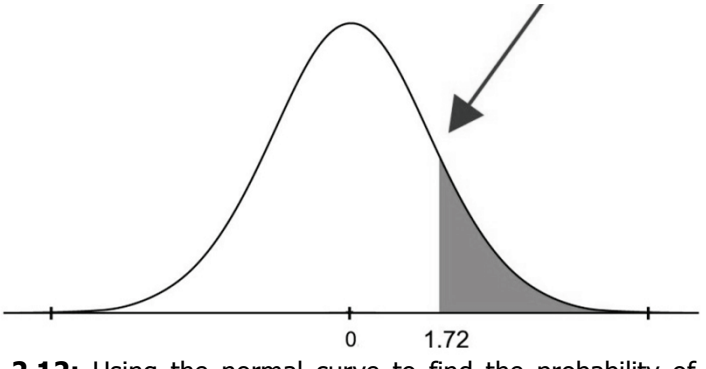


Figure 2.12: Using the normal curve to find the probability of the Z-score being greater than or equal to 7 or more.

Using the table in Figure 2.13, the area to the left of the Z-score (1.72) is 0.9573. Therefore the probability of obtaining a value of 7 or more is equal to one minus 0.9573 which equals 0.0427 or 4.27%.

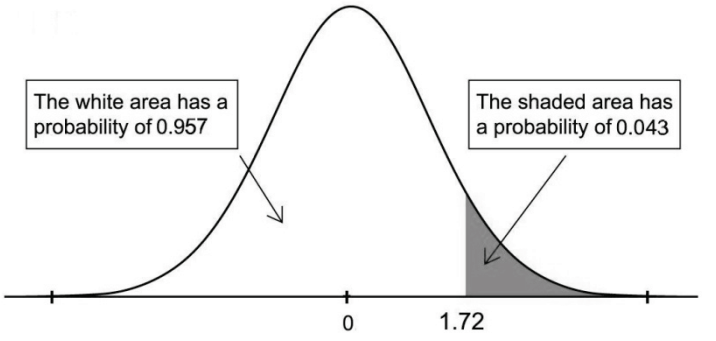


Figure 2.13:- The areas of the normal distribution and the probabilities associated there with.

With the use of Excel we do not have to resort to Z-score tables. Excel will calculate the numbers required and it will do so more speedily and also to a greater degree of accuracy.

2.6 The =normdist() function

Instead of using tables to look up Z-scores with which to compare a calculated test statistic to a theoretical statistic, a researcher can use Excel. Excel offers a function to calculate the probability of an observation being a member of a population with a given mean and a given standard deviation. This is the =normdist() function.

The form of this function in Excel is:

=normdist(x,mean,standard_dev,cumulative)

where x is the X-score.

Repeating Example 2 described previously but using the Excel =normdist() function the following is required in cell C6.

	A	B	C	D
1	**Normal Distribution**			
2	Observation		7	
3	Mean		3.47	
4	Standard Deviation		2.05	
5	Cumulative		TRUE	
6	Probability		0.957461	

Figure 2.14: The Excel spreadsheet used to calculate the probability of an X-score for normal distributions

In Figure 2.14 the formula in cell C6 is

=normdist(C2,C3,C4,C5)

where C2 = X-score (observed Value)
C3 = Mean to be tested against,
C4 = Standard deviation
C5 = Indicator point to use the probability density function. TRUE indicates the area under curve and FALSE indicates the ordinate[35].

The result shows a probability of the X-score being less than 7 to be 0.957461. Therefore, as before, to calculate the probability of the X-score being greater than 7 it is necessary to subtract the answer from one i.e. 1 - 0.957461=0.0427 or 4.27%.

Example 3

To look at another example, if we have a population with a mean of 40 and a standard deviation of 5, and we want to know the probability that an observation or outcome obtained from this population could be 30 or more we can use Excel as follows.

Figure 2.15 shows the result f using of the =normdist() function in cell B6.

[35] The FALSE indicator for this Excel function has not been used in this book. See the Excel Help function for more information about this option.

	A	B	C
1	**Normal Distribution**		
2	Observation	30	
3	Mean	40	
4	Standard deviation	5	
5	Cumulative	TRUE	
6	Probability	0.02275	

Figure 2.15: The calculation of the probability of the observation occurring from the probability distribution curve.

Once again in Figure 2.15 the formula in cell B6 is

=normdist(B2,B3,B4,B5)

where B2 = Observed Value,
B3 = Mean to be tested against,
B4 = Standard Deviation
B5 = Indicator point to use the probability density function. TRUE indicates the area under curve and FALSE indicates the ordinate.

The result shows a probability of the X-score being less than 30 to be 0.02275. Therefore, as before, to calculate the probability of the X-score being greater than 30 it is necessary to subtract the answer from one. i.e. 1 - 0.02275 = 0.97 or 97%.

Example 4

In another example we consider the same mean 40 and the same standard deviation of 5 but this time we want to know the probability that an observation or outcome obtained from this population could be 36 or more. We proceed as follows:

	A	B	C
13	**Normal Distribution**		
14	Observation	36	
15	Mean	40	
16	Standard deviation	5	
17	Cumulative	TRUE	
18	Probability	0.211855	

Figure 2.16: The calculation of the probability of the observation occurring from the probability distribution curve.

In Figure 2.16 the formula in cell B18 is

=normdist(B14,B15,B16,B17)
=0.211855

The Excel spreadsheet shows a probability of 0.0211855. As before to calculate the probability of the X-score being greater than 36, it is necessary to subtract the value returned by Excel from 1. This calculates to 78.8% (i.e. 1 - 0. 0211855 = 0.788).

= 79% (rounded up)

Example 5

In this example, we consider the distribution of year end exam results that follow a normal distribution with a mean of 55 and a standard deviation of 10. If the results followed this distribution in the next examination what mark would you expect to divide the class into two groups, one of which consisted of 95% of the pupils and the other 5% of the pupils.

The Z-value for 95% (or .95) of the curve is in the right hand tail and by consulting the tables it will be found to occur at 1.645 and this is represented in Figure 2.17.

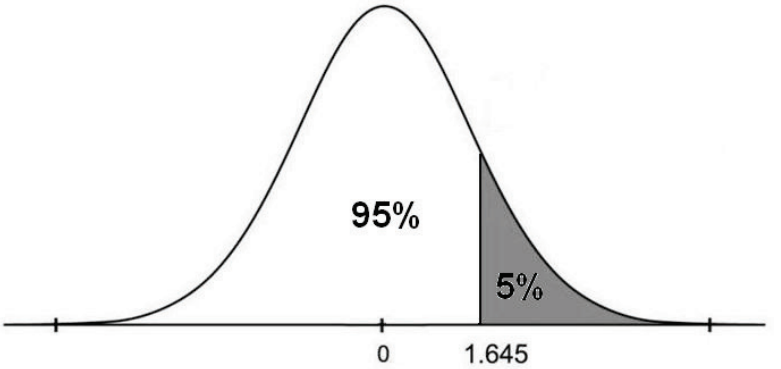


Figure 2.17: Z-value of 1.645 splits the total area into 95% and 5%.

The standardised value of the marks (X) is equal to $\frac{X-55}{10}$ which is set to equal 1.645 in this example.

This is represented by

$$\frac{X-55}{10} = 1.645$$

Multiplying both sides of this equation by 10 and rearranging to solve for X we find:

$$X = 55 + 1.645 * 10$$
$$X = 71.45$$

Example 6

The Jones twins are highly competitive. The first twin is in a group for mathematics which has been given a test for which the results are distributed normally with a mean of 65 and a standard deviation of 5. The first twin's score is 73. The second twin is in a different group which does a totally different type of mathematics test and whose results are distributed normally with a mean of 35 and a standard deviation of 11. The second twin scores 43. Which twin has done relatively better in their test?

First twin: Z-score = $\frac{73-65}{5}$ = 1.6 standard deviations

Second twin: Z-score = $\frac{43-35}{11}$ = 0.73 standard deviations

Using tables $P(Z \geq 1.6) = 1 - .9452 = .0548 = .055 = 5.5\%$ which means that this twin is in the top 5.5% of the class.

Using tables $P(Z \geq 0.73) = 1 - .7673 = .2327 = .234 = 23.4\%$ which means that this twin is in the top 23.4% of the class

Thus the first twin has performed relatively better to the peer group than the second twin.

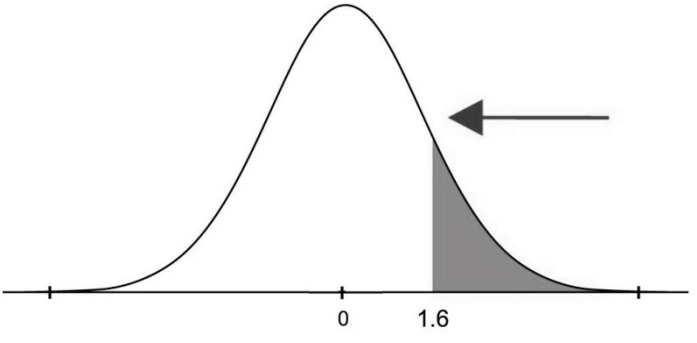


Figure 2.18: The shaded part of the curve shows the probability of first twin with a score of 73.

