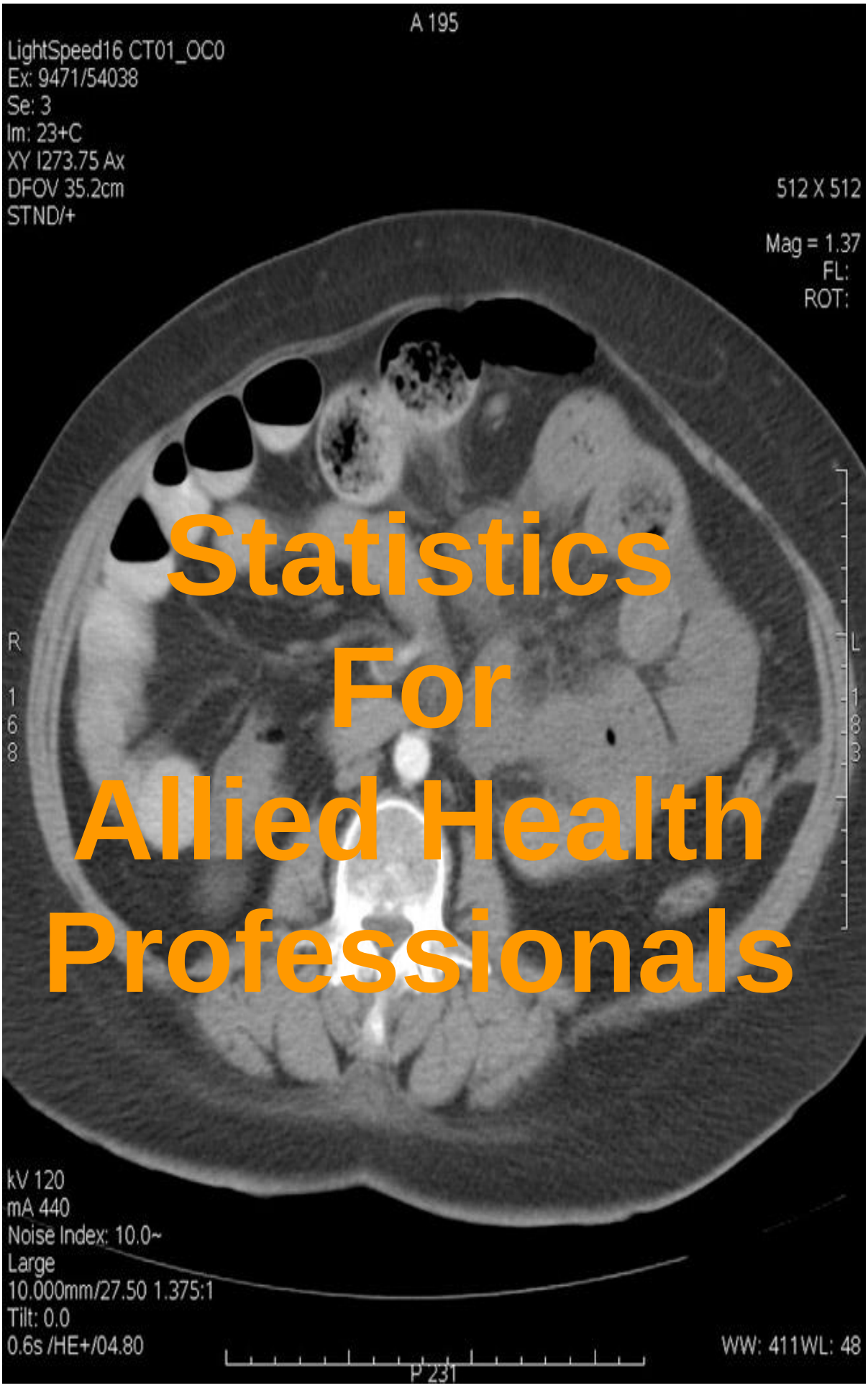


Statistics For Allied Health Professionals

CONTENTS

Contents

ABOUT THE AUTHOR

Dr Hartley is a Clinical Biochemist with a longstanding interest in statistics. HeI graduated with a B.Sc.(Hons) degree in Chemistry from the University of Manchester Institute of Science and Technology in 1969 and then took up a MRC funded research student position within the Medical Physics Department of the University of Leeds and the MRC Mineral Metabolism Unit. He was awarded the higher degree of PhD by the University of Leeds in 1974. The title of his thesis was 'The Metabolism of Copper, Zinc and the Four Major Cations in Man'.

Between 1972 and 1974, he carried out two years post-doctoral research at St Mary's Hospital in Portsmouth, UK, on a clinical trial of a new Total Parenteral Nutrition amino acid solution.

In 1974 he emigrated to Adelaide in South Australia, where he took up a post in the Clinical Chemistry Dept at the Institute of Medical and Veterinary Science. While there he extended his clinical nutrition, metabolic bone disease and statistical quality control work before moving to take up the post of Scientist Second in Charge of the Clinical Chemistry Dept. at the Royal Hobart Hospital in 1988.

He continues as a Senior Biochemist in the Department of Pathology at the Royal Hobart Hospital as well as being their Quality Manager since 2000.

In June 2004 he took up a 0.25 position in the School of Human Life Sciences, University of Tasmania, as a Senior Research Fellow.

He is the author of the book 'Computerised Quality Control – 2nd Ed' published by Ellis Horwood in 1990 as well as a large number of papers in scientific journals.

His research interests include nutritional biochemistry, trace elements, antioxidants, instrumental methods of analysis, laboratory computing and statistics.

Currently his three foci of interests are laboratory quality systems, biomedical statistics and the HPLC analysis of clinical samples for capsaicins. The latter involves the assay of capsaicins, the 'hot' components of chillis, in human plasma samples using HPLC.

Author's Contact Details

Dr Tom Hartley, Quality Manager and Senior Biochemist, Pathology Services, Royal Hobart Hospital, Hobart, Tasmania 7000, Australia

Dr Tom Hartley, Senior Research Fellow, School of Human Life Sciences, University of Tasmania, Launceston, Tasmania 7250, Australia.
Email Thomas.Hartley@utas.edu.au Phone +61 3 6222 8780
Website : www.medlabstats.com

PURPOSE OF THIS BOOK

Several years experience with teaching statistical concepts to a variety of Allied Health undergraduates has highlighted the need for a concise and approachable book on statistics. When these students graduate the aim is that they enter their professions with a

basic knowledge of project design, statistics and experience with the practical analysis of simple datasets using the readily available tools in Microsoft Excel. This book is designed so that the statistical concepts are described in the first half and the second half is devoted to illustrating how these statistical tests are accessed and used in Microsoft Excel. Where the latter does not proved the required tool then the reader is directed to www resources either on his own website or on others that have proved to be up to the job.

A FREQUENTLY ASKED QUESTIONS GUIDE TO USING THIS MONOGRAPH

Q : Why do I need to know about probability theory to understand statistics ?
Answer : Read Section 2.

Q : Researchers are always talking and writing about how their research was statistically significant. What do they mean by this?
Answer : Read Section 3.

Q : I don't usually measure anything on my patients but I do notice that they tend to fall into categories according to a few basic characteristics eg. women between 45 and 55 who are overweight usually present with or end up with Type 2 Diabetes. Is there a statistical method for confirming this kind of 'hunches'?
Answer : Read Sections 4, 5, 6 and 7

Q : When would I find it useful to plot a histogram of my data ?
Answer : Read Sections 8, 9 and 15.

Q : What kind of plots are useful in illustrating differences in patient outcomes or responses 'before' and 'after' treatments ?
Answer : Read Sections 8, 9 and 10.