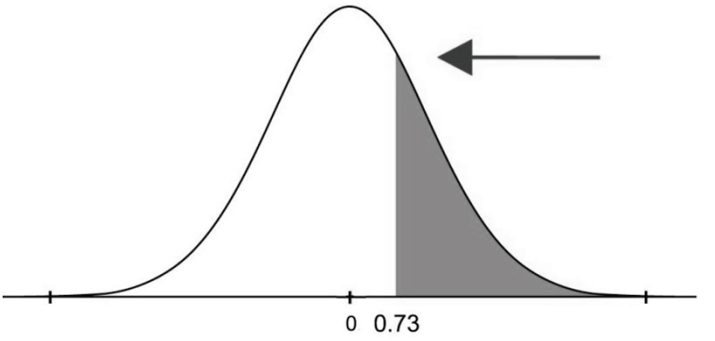


Figure 2.19: Shows the probability for the second twin with a score of 43.

2.7 Estimation and confidence interval

In the context of this book, estimation is the process of establishing the likely value of a variable. The most direct form of estimation is to establish a single-point value. As has been addressed in Part 1 of this book, the mean, the median or the mode may be used as single-point estimates. However, the use of these single statistics does not always contain enough information and the alternative, which is an interval estimate, may be required.

For example, looking at the results of Question 1 from the questionnaire in Part 1 of this book the question was:

The quality of the lectures is excellent								
Strongly Disagree								*Strongly Agree*
1. ☐	*2.* ☐	*3.* ☐	*4.* ☐	*5.* ☐	*6.* ☐	*7.* ☐	*8.* ☐	*9.* ☐

With a sample size of 30, the mean score for this question was 3.47 and the standard deviation was 2.05.

If a single point estimate of the true mean score (μ) for Question 1 is required then the average score $\overline{X}$ of 3.47 can be used. This is the "best" estimate using a single-point value. However, it is possible that more information about the estimate of the mean is required. Here an interval estimate can be calculated and in this case the single number is accompanied by confidence limits.

Recall that the average score for Question 1 of the sample of the 30 questionnaires returned is $\overline{X}$ = 3.47 with a standard deviation s=2.05. Because we are dealing with a sample mean the confidence interval which we require will use the standard error as described in Part 1. This is calculated as the standard deviation divided by the square root of the size of the sample. In this case the estimate of the standard error for Q1 is 0.374. To establish the 95% confidence limits we need to add 2 standard errors[36] to the calculated mean and subtract 2 standard errors from the calculated mean. In this case the results are 4.21 and 2.73 respectively. Thus we conclude with 95% confidence that the true lies between these two values.

We might also be interested in knowing what the position is at a confidence level of 99%. In this case we add 3 standard errors to the calculated mean and subtract 3 standard errors from the calculated mean, which produces results of 4.59 and 2.35 respectively. This technique of estimating the interval value of the responses can be used for all questions in the questionnaire in Part 1 if required.

2.8 Standard Error of the Mean

As indicated above the Standard Error (SE) was first mentioned in Part 1 of this book in the same section as the standard deviation, but its use was not described. Whereas the standard deviation applies to a whole population the standard error only applies to the sampling distribution of the means drawn from a population.

Many different samples of the same size may be drawn from a population. The number of samples can be very large if the values comprising the sample are allowed to be taken more than once from the population. If this is the case we refer to the samples as having been taken with replacement.

In Figure 2.20 below the samples are of the same size.

[36] Remember that the actual number is 1.96 but this is frequently rounded to 2.

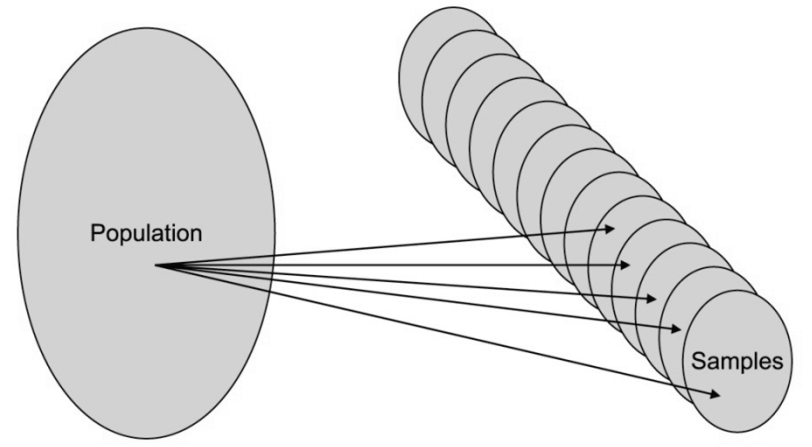


Figure 2.20: A large number of samples may be taken from a population especially if the sample elements are taken another time.

Since the samples are drawn from the same population the population standard deviation for each sample is σ sigma. Thus the population standard error $SE_{(P)}$ for samples of size n is

$$SE_{(P)} = \frac{\sigma}{\sqrt{n}}$$

In practice the value of σ is not known and thus is estimated from the observations from a single sample for which the inferences are to be based. The SE is calculated by dividing the standard deviation of the single sample on which inferences are to be made by the square root of the sample size (n) i.e.

$$SE = \frac{\text{Sample Standard Deviation}}{\sqrt{\text{Number of Observations}}}$$

$$SE = \frac{SD}{\sqrt{n}}$$

The SE will always be smaller than the standard deviation of the population (for $n \geq 2$).

Furthermore, in Figure 2.21 we show that the distribution of sample means will be closer to the mean of the population than the data points in the original population as the standard deviation of the sample means is smaller.

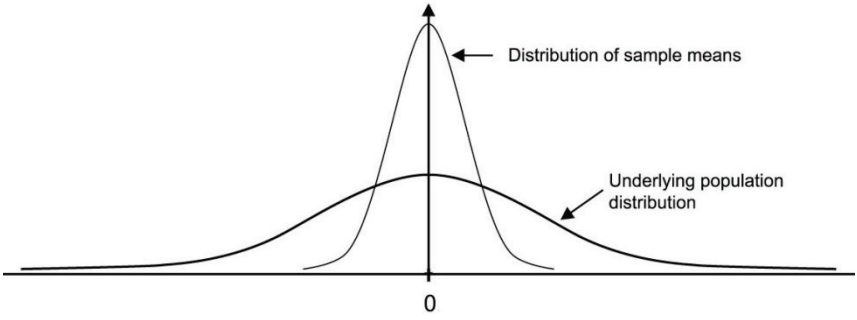


Figure 2.21: The distribution of the population and the distribution of sample means within that population

The sampling distribution of the mean of a sample of size N≥30 from a population with mean μ and standard deviation σ will be approximately normally distributed with mean μ and standard error = $\frac{\sigma}{\sqrt{n}}$. This follows from one of the most important principles in statistics which is referred to as the central limit theorem[37].

As a consequence the Z-score will be $Z = \frac{\left(\overline{X} - \mu\right)}{\left(\frac{\sigma}{\sqrt{n}}\right)}$ for sample means and will follow a standard normal distribution.

The standard error is used for interval estimation of the true mean μ and for carrying out tests of hypotheses concerning the true mean μ of a distribution.

In practice we usually do not know the value of σ and estimate it with the sample standard deviation referred to as s. At that point the Z-score (normal distribution) is not used, but instead a t (the so-called Student t) distribution is used.

The equivalent of the Z-score equation then becomes

[37] The central limit theorem states that with a large population the means of the samples taken will be normally distributed and that the mean of the means of the samples will be the mean of the population and the standard deviation of the distribution of the means will be standard deviation of the population from which the population was drawn divided by the square root of the sample size. For the central limit theorem to apply the sample should be greater than 30 and it should be small relative to the population.

$$t_{n-1} = \frac{\left(\overline{X} - \mu\right)}{\frac{s}{\sqrt{n}}}, \qquad v = n - 1 \text{ (where v=degrees of freedom)}$$

which follows a t-distribution with (n-1) degrees of freedom.

The use of the t-distribution[38], when the sample size is less than 30, actually depends on knowing that the random variable being studied follows a normal distribution. However, since the t-distribution or t-statistic is known to be robust with respect to the normal distribution assumption, we tend to use it without paying too much attention to this issue.

In general when we work with means of samples we use the standard error, but when we work with individual data points we use the standard deviation.

The larger the sample sizes in use the smaller the standard error.

2.9 Hypothesis testing

Hypothesis testing is a central concept in statistical reasoning and thinking and when this technique is mastered it will provide the researcher with a number of useful tools.

Hypothesis testing is an important way in which we can increase our confidence in the findings of our research and a number of examples will now be used to illustrate this. A hypothesis test which is often performed tests a claim concerning the "true" mean (or some other parameter) of the population.

In carrying out such a test one needs to specify the Null Hypothesis, the Alternative Hypothesis and the level of significance.

Before the test can be performed the researcher needs to draw a sample of acceptable or perhaps "credible"[39] size from the population and then compute the mean and the standard deviation of the sample.

[38]There are actually two assumptions necessary for the t-test (1) that the values in your sample should be independent of each other (2) the selection of any values from the population should not affect the selection of another and random sampling with replacement is regarded as the most appropriate way of doing this. The second assumption is that your population should be normally distributed. The t-test is preferred because it may be used in many situations where the population variability is not known.

[39] The "credible" size will depend on the hypothesis being tested and will vary considerably.

Then the appropriate test statistic needs to be computed. This is given by

$$t_{n-1} = \frac{\left(\overline{X} - \mu_0\right)}{\frac{s}{\sqrt{n}}}$$

where μ_0 is the value of the mean specified under the Null Hypothesis and (n-1) is a parameter of the t-distribution[40] which is referred to as the degrees of freedom. Increasingly the t-test is used in these circumstances as it provides usable results with both small and large samples.

For example, assuming that as a result of the original study the management of the Faculty sets a bench mark of 3.47 for the student quality evaluation. At a later point in time a new sample of 20 students are asked to complete the same questionnaire as before.

We are interested in the reply to Question 1. In this case the average for this new sample is 4.2 and the standard deviation is 2.0.

The research question is: *Does this new data set support* ***the claim*** *that the true average for the population is statistically significantly greater than 3.47 at the 5% level of significance?*

Note the hypothesis is always a claim which we then try to refute i.e. reject. A hypothesis is not proven, and it is technically not correct to say that a hypothesis is accepted. A hypothesis is either rejected or not rejected.

There are 5 steps in this type of hypothesis testing which are:-

Step 1: State the Null Hypothesis and the Alternative Hypothesis as:

The Null Hypothesis	H_0:	$\mu_0 = 3.47$
The Alternative[41] Hypothesis	H_1:	$\mu_1 > 3.47$

Note that this is a one-tailed test. We are only interested in the right side of the distribution probability function and this is because the Alternative Hypothesis H_1 is μ_1>3.47. This test is sometimes referred to as direc-

[40] Statistical packages such as SPSS always compute the t-statistic and its associated probability as the t-distribution which accommodates all sample sizes. For large sample sizes the t-distribution and the normal distribution can be considered identical.

[41] The Alternative Hypothesis is sometimes called the Research Hypothesis. Some researchers refer to the Null as the "Ho" and the Alternative as the "Ha".

tional. If the Alternative Hypothesis H_1 is $\mu \neq 3.47$ then we would have to use a two-tailed test. The Alternative Hypothesis H_1 is $\mu_1 > 3.47$ or $\mu_1 < 3.47$.

The difference between the use of hypotheses with regards to one and two-tailed tests will be discussed in some detail later.

Step 2: Establish the level of significance of $\alpha = 5\%$. This is the traditional level of significance used in social science. Other levels of significance are used in other branches of science.

Step 3: Decide on the test statistic, the level of significance and determine whether it is a one-tailed or two-tailed test.

The t-statistic is chosen because the sample size is less than 30 and the value of σ is unknown. This is a one-tailed test as we are interested in an alternative hypothesis that $\mu_1 > 3.47$ which indicates a bias in one direction.

Level of significance $\alpha = 0.05$ or 5%

Step 4: The test statistic is calculated using the formula previously described.

Sample mean = 4.2
μ (the hypothesis mean) = 3.47
Sample standard deviation = 2
Sample size (n) = 20
Degrees of freedom (n-1) = 19

$$t_{19} = \frac{(4.2 - 3.47)}{\left(\frac{2}{\sqrt{20}}\right)}$$

$$t_{19} = 1.63$$

Calc-t_{19} = 1.63

Using the t-tables on page 101 the theoretical value is 1.73 (Table-t). This needs to be compared to the Calc-t.

Note that different authors and teachers of statistics use different language to describe the calculated t and the t which is obtained from the tables. Here we use Calc-t when the number is produced by using the

formula and we use Table-t when the number is produced by looking at tables. This number from the tables is also referred to as the theoretical value of t and also as the critical value of t. In general, this statistic will be referred to as Critical-t in this book.

One of the benefits of using Excel is to avoid the use of the t-Tables. Excel will produce the number required faster and more accurately. When the Excel generated t value is used we have referred to it as the critical t.

Step 5: Compare the test statistic Calc-t = 1.63 to the Critical-t = 1.73.

The rule is that if the absolute value of Calc-t >Critical-t[42] then reject the Null Hypothesis, otherwise do not reject the hypothesis.

A table of t-values is provided in Figure 2.22. These are sometimes referred to as theoretical values.

df \ α	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	df \ α
1	3.077685	6.313749	12.70615	31.82096	63.6559	127.3211	318.2888	1
2	1.885619	2.919987	4.302656	6.964547	9.924988	14.08916	22.32846	2
3	1.637745	2.353363	3.182449	4.540707	5.840848	7.4532	10.21428	3
4	1.533206	2.131846	2.776451	3.746936	4.60408	5.59754	7.17293	4
5	1.475885	2.015049	2.570578	3.36493	4.032117	4.773319	5.893526	5
6	1.439755	1.943181	2.446914	3.142668	3.707428	4.316826	5.207548	6
7	1.414924	1.894578	2.364623	2.997949	3.499481	4.029353	4.785252	7
8	1.396816	1.859548	2.306006	2.896468	3.355381	3.832538	4.500762	8
9	1.383029	1.833114	2.262159	2.821434	3.249843	3.689638	4.29689	9
10	1.372184	1.812462	2.228139	2.763772	3.169262	3.581372	4.143658	10
11	1.36343	1.795884	2.200986	2.718079	3.105815	3.496607	4.024769	11
12	1.356218	1.782287	2.178813	2.68099	3.054538	3.428451	3.929599	12
13	1.350172	1.770932	2.160368	2.650304	3.012283	3.372479	3.852037	13
14	1.345031	1.761309	2.144789	2.624492	2.976849	3.325695	3.787427	14
15	1.340605	1.753051	2.131451	2.602483	2.946726	3.286041	3.732857	15
16	1.336757	1.745884	2.119905	2.583492	2.920788	3.251989	3.686146	16
17	1.333379	1.739606	2.109819	2.56694	2.898232	3.222449	3.645764	17
18	1.330391	1 734063	2.100924	2.552379	2.878442	3.196583	3.610476	18
19	1.327728	1.729131	2.093025	2.539482	2.860943	3.1737	3.579335	19
20	1.325341	1.724718	2.085962	2.527977	2.845336	3.1534	3.551831	20
21	1.323187	1.720744	2.079614	2.517645	2.831366	3.13521	3.527093	21
22	1.321237	1.717144	2.073875	2.508323	2.818761	3.118839	3.504974	22

Figure 2.22: A published table of one-tailed t-values for specified levels of significance and degrees of freedom

[42] The value of Table-t which is here extracted from a set of t-tables is also obtainable by using the appropriate function in Excel and it is sometimes referred to as the critical value of t.

Note that this table requires the use of degrees of freedom (dof or df or v) which are calculated as the sample size minus 1. The degrees of freedom are shown in Figure 2.22 in the first column to the left (v) and the level of significance is the first row in the same figure. The level of significance equates to the area under the curve and hence the letter A is used in the first row in Figure 2.22.

2.10 The =tinv() function

To calculate the Critical-t value used in the hypothesis testing example above the Excel function =tinv() is used.

=tinv() is used to find the t-value as a function of the probability and the degrees of freedom.

The format of the function is =tinv(probability, df)
Where

probability=the probability associated with the two-tailed t test by default.
df = the degrees of freedom which is equal to n-1

The formula in cell B11 in Figure 2.23 uses =tinv()

We can determine the theoretical (Critical-t) value at a specified level of significance, e.g. for a significance level of 5% the theoretical Critical-t value for the above example is 1.73. This was produced by entering =tinv(2*B5,B3-1). As the Calc-t value 1.63 is less than the Critical-t value 1.73 the Null Hypothesis **cannot** be rejected.

The rule is….**if the absolute value of Calc-t >Critical-t then reject.**

Note that the t-statistic has been typically used for small samples i.e. samples containing less than 30 data points. However, for large samples (i.e. with 30 or more data points) the t-distribution and the normal distribution are almost identical. Thus the t-distribution is sometimes used across the whole range of values.