Q : Researchers are always talking about their data being Normally Distributed. What do they mean by this ?
Answer : Read all of Section 11.

Q : If my data are Normally Distributed what statistical tests can I do on them ?
Answer : Read Sections 12, 13, 14, and 15.

Q : Researchers talk about being able to prove a cause and effect from their research. What do they really mean by this and how do they prove it ?
Answer : Read Sections 17, 18, 19, 20, 21, 22, 23 and24.

Q : In my work I suspect that the clinical effects I am observing are due to a combination of different causes. How would I go about proving this ?
Answer : Read Section 21 and 24.

Q : In my work I tend to find that my patients tend to fall into 'Low', 'Moderate" and 'High' risk categories. Are there a statistical method to deal more objectively with this kind of classification?
Answer : Read Sections 22 and 23.

Q : In my area of work we gather a lot of data from our patients using questionnaires. Is there a way of ensuring that are questionnaires are 'good' and 'unbiased' ?
Answer : Read Section 25.

Q : I have collected a lot of data on my patients. Is there a way of exploring that data for relationships that I probably have not realised are there ?
Answer : Read Section 26 first and see if that gives you some insights. Then you can also read Sections 21, 22, 23 and 24.

Q : I would like to apply for a research grant or a project grant but I am put off by the requirements to write up the statistical section.
Answer : Read Sections 27, 28.1 (if you are going to be making lab measurements) and 28.3 (if you are going to be making lab measurements of the same parameter using different instruments or methods). Once you have decided what your hypotheses are and then what types of data you are going to collect for analysis then you should read those sections that best fit your data types ie. choose the appropriate mix of parametric and non-parametric tests.

Q : I have time series data. What are the best statistical tests to use for analysing those kind of data ?
Answer read Sections 28.2 and 28.4

1: INTRODUCTION

Objectives in medical research are usually

to intervene for a beneficial effect
or
to intervene to determine the cause of a clinical symptom

The intervention is usually

an experimental drug
an experimental therapy
a new diagnostic test or investigation

The question always arises – how can we objectively decide that the intervention has been a success.

First we have to define what constitutes a success, a failure and an equivocal outcome.

Then a fairly intuitive next step would be to measure the frequencies of successes, failures and equivocal outcomes (neither successes or failures).

Then use statistics to analyze the numbers of successes, numbers of failures and number of equivocal outcomes against the hypothesis that these numbers are not related in any way to the intervention. This is the null hypothesis.

2: ELEMENTARY PROBABILITY THEORY

The simplest analogy is the tossing of a coin – it can land heads up or tails up. There is a 1/2 chance that there will be head and a 1/2 chance that there will be a tail. 1/2 equals 0.50. It is usual to use the letter p to signify probability and use the phrase …

the probability of a head is p=0.50
the probability of a tail is p=0.50

From these two phrases we can derive one of the fundamental rules of probability :

If there a n mutually exclusive events then
the sum of the individual probabilities must be ONE

$p_{HEAD} + p_{TAIL} = 0.50 + 0.50 = 1$

The next simplest analogy is a six sided dice – it can land with a 1, 2, 3, 4, 5 or 6 face up. This means that there is a 1/6 chance of landing with 1 facing up.

It follows from this that the value
of p for throwing a '1' is 0.1667,
of p for throwing a '2' is 0.1667,
of p for throwing a '3' is 0.1667,
of p for throwing a '3' is 0.1667,
of p for throwing a '4' is 0.1667,
of p for throwing a '5' is 0.1667
and of p for throwing a '6' is 0.1667.

And the p for throwing any number is 6 * 0.1667 = 1.000…. ! (So we know we have got our probability predictions right.)

---

3 : WHAT IS MEANT BY A 'SIGNIFICANT DIFFERENCE'.

In clinical research we are usually involved in activities in which we want to cause a change in a parameter – these fall under the general heading of intervention studies. In this scenario the rule is that we need to look at our statistical results with the objective of finding those parameters that have changed with a statistical p value equal to or less than 0.05 then we can be 95% confident that we have detected a significant difference. Many statistical tests report p values for a 'one tailed' and a 'two tailed' test. You can use the 'one tailed' test only in scenarios where it is clear that the parameter you have tested statistically could only go one way because of the study design eg. body weight will fall in a 'hunger strike' study! But in other studies you cannot be so dogmatic and you should then be more 'open minded' and opt for the two tailed p value. For example you may be studying a blood pressure lowering drug – the company assures you that based on their studies of 'monotherapy' it lowers blood pressure – but you want to use it in a more realistic scenario – use it in combination with other drugs usually co-prescribed to a hypertensive eg a diuretic. Under that scenario you cannot be dogmatic – it may well cause no change or an increase in blood pressure in some of your test subjects. So if in doubt always use the p value from a 'two tailed' test.

Other researchers undertake observational studies ie. non-intervention studies. Usually they have a question and they then go out and collect data that they reasonable expect to provide and answer to their question eg. Does socioeconomic status affect blood pressure ? They would usually start from the premise that socioeconomic status does not affect blood pressure – they take the 'null' hypothesis route. Under this scenario they are usually hoping that all their 'two tailed' statistical tests will return p values greater than 0.05. When this is observed they can be confident that there is a less than 5% chance that socioeconomic status affects blood pressure. But they may find, for example, within some of the parameters they have measured that there is a significant difference between say the number of cigarettes smoked between two socioeconomic status groups. Then if they look at the blood pressures in these two subgroups they may find that there are significant differences – p values equal to or less than 0.05. Regardless of the direction of the difference researchers involved in non-intervention studies must always apply 'two tailed' tests.

4: CHI SQUARE TEST

This is the test to use when we want to check that the **probabilities** that we observed in a clinical trial or a new clinical investigation support the hypothesis that the intervention **has improved the probabilities of clinical improvement** or diagnosis.

We tabulate our data in the following format :

TABLE 1

	TREATMENTS			
OUTCOMES counted as number of Patients	Physiotherapy only	Physio plus Antibiotics	Antibiotics only	ROW TOTALS
Improved				
No Change				
Deteriorated				
COLUMN TOTALS				Grand Total

The formula for the Chi Square Test is

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

This formula is saying *'for every observed frequency of an event (O) subtract the expected frequency of that event (E), square it and divide it by the expected frequency of that event'*. You then sum all those calculations to get Chi Squared.

How do you know what the expected frequency of the events are in each box of the table ? The rule is to multiply the row total by the column total and divide by the grand total. This is best left to the computer program you use for calculating Chi Squares.

What you do need to realise is that if the observed frequencies in each square equalled the expected frequencies in each square then the value of Chi Square would be ZERO…a result that would mean that your treatments had had NO EFFECT AT ALL.

In Medical Research we are usually hoping for large values of Chi Square – signifying our intervention has been significantly effective!

Statisticians have modelled this type of experiment with all possible outcomes and drawn up Statistical Tables which show with what probability a particular value of Chi Square can be expected to occur, See the Table in APPENDIX ONE

The only missing item that you need to calculate is the number of degrees of freedom for your study. This is easily calculated from the formula ..

Degrees of Freedom for Chi Square Test =

(Number of Data Columns in the Table minus One)

multiplied by

(Number of Data Rows in the Table minus One)

In this example we have three treatments and three possible outcomes which means we have FOUR degrees of freedom; (3-1) * (3-1) = 2 * 2 = 4

Reading across the row corresponding to FOUR degrees of freedom in APPENDIX ONE we have three values 7.78, 9.49 and 13.28.

Imagine that we have a Chi Square result of 8.2 which is to the left of the value 9.49 we can state that the p value for this Chi Square Test is >0.05 and <0.10.

Any statistical test that returns a 'p' value of > 0.05 is a sign that there is NO statistically significant finding using the test.