	A	B	C	D	E	F	G
1	Population mean	3.47					
2	Sample mean	4.2					
3	Count	20					
4	Standard dev.	2					
5	Significance level	5%					
6							
7	t calculated	1.632					
8	t critical value	1.729					
9	Decision	Do not reject null hypothesis					
10							
11	B7 = ABS(B2-B1)/(B4/SQRT(B3))						
12	B8=TINV(2*B5,B3-1)						
13	B9=IF(B7<B8,"Do not reject null hypothesis","Reject null hypothesis")						
14							
15	As Excel produces two-tailed t-critical values by default,						
16	a one-tailed t-critical value can be returned by replacing						
17	the probability with 2*probability.						
18							

Figure 2.23: Using =tinv()

2.11 More examples using the t-statistic

Example 7

To further illustrate these principles some additional examples are now provided.

Surveys over previous years have revealed that satisfaction with working conditions at the University was normally distributed with a value of 4.1 on a scale of 1 to 7.

As a result the University has introduced several new initiatives regarding pay and benefits over the past 2 years.

A new survey of 36 faculty members has been performed and the current average satisfaction score for the faculty was 4.5. The standard deviation for this sample is 1.9.

Does the new survey suggest that there is a significant increase in the level of satisfaction? You want to work at a level of confidence of 95% which is the reciprocal of a level of significance of 5%.

Steps in the Solution

Step 1: State the Null Hypothesis and the Alternative Hypothesis

The Null Hypothesis H_0: $\mu=4.1$
The Alternative H_1: $\mu>4.1$

Step 2: Decide on the test statistic, the level of significance and determine whether it is a one-tailed or two-tailed test.

The t-statistic is chosen although the sample size is greater than 30 but the value of σ is unknown. This is a one-tailed test as we are interested in an alternative hypothesis that μ>4.1 which indicates a bias in one direction.

Level of significance α = 0.05 or 5%

Step 3: Calculate the value of t.

$$t_{n-1} = \frac{\left(\overline{X} - \mu_0\right)}{\frac{s}{\sqrt{n}}}$$

$$t_{35} = \frac{\left(4.5 - 4.1\right)}{\frac{1.9}{\sqrt{36}}}$$

$$t_{35} = \frac{0.4}{0.32}$$

$$\text{Calc - } t_{35} = 1.25$$

Step 4: Use the test statistic (remember the degrees of freedom is the sample size (36) minus 1).

The rule is….**if the absolute value of Calc-t >Critical-t then reject.**

From t-tables (see Appendix 2) or from Excel we see that t_{35} (α = 0.05) = 1.69. As our calculated t=1.25 the Null Hypothesis cannot be rejected at the 5% level of significance. Therefore, there does not seem to have been any improvement in the faculty's perception of working conditions.

Excel can be used to perform the calculations in this example, as shown in Figure 2.24. In this spreadsheet the data supplied in this example are entered in column B from row 1 to 5. These are the population mean of 4.1, the sample mean of 4.5, the sample size (which is called the count in Excel) of 36, the standard deviation of 1.9 and the reciprocal of the level of significance (called the confidence level) which is 95%.

	A	B	C	D	E	F	G
1	Population mean	4.1					
2	Sample mean	4.5					
3	Count	36					
4	Standard dev.	1.9					
5	Significance level	5%					
6							
7	t calculated	1.263					
8	t critical value	1.69					
9	Decision	Do not reject null hypothesis					
10							
11	B7 = ABS(B2-B1)/(B4/SQRT(B3))						
12	B8=TINV(2*B5,B3-1)						
13	B9=IF(B7<B8,"Do not reject null hypothesis","Reject null hypothesis")						
14							

Figure 2.24: Excel spreadsheet for *Example 7* designed to perform a t-test at the 5% level of significance.

The calculated t-value is shown in cell B7 and is 1.263.

The rule now becomes…. if the absolute value of Calc-t >Critical-t then reject.

Note the term probability in Excel is synonymous with the term level of confidence, which in turn is related to the level of significance. Also note that the **=tinv() performs a 2-tailed test** and therefore it is necessary to multiply the probability by 2 before it works in a 1-tailed environment. The t-critical value is shown in cell B8 and is 1.690.

In cell B9 the =if() function is used to interpret the results of the calculation and the table value. Note the format of the =if() in excel is

=if(logical_test,value_if_true,value_if_false)

It is useful to now consider whether the Null Hypothesis could be rejected at the 10% level. From the tables or from the computer we see

that t_{35} (α=0.10) = 1.306. Therefore the Null Hypothesis cannot be rejected at the 10% level of significance either.

The Excel calculations are shown in Figure 2.25:

	A	B	C	D	E	F	G
1	Population mean	4.1					
2	Sample mean	4.5					
3	Count	36					
4	Standard dev.	1.9					
5	Significance level	10%					
6							
7	t calculated	1.263					
8	t critical value	1.306					
9	Decision	Do not reject null hypothesis					
10							
11	B7 = ABS(B2-B1)/(B4/SQRT(B3))						
12	B8=TINV(2*B5,B3-1)						
13	B9=IF(B7<B8,"Do not reject null hypothesis","Reject null hypothesis")						
14							

Figure 2.25: Excel spreadsheet for example number 7 designed to perform a t-test at the 10% level of significance.

Example 8

Surveys over previous years have revealed that the mean satisfaction with working conditions at the University was 4.1 on a scale of 1 to 7.

As a result the University has introduced several new initiatives regarding pay and benefits over the past 2 years.

A new survey of 36 faculty members has been performed and the current average satisfaction score for the faculty was 5.5. The standard deviation for this sample is 1.9.

Does the new survey reveal that there is a significant increase in the level of satisfaction? You want to work at a significance level of 5%.

Steps in the Solution

Step 1: State the Null Hypothesis and the Alternative Hypothesis

Null Hypothesis H_0: μ=4.1 and
Alternative Hypothesis H_1: μ>4.1

Step 2: Decide on the test statistic, the level of significance and determine whether it is a one-tailed or two-tailed test.

The t-statistic is chosen as the sample size is greater than 30 but the value of σ is unknown. This is a one-tailed test as we are interested in an alternative hypothesis that μ>4.1 which indicates a bias in one direction.

Level of significance α = 0.05 or 5%

Step 3: Calculate the t.

$$t_{n-1} = \frac{\left(\overline{X} - \mu_0\right)}{\frac{s}{\sqrt{n}}}$$

$$t_{35} = \frac{(5.5 - 4.1)}{\frac{1.9}{\sqrt{36}}}$$

$$t_{35} = \frac{1.4}{0.32}$$

$$\text{Calc - } t_{35} = 4.4$$

Step 4: Use the test statistic (remember the degrees of freedom is the sample size 36-1=35).

The rule is**....if the absolute value of Calc-t >Critical-t then reject**.

From t-tables (page 100) or from the computer we see that t_{35} (α=0.10) = 1.690. As our Calc-t = 4.4 the Null Hypothesis can be rejected. Therefore there seems to have been an improvement in the faculty's perception of working conditions.

Figure 2.26 shows how this example would be calculated using Excel.

	A	B
1	Population mean	4.1
2	Sample mean	5.5
3	Count	36
4	Standard dev.	1.9
5	Significance level	5%
6		
7	t calculated	4.421
8	t critical value	1.690
9	Decision	Reject null hypothesis
10		
11	B7 = ABS(B2-B1)/(B4/SQRT(B3)	
12	B8 = TINV(2*B5,B3-1)	
13	B9 = IF(B7<B8,"Do not reject null	
14	hypothesis","Reject null hypothesis")	
15		

Figure 2.26: Excel spreadsheet for *Example 8*

2.12 Paired samples t-test to compare means

The paired samples t-test is used to compare the values of means from two related samples. One of the most important applications of this is the 'before and after' scenario. In such cases the difference between the means of the samples is not likely to be equal to zero as there will be sampling variation. The Hypothesis test will answer the question "Is the observed difference sufficiently large to reject the Null Hypothesis?"

In Figure 2.27 below the examination scores for 10 candidates have been recorded before and after the candidates participated in a revision course. The scores are shown in Rows 3 and 4 of the spreadsheet. It is important to note that under these circumstances the data is not considered independent.

Does this evidence suggest that the revision course had an effect on the performance of the candidates?

Note that in this case we are looking at the same sample before and after an intervention.

Steps in the Solution

Step 1 – Pair the data: Subtract the scores before and after the intervention (revision course). Average the differences and calculate their standard deviation. This is shown as the Difference in Figure 2.27 below.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Candidates	1	2	3	4	5	6	7	8	9	10	
2	Score before	38	41	52	27	18	19	14	50	38	40	
3	Score after	40	45	49	30	24	24	19	49	36	39	
4	Difference	2	4	-3	3	6	5	5	-1	-2	-1	
5												

Figure 2.27: The data is paired

d is defined as the "score after" minus the "score before".
i.e. $d = X_{before} - X_{after}$

$\overline{d} = 1.8$ where $\overline{d}$ = mean of the values of d

$s_d = 3.293$ where s_d = standard deviation of the values of d

Step 2: State the Null Hypothesis and the Alternative Hypothesis

The Null Hypothesis H_0: $\mu_d = 0.0$

The Alternative Hypothesis H_1: $\mu_d > 0.0$

Step 3: Decide on the test statistic, the level of significance and determine whether it is a one-tailed or two-tailed test.

The t-statistic is chosen because the sample size <30. This is a one-tailed test as we are interested in an alternative hypothesis that $\mu_d > 0.0$ which indicates a bias in one direction.

Level of significance α = 0.05 or 5%

Step 4: Calculate the t-statistic.

$$t_{n-1} = \frac{\left(\overline{X} - \mu_0\right)}{\frac{s}{n}}$$

$$t_{n-1} = \frac{\left(1.8 - 0\right)}{\frac{3.293}{\sqrt{10}}}$$

$$\text{Calc - } t_9 = 1.73$$

Step 5: Use the test statistic (remember the degrees of freedom is the sample size (10) minus 1)).

The rule is**....if the absolute value of Calc-t >Critical-t then reject**.

From Excel (or from the tables on P101) we see that Critical-t_9=1.833. As our calc-t = 1.73 the Null Hypothesis cannot be rejected. Therefore it seems that there is insufficient evidence to suggest that the revision course was effective.

The Excel calculations are shown in Figure 2.28.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Candidates	1	2	3	4	5	6	7	8	9	10	
2	Score before	38	41	52	27	18	19	14	50	38	40	
3	Score after	40	45	49	30	24	24	19	49	36	39	
4	Difference	2	4	-3	3	6	5	5	-1	-2	-1	
5												
6	Average difference	1.8										
7	Standard dev.	3.293										
8	Count	10										
9	Significance level	5%										
10												
11	t calculated	1.728										
12	t critical value	1.833										
13	Decision	Do not reject null hypothesis										
14												
15	B6=AVERAGE(B4:K4)											
16	B7=STDEV(B4:K4)											
17	B11 = ABS(B6-0)/(B7/SQRT(B8))											
18	B12=TINV(2*B9,B8-1)											
19	B13=IF(B7<B8,"Do not reject null hypothesis","Reject null hypothesis")											
20												

Figure 2.28: Excel spreadsheet showing hypothesis testing for a paired t-test

2.13 The t-test for independent samples

In most cases, samples will be independent rather than paired, and if we want to test the significance of the difference between their means, we must use a different method to the one presented above.

For example, the senior management of University A claim that on average their reward package is better than the reward package offered by their competitor University B. A random sample of 20 employees from University A and a random sample of 30 employees from University B were asked to score on several dimensions the reward scheme offered to them and the following scores were recorded as shown in Figure 2.29.

University A scores		University B scores		
121	125	125	122	135
120	120	125	130	130
119	125	140	130	130
100	140	90	135	135
128	120	140	150	130
122	120	100	130	150
119	87	90	145	89
115	93	100	125	140
125	126	110	128	130
125	120	135	128	80

Figure 2.29: Independent data which is not paired i.e. two different samples of different sizes[43].

Does this evidence substantiate University A's claim? You should work at a significance level of 5%[44].

Steps in the Solution

Step 1: Calculate the mean, the standard deviation and the variance (the standard deviation squared i.e. S^2) for each sample.

$\overline{X}_1 = 118.5$ $S_1 = 12.15$ $S_1^2 = 147.63$

$\overline{X}_2 = 124.23$ $S_2 = 17.71$ $S_2^2 = 350.19$

Step 2: State the Null Hypothesis and the Alternative Hypothesis.

We will test the Null Hypothesis that there is no difference between the mean scores of the reward packages offered, against the alternative hypothesis that University A's reward package is better than University B's reward package:

Null hypothesis H_0: $\mu_1 = \mu_2$

Alternative hypothesis H_1: $\mu_1 > \mu_2$

[43] For two samples to be independent they do not have to be of different sizes.

[44] This test using independent samples relies on an assumption of equal variances and these need to be tested for in advance by means of the F-test. If both samples have more than 30 data elements this is unnecessary.

As we are in fact assuming that both samples have been drawn from the same population, the estimated variance of this population (or pooled variance) is:

$$\hat{S}^2_{pooled} = \frac{(n_1 - 1) * {s_1}^2 + (n_2 - 1) * {s_2}^2}{n_1 + n_2 - 2}$$

$$\hat{S}^2_{pooled} = \frac{(20 - 1) * 147.63 + (30 - 1) * 350.19}{20 + 30 - 2}$$

$$\hat{S}^2_{pooled} = 270.01$$

Step 3: Decide on the test statistic, the level of significance and determine whether it is a one-tailed or two-tailed test.

The t-statistic is chosen as the value of α is not known and S is used as an estimate. This is a one-tailed test as we are interested in an alternative hypothesis that $\mu_1 > \mu_2$ which indicates a bias in one direction.

Level of significance α = 0.05 or 5%
Here the df is (n_1+n_2-2).

Step 4: Calculate the t.

$$t_{(n_1+n_2-2)} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\hat{S}^2_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$t_{(20+30-2)} = \frac{118.5 - 124.23}{\sqrt{270.01\left(\frac{1}{20} + \frac{1}{30}\right)}}$$

$$\text{Calc-}t_{48} = -1.209$$

Step 5: Use the test statistic (remember the degrees of freedom).

The rule is**....if the absolute value of Calc-t >Critical-t then reject**.

From t-tables on Page 101 or from the computer we see that Critical-t_{48} (α=0.05) = 1.677. As our calc-t = 1.209 the test statistic is not significant and the Null Hypothesis cannot be rejected. Therefore there is insufficient evidence to suggest that the University A's reward package is better. The calculations in Excel are shown in Figure 2.30 below.

	A	B	C	D	E	F	G
1		A	B		A	B	
2	Mean	118.50	124.23		121	125	
3	Standard dev.	12.15	18.71		120	125	
4	Significance level	5%			119	140	
5	t-test: Two samples assuming equal variances				100	90	
6		A	B		128	140	
7	Observations	20	30		122	100	
8	Variance	147.63	350.19		119	90	
9	Pooled variance	270.01			115	100	
10	Hypothesized mean difference	0			125	110	
11	df	48			125	135	
12	t calculated	1.209			125	135	
13	t critical value	1.677			120	130	
14	P(T<=t) one-tail	0.116			125	130	
15					140	135	
16	Decision	Do not reject null hypothesis			120	130	
17					120	150	
18	B7=COUNT(E2:E31)				87	89	
19	B8=VAR(E2:E21)				93	140	
20	B9=((B7-1)*B3^2+(C7-1)*C3^2)/((B7-1)+(C7-1))				126	130	
21	B11=B7+C7-2				120	80	
22	B12=ABS(B2-C2)/SQRT((B9*(1/B7+1/C7)))					122	
23	B13=TINV(2*B4,B11)					130	
24	B14=TDIST(B12,B11,1)					130	
25						135	
26	Notice that the two samplemean values (variance)are					150	
27	118.5(147.63 and 130.33(261.95). The one-tailed calculated					130	
28	t-statistic is 1.209 and the highlighted p-value for this test is					145	
29	P=0.116. Since the p-value is higher than 0.05, this does					125	
30	not provide evidence to reject the null hypothesis of equal					128	
31	means at the 5% level of significance.					128	
32							

Figure 2.30: Excel spreadsheet with hypothesis testing for independent samples.

2.14 Right-tailed and Left-tailed hypothesis tests

All the examples above used only one-tailed t tests. These tests took the form of Null Hypothesis H_0: $\mu1 = \mu2$ and Alternative Hypothesis H_1: $\mu1 > \mu2$. These tests are referred to as right-tailed tests. If the tests had the form of Null Hypothesis H_0: $\mu1 = \mu2$ and Alternative Hypothesis H_1: $\mu1 < \mu2$ then the tests are referred to as left-tailed tests.

Figure 2.31 shows a graph of a right-tailed test and Figure 2.32 shows a graph of a left-tailed test.

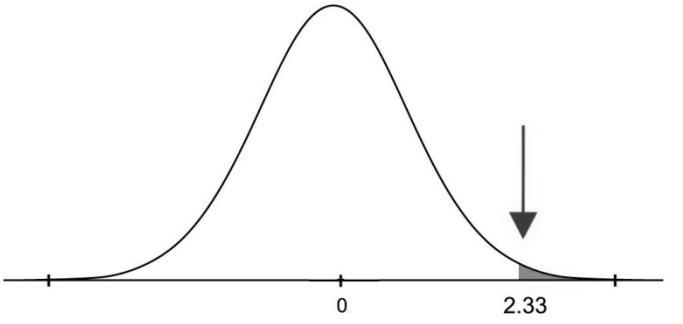


Figure 2.31: A right-tailed test

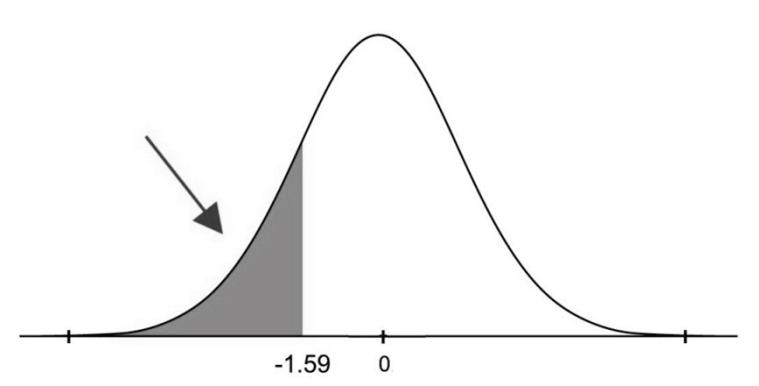


Figure 2.32: A left-tailed test

2.15 Two-tailed Hypothesis tests

A two-tailed test is carried out when the researcher is looking for any change from an expected result, not specifically an increase or a decrease.