Any statistical test that returns a 'p' value of < 0.05 is a sign that there IS a statistically significant finding using the test.

So in conclusion we can say that there was no statistically significant evidence to support the hypothesis that the three treatments had any significantly different outcome.

4.1 : ONLINE CHI SQUARE CALCULATOR

There is an online calculator for doing Chi Square testing at:

www.physics.csbsju.edu/stats/contingency.html

or you can download one based upon an Excel Spreadsheet from the author's website :
www.medlabstats.com/alliedhealth

5: WILCOXON RANK SUM TEST ON INDEPENDENT SAMPLES

In the Introduction I made the comment

"First we have to define what constitutes a success, a failure and an equivocal outcome.

Then a fairly intuitive next step would be to measure the frequencies of successes, failures and equivocal outcomes (neither successes or failures)"

I am now going to add a second *fairly intuitive* next step and that is just to look at the data in terms of the relative sizes of the responses and see if these correspond to your classification or hypothesis.

Suppose you had a fitness intervention programmes, which had the aim of getting the participants to loose a bit of weight, and these were being run at two gyms. Lets call the two groups East Gym and North Gym and look at their weight losses six months into their participation, see Table 2

TABLE 2

EAST GYM		NORTH GYM	
Andrew	7.5	Fred	7.6
Ben	7.7	Geoff	8.2
Carl	8.3	Harry	8.5
Dennis	8.6	Ian	8.8
Eddie	8.9	John	9.1

We would agree that if we sorted all the weight losses in ascending order AND if there was no difference between the gym groups then the results should alternate East, North, East, North …… So if we then went on to add up the ranks for the two gyms we would expect to identical results. When we do this on small samples like those shown in Table 2 we get Rank Sums of 24 and 30 for the East and North gyms respectively; not exactly equal but this illustrates the 'small sample size' effect. This kind of 'phenomenon' was studied by a famous statistician Frank Wilcoxon who devised the simple and elegant Wilcoxon-Rank Sum Test. He then went on to devise the Table in APPENDIX TWO that covers a range of possible situations of groups of various sizes. This is read as follows: Read across to 5 in the category 'Number of Subjects in Group with Fewest Members' and then down to row 5 which equals the 'Number of Subjects in the Group with the Most Members'. In that box you can see two numbers – 17 and 38. These are the two limits that you interpret your calculated rank sum against. Because 24 and 30 are both INSIDE these limits you can state that there was no effect of exercise on their final weight losses of the people in this study at the p<=0.05 level of significance.

If there are unequal numbers of individuals in the two groups then the rule for interpretation changes to ***test the rank sum of the SMALLEST GROUP against this range.*** If that is outside the range quoted in APPENDIX TWO then we can say that we have observed a statistically significant difference in the body weights of the two groups at the p<=0.05 level of significance.

The Table in APPENDIX TWO can only be used for datasets up to 15 rows so for larger dataset you will need to carry out the following calculations

$$\text{Theoretical Mean of the Ranks} = (N_A\ (N_A+N_B+1))/2$$

where N_A and N_B are the number of rows of data in the two datasets, A and B.

The Theoretical Standard Deviation, SD,

$$= \text{Square Root of } [(N_A*N_B)*(N_A+N_B+1)/12]$$

Conventionally the 95% confidence limits are now calculated as:

$$\text{Theoretical Mean of the Ranks} - 1.97*\text{Theoretical SD}$$

and

$$\text{Theoretical Mean of the Ranks} + 1.97*\text{Theoretical SD}$$

(Hint : Always work to four figures).

If both values calculated from a dataset fall **OUTSIDE** this acceptance range, we can say with 95% confidence that the two groups of patients investigated were different. ie. the treatment had had a statistically significant effect.

There is a downloadable calculator for the Wilcoxon Rank Sum Test at

www.medlabstats.com/alliedhealth

The Mann-Whitney U Test is almost identical and a web based calculator is available at

http://faculty.vassar.edu/lowry/utest.html

and the Table for interpreting the significance of the Mann-Whitney U Test result is in APPENDIX THREE.

6: THE MANN-WHITNEY U-TEST

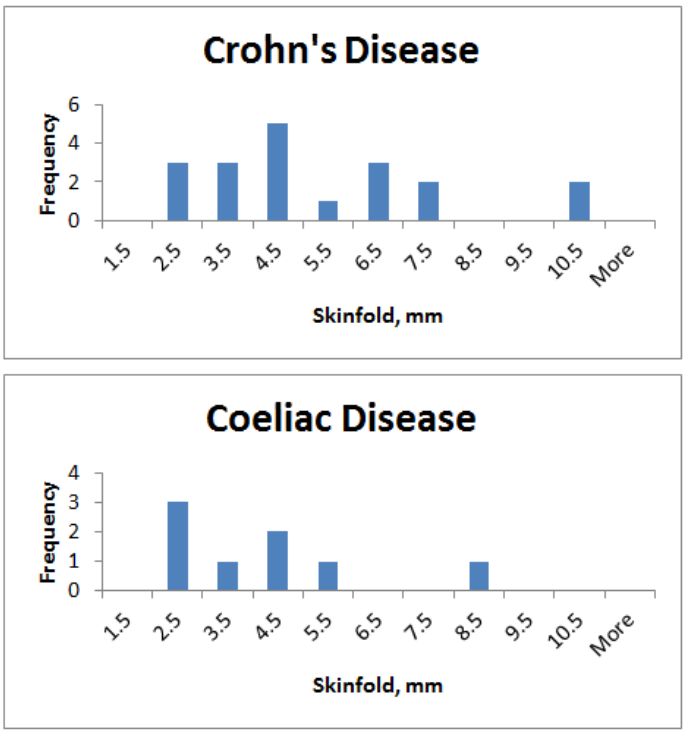
Not all datasets can be expected to fit a Normal Distribution so there is a branch of statistics called Non-parametric Statistics that is dedicated to solving problems on such datasets. An example dataset published by Bland and Altman (BMJ 1996; 312 : 1153) is shown in Table 13. Skinfold thickness measurements are taken as a measure of subcutaneous fat reserves and are regarded as a nutritional index. Poorly nourished individuals have reduced skinfold thicknesses. Because both Crohn's Disease and Coeliac disease are characterised by fat malabsorption this dataset can be used to assess the relative differences in fat malabsorption between the two groups.

TABLE 13

Crohn's Disease (A)		Coeliac Disease (B)
1.8	4.2	1.8
2.2	4.4	2.0
2.4	4.8	2.0
2.5	5.6	2.0
2.8	6.0	3.0
2.8	6.2	3.8
3.2	6.6	4.2
3.6	7.0	5.4
3.8	10.0	7.6
4.0	10.4	

One of the first things to do with a small dataset like this is to plot the histograms of the data using Excel.

FIGURE 10



It is immediately obvious from Figure 10 that the datasets are no way near fitting the characteristic bell shaped Normal Distribution curve and the main reason for this is the paucity of data points. There is, however, an impression that the Coeliac data are clustering lower than the Crohn's disease data. Is this statistically significant ?

Two statisticians – Henry Mann and Donald Whitney – developed a statistical test now known as the Mann-Whitney U Test to analyse these type of datasets. Essentially the theory is If you have two columns of data

- and you pool them,
- then sort them into ascending order,
- then add up the ranks for the items that came from the 'first' column
- then add up the ranks for the items that came from the 'second' column
- and compare that with the sums of the ranks of the items that came from the 'second' column
- then compare the these two sums
- if they are about the same then the two columns of data are probably closely overlapped.

They modelled this concept for all types of overlapping, non-overlapping and unequal columns of data and came up with this 'U' statistic that they could assign a 'p' significance value to.