These tests take the form of a Null Hypothesis H_0: $\mu 1 = \mu 2$ and the Alternative Hypothesis H_1: $\mu 1 \neq \mu 2$. These tests are referred to as two-tailed tests.

If the significance level is α %, then the critical region is in two parts, half in the lower tail and half in the upper tail.

Figure 2.33 shows that the calculated statistic could be either larger or smaller than the table/Excel statistic.

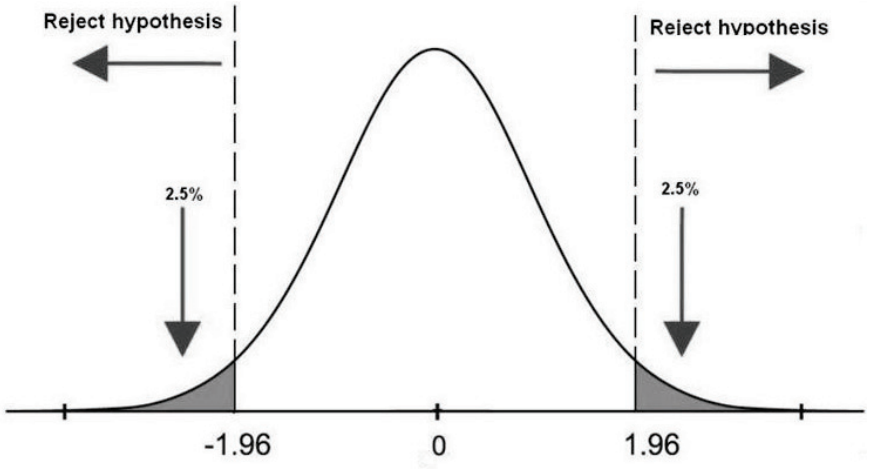


Figure 2.33: A two-tailed test where Z-values 1.96 and -1.96 define the critical regions (shaded in grey above) with an α of 0.05 that is 0.025 in each tail.

As an example, in the past the mean performance score for employees was 60 on a 100 point scale[45]. We have introduced a new controversial training programme which has been opposed by some unions and we wish to establish whether this has had any impact on the employees' performance. We take a single sample of 25 observations with a mean score of 55 and a standard deviation of 15.

We want to test to see if the new mean is significantly different from the original at the 5% level of significance.

Steps in the solution

Step 1: Research question is….*Is the new mean significantly different to the old mean?*

State the Null Hypothesis and the Alternative Hypothesis

Null Hypothesis	H_0:	$\mu = 60$ and
Alternative Hypothesis	H_1:	$\mu \neq 60$

Step 2: Decide on the test statistic, the level of significance and determine whether it is a one-tailed or two-tailed test.

[45] We are assuming that the true standard deviation, i.e. the population standard deviation is not known.

The t-statistic is chosen as the sample size is less than 30 and the value of σ is unknown. This is a two-tailed test as we are interested in an alternative hypothesis where μ≠60.

Level of significance α = 0.05 or 5%.

Step 3: Calculate the t-statistic.

$$t_{n-1} = \frac{\left(\overline{X} - \mu_0\right)}{\frac{s}{\sqrt{n}}}$$

$$t_{24} = \frac{(55 - 60)}{\frac{15}{\sqrt{25}}}$$

$$Absolute \quad Calc - t_{24} = -1.67$$

When consulting Critical-t we can see that it only provides the area in the right hand tail of the distribution. Thus as we are using a two-tailed test with a 5% significance, we need to look for 2.5% in either tail. This means that we consult the t-tables for df=24 and 2.5%. This returns a value of 2.06.

Thus use the rule ... **If the absolute value of Calc-t>Critical-t then reject**.

Step 4: Since the absolute value of -1.67 is < Critical-t, we do not reject the Null Hypothesis. This is translated into the fact that the new controversial training programme does not appear to have had any significant impact on performance.

Figure 2.34 shows the above example in Excel.

	A	B
1	Sample mean	55
2	New mean	60
3	Count	25
4	Standard dev.	15
5	Significance level	5%
6		
7	t calculated	1.667
8	t critical value	2.064
9	Decision	Do not reject null hypothesis
10		
11	B7 = ABS(B2-B1)/(B4/SQRT(B3))	
12	B8 = TINV(B5,B3-1)	
13	B9 = IF(B7<B8,"Do not reject null	
14	hypothesis","Reject null hypothesis")	
15		

Figure 2.34: Two-tailed test

Note that the formula in cell B8 does not require the probability to be multiplied by 2 as Excel defaults to a two-tailed test.

2.16 The use of P-values

The P-value, which is increasingly supplied as an integral part of computer reports is effectively a short cut for hypothesis testing. The P-value is a probability value and thus it has to be between 0 and 1.

The P-value associated with a test is the probability that we obtain the observed value of the test statistic or a value greater in the direction of the alternative hypothesis calculated when the Null Hypothesis is true (has not been rejected).

The P-value is an estimate of erroneously rejecting the Null Hypothesis on the basis of the sample data. This applies to both one and two-tailed tests.

In the case of a one-tailed test we determine the area under the t-distribution in the appropriate direction. To do this we determine the value from the tables.

For the previous example involving a two-tailed test we can now determine the P-value as follows.

Steps in the solution

The required P-value is the area $A_1 + A_2$ under the curve as shown in Figure 2.35. This includes the area to the left of the calculated t-value -1.67, but also the area to the right of 1.67 as the test is two-tailed.

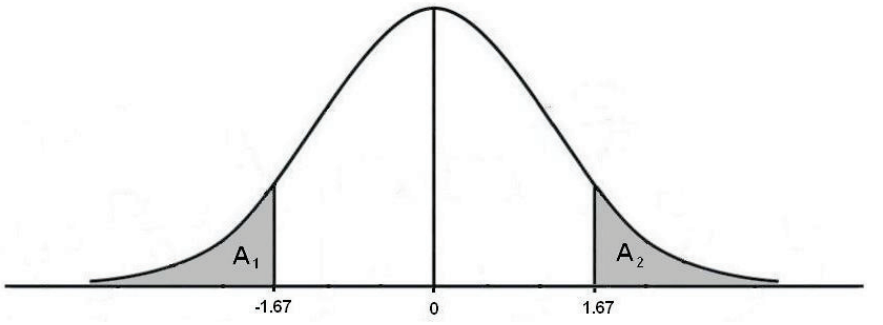


Figure 2.35: Required area for the two-tailed test when t-calc = $\pm$ 1.67.

As the graph is symmetrical, the required area $A_1 + A_2 = 2^* A_2$. We now calculate the area A_2 as follows using linear interpolation.

Step 1: From the t-tables on p.101, it is noted that it is not possible to find an exact area for the number 1.67 with 24 degrees of freedom.

Step 2: Linear interpolation

1.67 lies between 1.3278 (representing an area of 10%) and 1.7109 (representing an area of 5%) as shown in Figure 2.36.

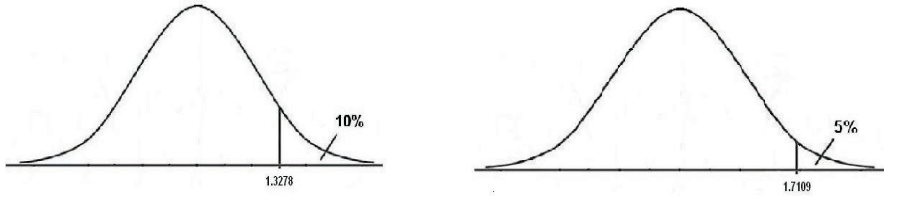


Figure 2.36: Areas under the curve representing 10% and 5%

The difference between these areas = 10% - 5% = 5%

Step 3: The area between 1.3278 and 1.67 is now calculated as a fraction of 5% (.05).

$$Area = \frac{1.67 - 1.3178}{1.7109 - 1.3178} * .05$$
$$= .04479$$
$$= 4.479\%$$

This value is represented by the shaded area in Figure 2.37.