There is an online calculator at

http://elegans.som.vcu.edu/~leon/stats/utest.html

When you put these data into their calculator then you get the following answer (see Figure 11).

FIGURE 11

U Test Results

n_1	n_2	U	P (two-tailed)	P (one-tailed)
20	9	120.5	0.1530786	0.0765393
normal approx z = 1.43778			0.1504954*	0.0752477*

*These values are approximate.

The two samples are not significantly different (P >= 0.05, two-tailed test).

Interestingly if you delete the highest point of 7.6 from the Coeliac dataset because "it looks like an outlier" then you get a "significant difference", (see Figure 12.)

FIGURE 12

U Test Results

n_1	n_2	U	P (two-tailed)	P (one-tailed)
20	8	118.5	0.0487486	0.0243743
normal approx z = 1.95791			0.0502408*	0.0251204*

*These values are approximate.

The difference between the two samples is marginally significant (P < 0.05, two-tailed test).

This program automatically adds comments about whether the Mann Whitney U Test has produced evidence of a significant difference between the two datasets.

7: KRUSKAL-WALLIS TEST

8: HISTOGRAMS AND MEDIANS

A useful technique for getting an idea of what your data set looks like is to use a histogram. A histogram is a graphical representation of your data sorted into size “buckets” along the x-axis and count of the number of values that fall into that “bucket” along the y-axis. The correct statistical term for “bucket” is interval or class; Microsoft Excel confuses the issue further by using the term “bins”.

The first decision we have to make what size of interval we are we going to use. If we use too large an interval then we will lose any idea of the fine detail in the profile of our data. Conversely if we use too large a size for the interval then we will get less idea of the overall frequency profile in our data.

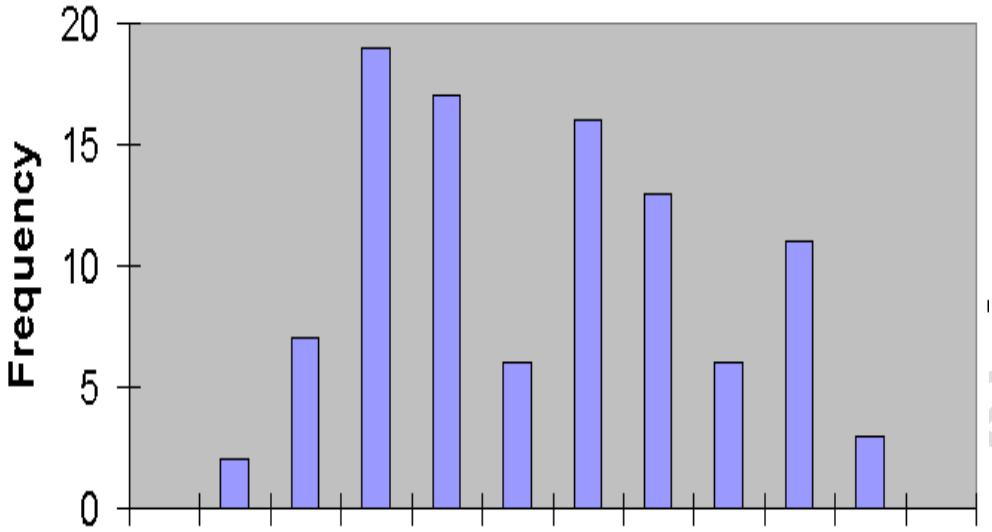


FIGURE 1
HISTOGRAM OF THE BODY WEIGHTS OF 100 ARMY RECRUITS

The rule of thumb in deciding on the best interval size is to divide the range of the data by the square root of the number of items in your dataset. You then have to use some subjective judgement to round that interval into a more usable vale. For example if the calculation suggests an interval size of 0.28 then you are best going to one of 0.25. When you use the Histogram Tool in Microsoft Excel you also get a printout of the frequencies in each interval. Table 3 gives an example where the body weights of 100 Army recruits were analyzed using an interval size of 5 kgs starting from 51 kgs.

TABLE 3
CANDIDATE WEIGHTS CLASSIFIED INTO 5kg WIDE INTERVALS

Interval Size in kgs	Number of Candidates in Interval
51-55	2
56-60	7
61-65	19
66-70	17
71-75	6
76-80	16
81-85	13
86-90	6
91-95	11
96-100	3

We can use Table 3 to put together some useful general descriptors of a dataset. By convention the definition of a clinical "Normal Range" is the range of values we find in 95% of our sample of the population. The more acceptable term in place of Normal Range is **Reference Interval**

We have a sample of 100 candidates in this example. So our normal range should cover 95 of our candidates. Strictly speaking this would mean that we ignore the weights of the lowest 2 ½ candidates and ignore the weights of the top 2 ½ heaviest candidates. We cannot do this in practice so round up to the nearest whole numbers – ignore the weights of candidates, 1, 2 and 3 and candidates 98, 99 and 100.

As a result we get a Non- Parametric Reference Interval of 58 - 95 kgs.

This Normal Range has been determined using a Non-Parametric Technique and this should be mentioned when reporting it.

Another way to determine the Reference Interval is to use Parametric Statistics. The definition of a Reference Interval in Parametric Statistics is:

"Mean plus and minus 1.96 times the Standard Deviation"

To calculate a Standard Deviation you will need an electronic calculator that has statistical functions programmed into it. Alternatively you can use the Descriptive Statistics Tool in Microsoft Excel. Either way we will get the following results

Mean	=	74.81
Standard Deviation (SD)	=	11.77

With these two results we calculate :

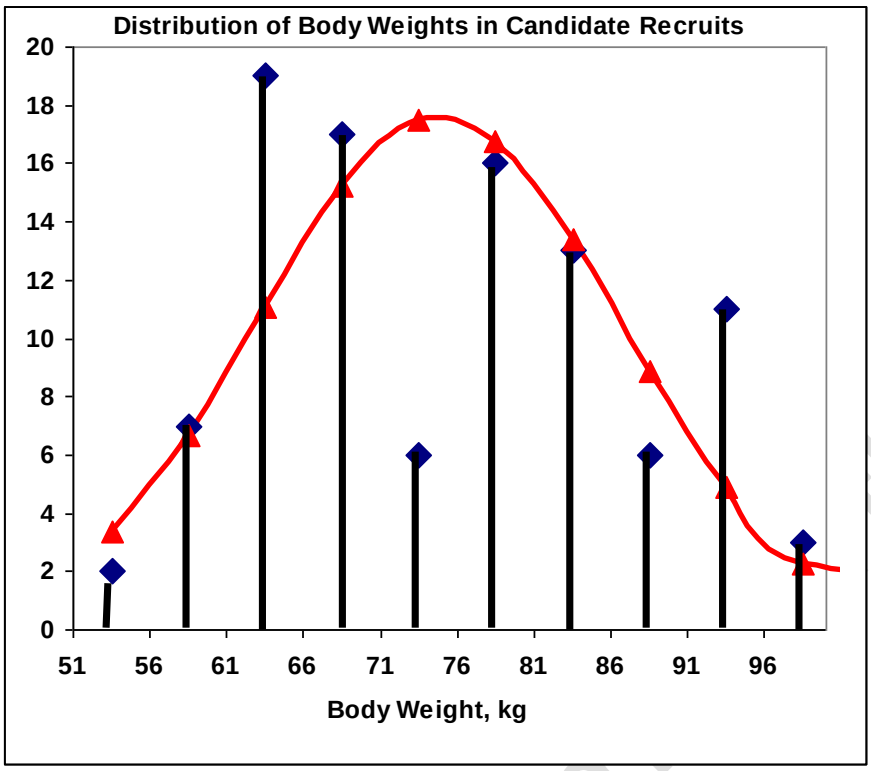
74.81 – (1.96 x 11.77) = 51.7 kg
74.81 + (1.96 x 11.77) = 97.9 kg

which leads to a **Parametric Reference Interval of 75 – 98 kgs.**

This is wider than the Non-Parametric Normal Range. We will not go into the detail here as to why this is but the reason is best illustrated by looking at Figure 2 which shows the "Normal Probability Curve" calculated from the mean of 74.81 and SD of 11.77 superimposed on the histogram of our data.

The calculated "Normal Probability Curve" is always a symmetrical bell shaped curve shown in red on this diagram. Our example dataset is not very symmetrical as we can see from the histogram. It is always very important to do a comparison like this because more often than not data collected during Allied Health Research Projects are not "Normally Distributed" and when you observe this during your work you should immediately switch over to using the Non-Parametric Statistics tests.

FIGURE 2



In Parametric Statistics the measure of the central value of a data set is always taken as the **mean.**

In Non-Parametric Statistics it is the **median** which is the value in the dataset at which 50% of the values in the data set are less than that value and 50% of the values in the dataset are greater than that value.

When we have an odd number of data items, e.g. 27 then the median is the 14th value; the first 13 values in a sorted dataset are less than the 14th value and the 15th value to the 27th value are greater than the 14th value. When you have an even number of data items eg 100 then the median value lies half way between a couple of values – in this examples between the 50th and 51st data points.

One final descriptor which can be of use is the mode. The **mode of a histogram** is the class which has the highest "frequency" in it. In Table 2 data it occurs in the class 61-65 kgs. That class contains 19 candidates.

The **mode of a dataset** is the value in the dataset which occurs with the greatest frequency. The mode of this dataset is 78 kg because there are 9 candidates with this body weight.

Why are they different? Because the mode of the histogram is dependent on the interval size and the choice of starting value for the series of intervals. This becomes apparent if we redraw Table 3, keeping the interval size the same but changing the starting value to 53 kg, see Table 4.

TABLE 4 : CANDIDATE WEIGHTS CLASSIFIED INTO 5kg WIDE INTERVALS STARTING AT 53 kg

Interval Size in kg	Number of Candidates in Interval
53-58	4
59-64	18
65-70	23
71-76	6
77-82	20
83-88	10
89-94	14
95-100	5

Under this classification the modal class of our new histogram appears as the 65-70 kg class. This illustrates the important point that you should be very alert to how graphs in statistics can give a "false" impression!

9: QUARTILES AND BOX AND WHISKER PLOTS

The Box and Whisker plot is perhaps a quicker way of getting an idea of the 'symmetry' of your dataset than is the histogram.

First you sort your data in order of size from the minimum to maximum. You then divide the series into four equal parts ie if you have 100 data points you divide the sorted series into four groups of 25.

The FIRST QUARTILE lies between the 25^{th} and 26^{th} values
The SECOND QUARTILE lies between the 50^{th} and 51^{st} values
The THIRD QUARTILE lies between the 75^{th} and 76^{th} values

(Notice that the SECOND QUARTILE is identical to the MEDIAN discussed in the previous section)

For a generalised series of ordered data **any percentile** can be calculated using these linear interpolation formulae as are recommended by NIST and used by Microsoft Excel.

So to estimate the value, V_P of the p^{th} percentile of an ascending ordered dataset containing N elements with values $V_1, V_2, ..., V_N$ we first calculate n

$$n = \frac{p}{100}(N-1)+1$$

n is then split into its integer component, k and its decimal component, d

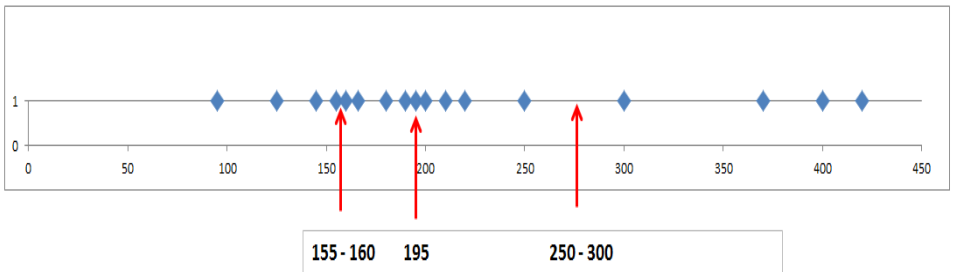
V_p is then calculated as:

$$v_k + d(v_{k+1} - v_k)$$

For example if we have a simple series of seventeen serum creatinine results :

95, 125, 145, *155, 160*, 166, 180, 190, 195, 200, 210, 220, *250, 300*, 370, 400, 420

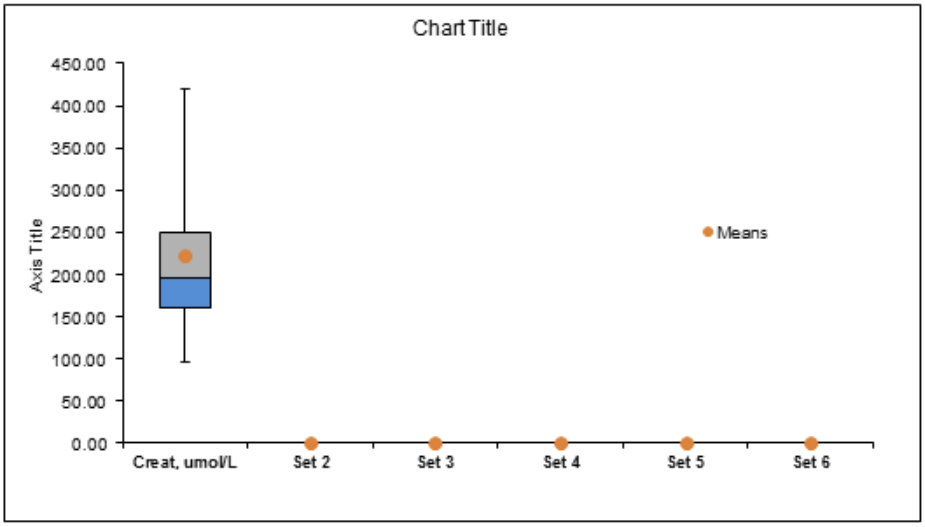
FIGURE 13



In this series the median is easy to determine … it is the (0.5 + 17/2) = 9th item; ie 195.

The first and third quartiles fall between the 4th and 5th values (155 and 160) and the 13th and 14th values (250 and 300). Simple linear interpolation would give a fairly good estimate of the first quartile because the 4th and 5th values are numerically very close to each other….. (155 + 160)/2 = 157.5. Excel gives an answer of 160. In contrast simple linear interpolation between the 13th and 14th data points gives a value of 275 whereas Excel gives a value of 250. So it is always best to use the formulae and procedure that Excel uses. (You will not be expected to do this in an exam.) In the lecture we will use an Excel template that has been set up for you to determine the quartiles and vertical 'Box and Whisker Plots' for up to six variables by 100 points per variable, (see Figure 14)

FIGURE 14



	Creat, umol/L	Set 2	Set 3	Set 4	Set 5	Set 6
Max	420.00	0.00	0	0	0	0
75th Percentile	250.00	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!
Median	195.00	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!
25th Percentile	160.00	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!
Min	95.00	0.00	0	0	0	0
Mean	222.41	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
Number of points	17	0	0	0	0	0