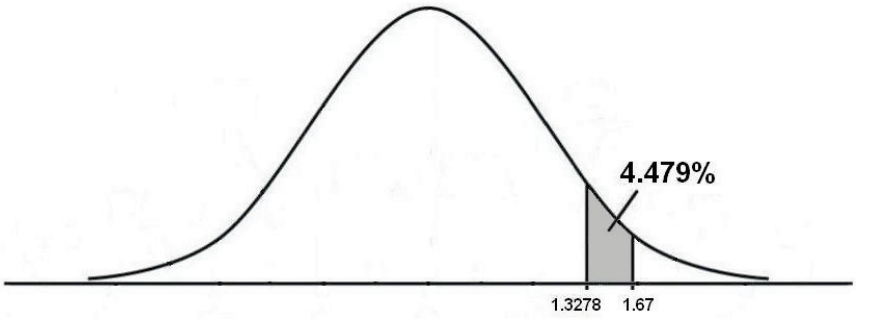


Figure 2.37: Area under the curve between 1.378 and 1.67.

Step 4: Finally, the required area A_2 (to the right of 1.67) is now obtained by subtracting 4.479% from 10% (area to the right of 1.3278).

$$A_2 = 10.00 - 4.479 = 5.521\%$$

This corresponds to a P-value of 0.5521

Step 5: As we have a two-tailed test, the associated P-value will be 2* A_2 = 2 * .05521 = .1104 or 11.04%.

Step 6: Comparing the P-value with the level of significance 0.05 (5%) and as the P-value is greater than the level of significance we conclude that there is not support for rejecting the Null Hypothesis. If the hypothesis is rejected then there is a one in eleven chance that this decision is incorrect, which is higher than the one in 20 (P=0.05) that we are working with in this example and which is traditionally used.

The P-value is another way of thinking about statistical significance. The P-value gives us the level of confidence with which we can reject the Null Hypothesis. If we were to reject the Null Hypothesis in this example we would expect to find that 11 times out of 100 we would be in error. One way of recalling the application of the P-value is to remember *If the P-value is low then the hypothesis must go!*

P-values have only been in regular use in recent years. The reason for this is that it was formerly believed that the significance level should be decided for a test before the data analysis was performed. The reason for this was that the result of the test should be decided exclusively on the pre-chosen significance level. It was believed that if the P-value was known and the results of the test were marginal, researchers might change the significance level to suit their objectives and comply with the data.

2.17 A test for the normal distribution

In this book we have referred to data being normally distributed which is an assumption we need to make in order to be able to use the Z-test and the t-test. However we have not yet addressed the issue of how one might know if it is reasonable to assume that a given data set is normally distributed. We will use two different approaches to consider this matter. The first is the shape of the graph of the data and the second is the value of a series of statistics including the coefficient of skewness and kurtosis.

Data distributions which are skewed would not be considered normally distributed. Two examples of skewed data distributions are given in Figure 2.38 and Figure 2.39.

In Figure 2.38 the data distribution is skewed negatively which means that the mean is to the left of the median.

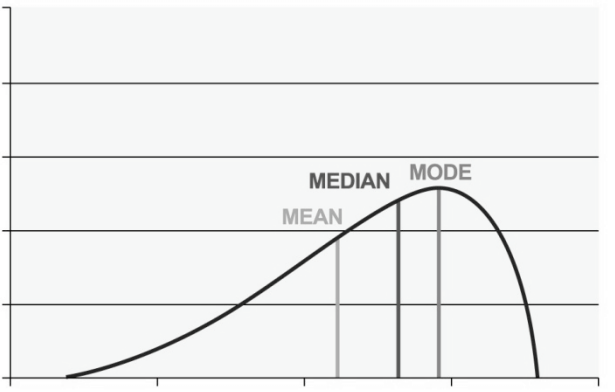


Figure 2.38: A negatively skewed distribution.

In Figure 2.39 the data distribution is skewed positively which means that the mean is to the right of the median.

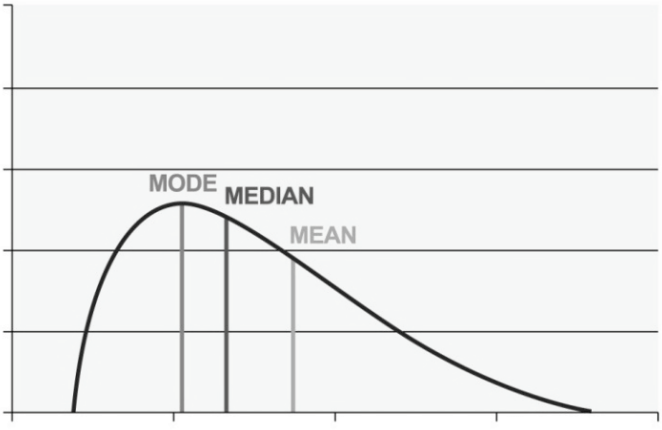


Figure 2.39: A positively skewed distribution.

The data as represented by these graphs could not be treated as normally distributed.

Consider the following set of 30 data points or elements.

	A	B	C	D	E	F	G	H	I	J
4	1	1	2	2	3	3	4	4	4	5
5	5	5	6	6	6	6	6	5	5	5
6	5	4	4	3	3	2	1	1	1	4

Figure 2.40: Sample of data

The first task is to graph this data set as a bar chart. To do this move the data into one row and then highlight this line and chose the **Bar Chart** option from the **Chart Wizard**. The bar chart will appear as shown in Figure 2.41.

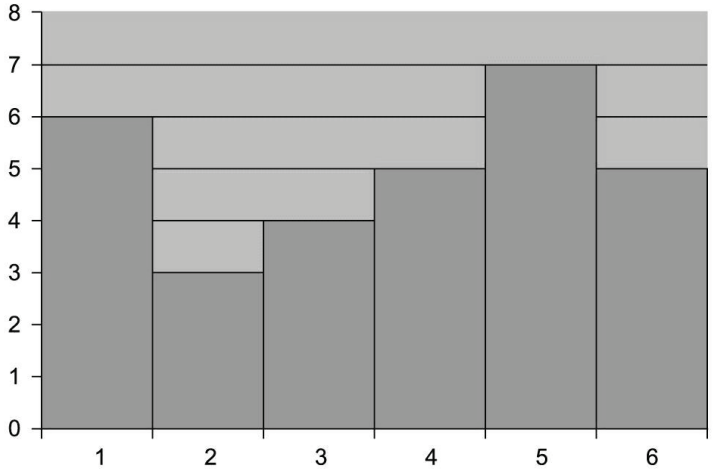


Figure 2.41: A bar chart of the data

One of the main characteristics of a normal distribution is that the data is distributed under a bell shaped curve which is symmetrical i.e. it will give an equal number of data points on each side of a central point which is determined by the mean. From Figure 2.41 it may be seen that the data is approximately bell shaped. There is no one central data point in this data and there are more data points to the right of the centre of the data than there are to the left.

By eyeballing the bar chart we can see that the distribution of the data is not perfectly normal. However, the bar chart is approximately normal and therefore we should look at some of the other indicators of a normal distribution.

The first statistics we consider are the mean, median and the mode.

Mean	3.9
Median	4
Mode	5

Figure 2.42: Measures of average

If a data distribution is normal then the mean, the median and the mode should be equal. Clearly this is not the case. However there is not a great difference between the values of these three statistics.

The next step is to calculate the coefficient of skewness and kurtosis.

Skewness	0.3
Kurtosis	1.3

Figure 2.43: Measures of shape

A normal distribution will have a skewness coefficient of 0 and thus the above score of 0.3 approximates a normal distribution.

A normal distribution will have a kurtosis coefficient of 1.3 if calculated in Excel and thus the above score of 1.3 is close to a normal distribution.

The question now is, although the above data set is not strictly normal, does it sufficiently approximate a normal distribution to allow the techniques that assume the normal distribution to be used?

In general the normal distribution techniques are often sufficiently robust to be usable with the above data. If it were decided that a more formal test for normality was required other techniques could be used.

In the social sciences, many scales and measures have scores that are positively or negatively skewed. This could reflect the underlying construct being measured as might be the case when respondents record their scores on life satisfaction questionnaire items giving rise to a negative skew for the distribution, with the majority of the respondents being reasonably happy with their situation in life. The incomes of salary earners are generally positively skewed with most earners at the lower end of the scale.

There are more formal ways of deciding if a data set is normally distributed. One of these includes the use of the chi-square test and this is beyond the scope of the book.

2.18 Summary-Part 2

This Part moves the learner from relatively simple statistical concepts towards the more powerful ideas and techniques. The material here introduces the learner to inferential statistics. This has been done by addressing data distributions and then moving on to the probability distribution which is most frequently used i.e. the normal distribution and the use of the Z score. Estimation, both point and interval, has also been addressed. The use of the t-distribution has been demonstrated.

The central concept of hypothesis testing is then examined and applied to paired t-tests and t-tests on independent samples.