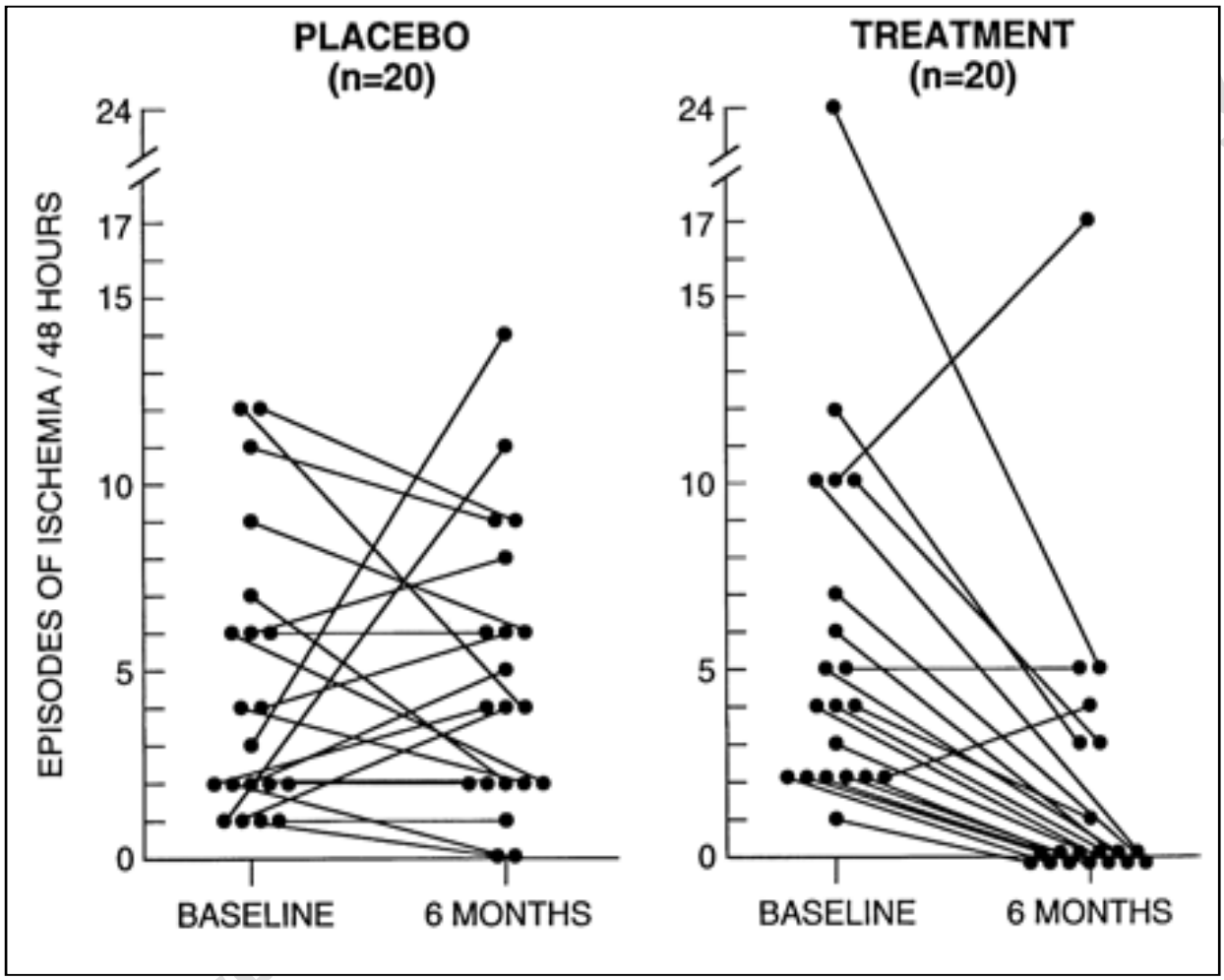
This Excel spreadsheet template can be downloaded from

www.medlabstats.com/alliedhealth

10: BUMP PLOTS

These are useful plots for visualising trends in paired data typically the data from a group of patients on entering a clinical trial and their corresponding data after a period of treatment. Figure xxxxx shows a typical example from the literature :

FIGURE xxxxxx



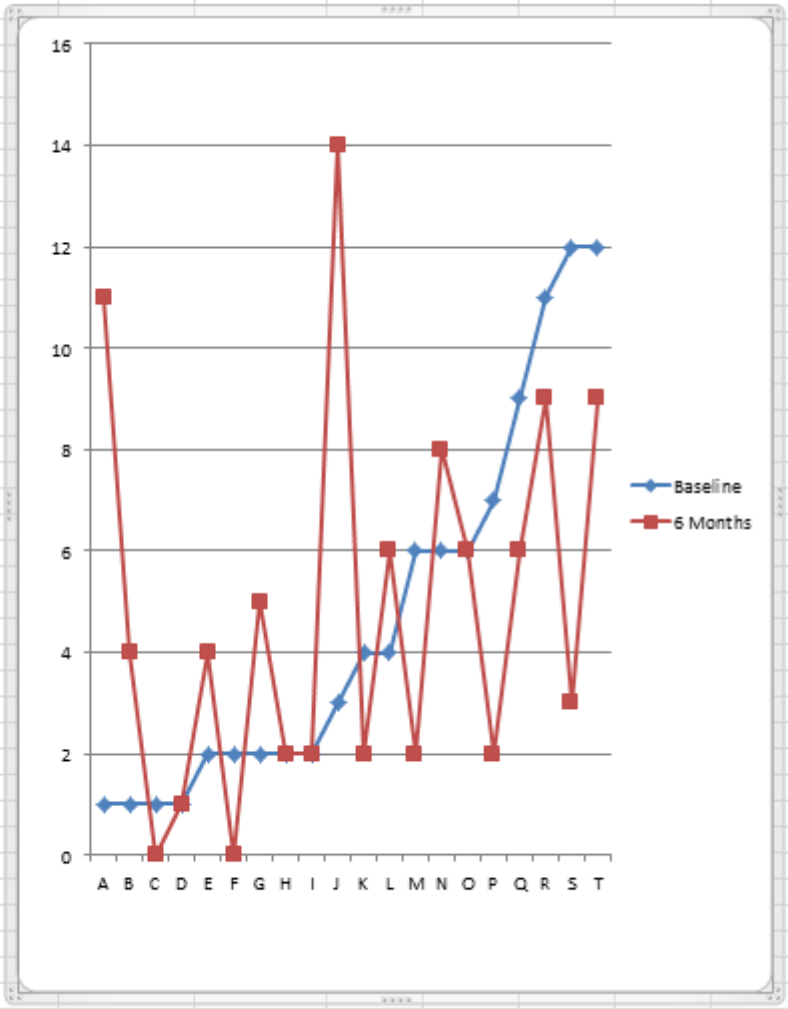
Without any formal statistical testing it is clear that the placebo had no beneficial effect but that the treatment did. The obvious statistical test to use in this type of clinical trial is the Wilcoxon Rank Sum Test described in an earlier section of this monograph.

Bump Plots can be produced using Excel as follows :

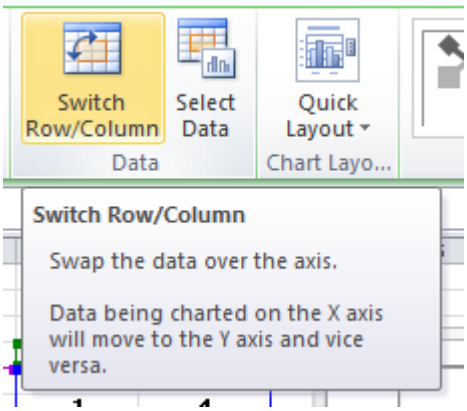
(1) Highlight the data and select a line graph Figure xxxx

FIGURE XXXX

	Baseline	6 Months
A	1	11
B	1	4
C	1	0
D	1	1
E	2	4
F	2	0
G	2	5
H	2	2
I	2	2
J	3	14
K	4	2
L	4	6
M	6	2
N	6	8
O	6	6
P	7	2
Q	9	6
R	11	9
S	12	3
T	12	9

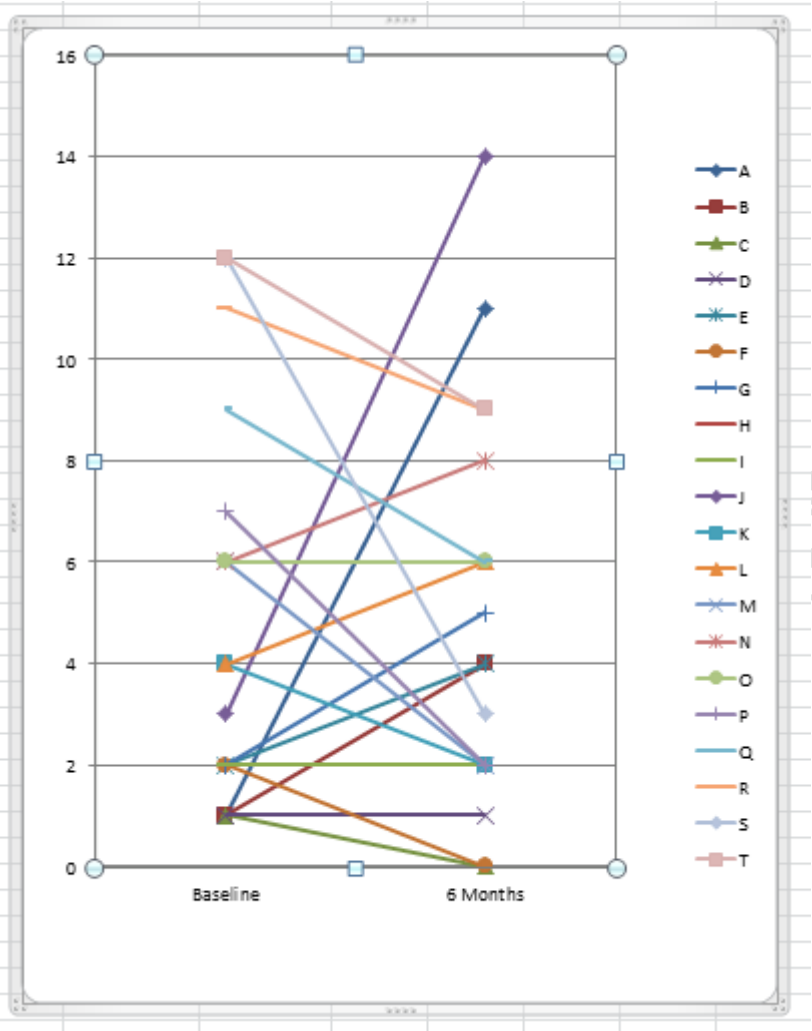


(2) Now click on the Switch Row/Column button, see Figure xxxxx



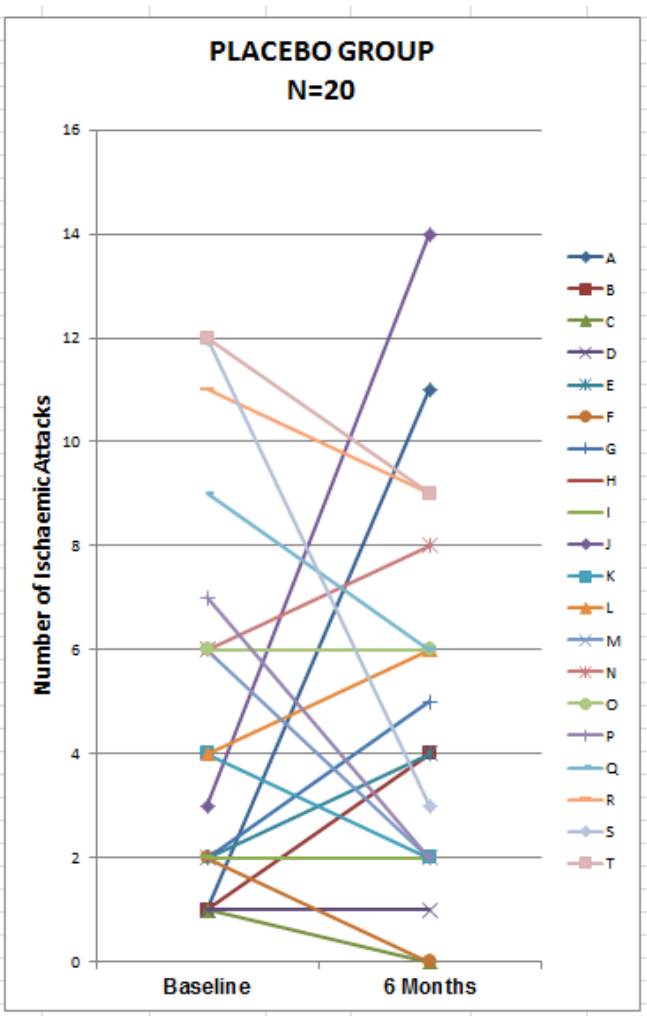
This gives the result shown in Figure xxxx which is very close to the appearance of the publication version of the bump plot

FIGURE xxxxxx



Further manipulation with the titles to the axes etc etc will give you are very reasonable looking bump plot; see Figure xxxxxx.

FIGURE xxxxx



11: THE NORMAL DISTRIBUTION OF FREQUENCIES, ITS MEAN AND ITS STANDARD DEVIATION.

11.1 CHARACTERISTICS OF THE NORMAL DISTRIBUTION

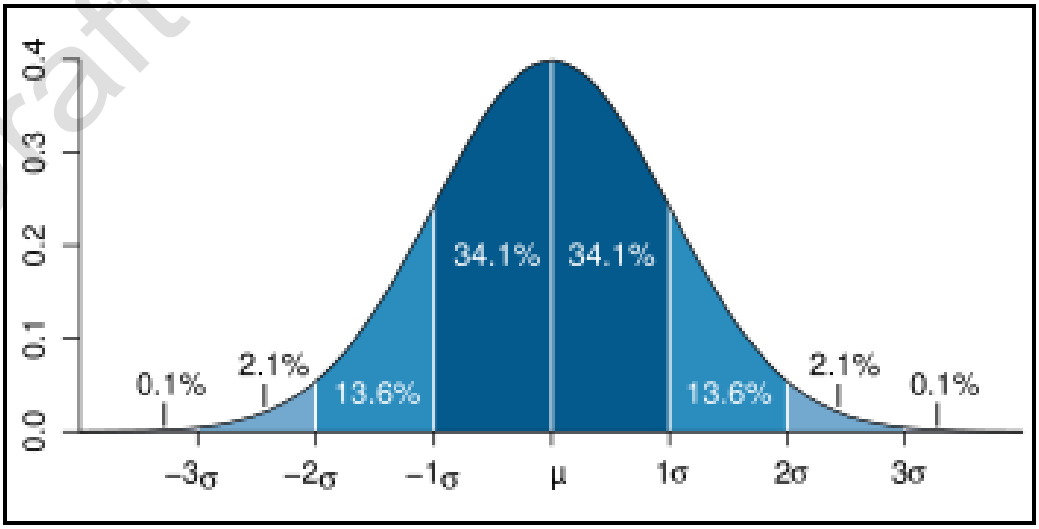


FIGURE 5

- It is a frequency distribution that has a characteristic Bell Shaped. The frequencies of a particular X value are shown on the Y axis and the values of the measured parameter are shown on the X axis.