The earlier section of this Part only addresses one-tailed tests and so towards the end, two-tailed tests are introduced as are the P-value and a test to establish if a data set is normally distributed.

This Part of the book relies on the use of a material number of worked examples as this is the way that learners will acquire the skills needed to be proficient with these techniques.

Readers are reminded to save their workings in Excel and to make paper copies or printouts where they require further backup.

List of Excel functions used for the first time in this Part of the book.

Function	Result
=frequency()	This is an array function which calculates how often values occur for a specified range of values.
=normdist()	Calculates the P-values for the Normal Distribution. This is a one-tailed probability.
=tdist()	Calculates the P-values for the t Distribution
=tinv()	Calculated the theoretical or table values from the t-Distribution. This is two-tailed.
=abs()	Returns the absolute or positive value of a calculation irrespective of the result of the calculation.

More details of these functions are provided in the Help command within the spreadsheet.

Self test 2

No.	Question	Answer
1	What is a data frequency table?	
2	What is a useful heuristic for calculating the width of a data interval for a frequency table?	
3	Give 4 characteristics of a normal distribution.	
4	What do we mean by a standardized normal distribution?	
5	Define the Z-score?	
6	What are normal distribution tables?	
7	Define the term test statistic.	
8	What is the Null Hypothesis?	
9	When can we say we have proved the Null Hypothesis?	
10	Explain the meaning of μ	

No.	Question	Answer
11	Explain the meaning of σ	
12	Explain the meaning of $\overline{X}$	
13	When is the t-test preferable to a Z-test?	
14	What is meant by paired t-tests?	
15	When is a t-test said to be independent?	
16	What does it mean to say that the hypothesis is always a claim?	
17	What does the theoretical value of t mean?	
18	What does a P-value of 0.05 mean?	
19	When is it necessary to use a two-tailed test?	
20	In what ways are normal distribution and the t-distribution similar and in what way do they differ?	

Assignment No 2

1. What is the difference between sample statistics and population parameters?
2. What do you understand by point estimation and interval estimation?
3. How do you interpret the confidence levels associated with interval estimation?
4. What do you understand by the term "level of significance"?
5. What is the difference between statistical significance and practical significance?
6. Given that a new process does not appear to have an adequate level of output how would you formulate and express a Null Hypothesis and an Alternative Hypothesis? How would you design such a study?
7. Explain the difference between a one-tailed and a two-tailed test.
8. A new approach to instructing students is adopted by the Institute. You are asked to design a test to establish if it has had an impact on the performance of the students. Is this going to be a one-tailed or a two-tailed hypothesis test?

Additional exercises

1. Display the data below as a frequency distribution and draw an appropriate graph of the frequency distribution.

	A	B	C	D	E	F	G	H	I	J
1	4	29	113	121	168	106	56	33	190	58
2	56	63	113	170	89	138	117	47	77	170
3	52	195	29	106	118	70	116	97	149	135
4	180	33	78	151	86	42	7	175	91	182
5	56	44	98	124	72	150	193	128	103	39
6	109	111	145	103	173	192	107	156	172	1
7	134	157	78	160	43	85	138	161	91	138
8	183	193	51	79	130	189	88	52	113	109
9	155	111	165	132	37	184	75	20	29	150
10	10	150	61	97	77	97	174	60	68	11
11	18	27	150	191	194	146	37	39	94	46
12	31	73	1	145	96	60	152	8	194	155
13	46	193	36	103	80	28	110	50	17	5
14	188	14	86	159	149	150	41	151	24	134
15	149	165	91	117	4	35	119	115	92	33
16										

2. What heuristic have you used to establish the width of the intervals you have used? Produce a second frequency distribution and graph using this data and comment on the difference between them.

3. Using the data supplied below comment on whether you would consider this data set to be normally distributed. Explain your answer.

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

4. The number of pupils completing their secondary education in the county each year is normally distributed with mean 14000 and standard deviation 750. What is the probability that 16000 will complete their secondary education this year?

5. Sean is in a large school where the scores for mathematics are normally distributed with a mean (μ) of 55 and a standard deviation (σ) of 10. Sean has passed his mathematics exam with a score of 80. Carole is from a similar school where the scores for mathematics are normally distributed but they take a different examination which has a mean of 30 and a standard deviation of 5. Carole's score is 42. Who has done better in the mathematics exams?

6. Explain the meaning of the Null Hypothesis (H_0) and the Alternative Hypothesis (H_1).

7. The Widget Company Limited has been making widgets for many years and its output is normally distributed. Historical records show that the mean weight of widgets is 350 and the standard deviation is 50. A new widget is produced with a weight of 450. With a significance level of 5% is there support to claim that the 450 widget belongs to the same group as the others?

8. A group of 10 students is asked to rate the quality of the food in the student refectory on a scale of 1 to 10 where 1 is very poor and 10 is excellent. The replies of the students are shown below. A new supervisor is appointed in the student refectory and he introduces a new menu as well as a more friendly attitude towards the students. The opinions of the same group of students are canvassed again and the results of the second set of interviews are also shown below. Is there evidence to support the hypothe-

sis that the new supervisor has improved the students' rating of the refectory? Work at a significance level of 5%.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Test before intervention	4	5	3	6	5	4	3	4	3	3
Test after interverntion	4	3	6	6	7	6	5	6	4	4

9. The IQ of physics graduates is recorded at two different universities. At the first university there are 20 physics graduates and at the second university there are 25 physics graduates this year. Using the data supplied below examine the hypothesis that the graduates from the first university have scored better in their IQ tests. Work at a significance level of 5%.

University A	123	170	130	160	173	170	142	126	148	143
	144	166	166	152	167	147	144	162	139	148
University B	152	169	156	175	169	136	173	158	135	139
	155	149	174	167	146	172	159	139	136	142
	155	146	169	172	172					

10. Explain why some researchers or statisticians believe that using the t-test for hypothesis testing is better than using the Z-score.

11. A factory produces widgets with a mean diameter of 160 mm and a standard deviation of 11mm. New technology has been introduced and is supposed to improve the accuracy of our process. We take a sample of 25 widgets with a mean of 155 and a standard deviation of 15. Has our process been improved? Work at a significance level of 5%.

12. Under what set of circumstances may a researcher say that he/she has proved his/her hypothesis?

13. Explain the importance of the P-value.

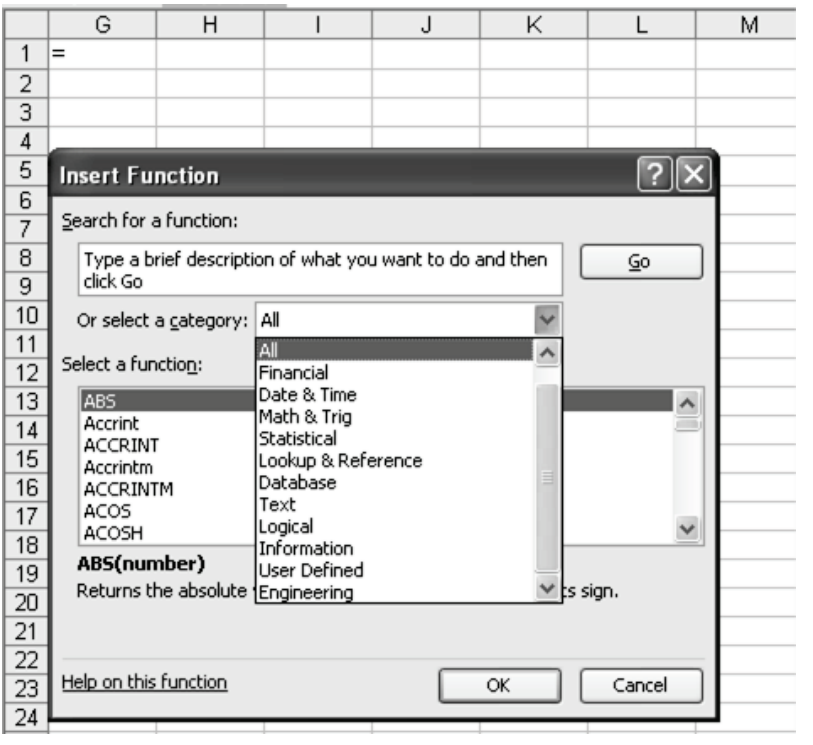
14. Formulate a Null Hypothesis and Alternative Hypothesis which will require a two-tailed test.

15. How might a researcher test a data set in order to establish if it is normally distributed?

A Note on Excel Functions

There are approximately 365 functions in Excel which are divided into 11 categories. In addition it is possible to purchase more in the form of add-ins which extend the power of the spreadsheet.

Excel provides some help in choosing the right function by using the Insert Function commands but to employ functions effectively you need to be acquainted with the mathematics behind the function.



The form of an Excel function is always **=FunctionName(Argument)** where the Argument may be a cell, a group of cells or other forms of data.