- The frequency distribution is described by the continuous function :

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{-(X-m)^2/2\sigma^2}$$

Where Y is the frequency, N is the total number of data items in a series, X is a value on the x axis, m is the Mean of the Data, sigma is the Standard Deviation of the dataset and Pi is the mathematical constant 3.142.

- The frequency distribution is symmetrical about the mean of the data.

- It has a property called the **Variance** which is equal to Mean of the Sum of the Squares of the differences of any 'X' value from the Mean of the Data, m.

$$\textbf{Variance} = \frac{\sum_{1}^{N}(X - m)^2}{N}$$

N is the total number of data items.

- It has a property called the **Standard Deviation**, signified by the Greek letter sigma (σ) which is equal to the Square Root of the Variance.

- The Standard Deviation divided by the Mean and multiplied by 100 is the **Coefficient of Variation**.

- The Standard Deviation divided by the Square Root of the Number of Observations is the **Standard Error of the Mean**.

- The area under the Normal Error Curve is equal to One (or 100%).

- 34.1% of the observations lie between the Mean and plus one Standard Deviation from the Mean, see Figure 6.

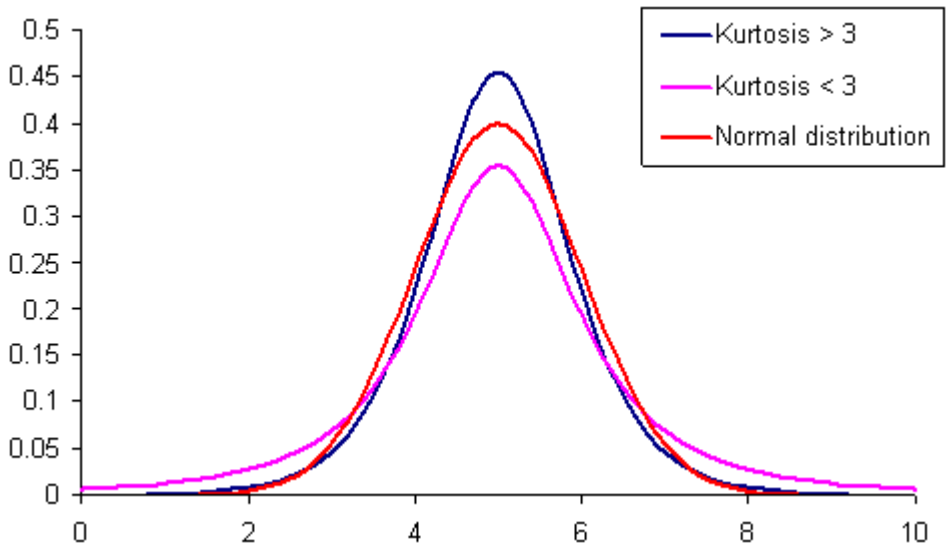
- 47.7% of the observations lie between the Mean and plus two Standard Deviations from the Mean
- 49.8% of the observations lie between the Mean and plus three Standard Deviations from the Mean
- Addition of separate estimates of a population's Variances is allowed.
- Addition of separate estimates of a population's Standard Deviation is not allowed.
- ONLY DATASETS THAT CONFORM TO THESE CHARACTERISTICS CAN BE ANALYZED USING PARAMETRIC STATISTICAL TESTS.

11.2 KURTOSIS

Kurtosis, K, is a measure of the peakedness of a distribution. The (normalized) kurtosis statistic is calculated from the formula:

$$K = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sigma^4}$$

The division by σ^4 factor is used to make the kurtosis a pure number. Kurtosis is often known as the fourth moment about the mean because the differences between each point and the mean are raised to the power 4; (it is given the symbol μ_4 for a population and m_4 for a sample. Because it is raised to the power 4 it is very sensitive to the values of the data points well away from the mean ie. data points that fall in the two tails of the distribution.) By definition Normal distribution with a conventional profile has a Kurtosis of 3.



It must be remembered that in Excel when you use the KURT() function it is giving the solution of the following formula :

$$K = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sigma^4} - 3$$

This is identical to the formula given above apart from the fact that there is a '*minus 3*' term This means that in Excel a Normal distribution has a Kurtosis of ZERO. In some statistical texts you will find this 'Excel approach' referred to more correctly as the Excess Kurtosis

There are three special terms that are often used in relation to kurtosis viz.

- A **Leptokurtic** distribution – this has a kurtosis > 3 (or greater than zero in Excel) and means that has more "peakedness" . (Lepto- comes from the Greek meaning fine or slight
- A **Mesokurtic** distribution – this has a kurtosis = 3 (or equal to zero in Excel) and means that it has the conventional profile of the Normal Distribution. (Meso- comes from the Greek meaning middle)
- A **Platykurtic** distribution – this has a kurtosis < 3 (or less than zero in Excel) and means that it has a more flattened profile than you see in a conventional Normal Distribution. (Platy- comes from the Greek meaning flat or broad)

In Figures xxxx, yyy and zzz are some illustrations of datasets that have a kurtosis of approximately zero, approximately plus one and approximately minus one respectively. The idea here is to communicate an appreciation that if you add more data to the middle of a previously mesokurtic distribution (Figure xxxx) then it becomes Leptokurtic, see Figure yyyy. Addition and subtraction of data points from the original mesokurtic distribution dataset are indicated by the yellow highlighting in Figures yy and zzz. So to make a mesokurtic distribution become a leptokurtic distribution does not take much – just the movement of two items from either side to the centre and the addition of two additional data points to the centre is enough to get an almost +1 change in the kurtosis. To move the mesokurtic distribution to a platykurtic distribution (Figure zzz) takes considerably more effort. In this example 15 points were removed from the centre of the mesokurtic distribution followed by the addition of two data points into the tails plus extension of the tails out so as to accommodate data points beyond the previous maximum and minimum ie. the addition of the data points 20 and 65. IN all three Figures the shape of the calculated distribution for the corresponding mean and SD of the actual data in the histograms is shown to give a visual appreciation of the 'goodness of fit' of the data to the theoretical curve; in all examples it looks pretty good considering we have relatively small sample sizes ie. values of N as shown in the boxes.

FIGURE xxx

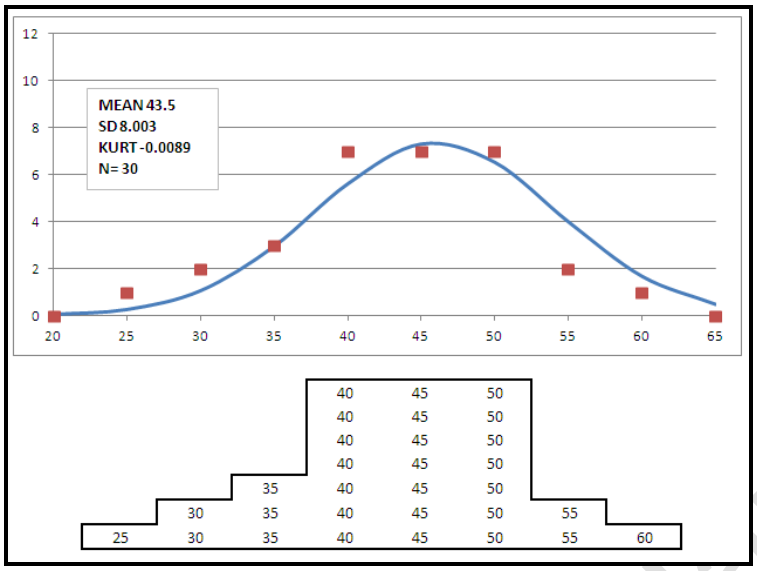


FIGURE yyy

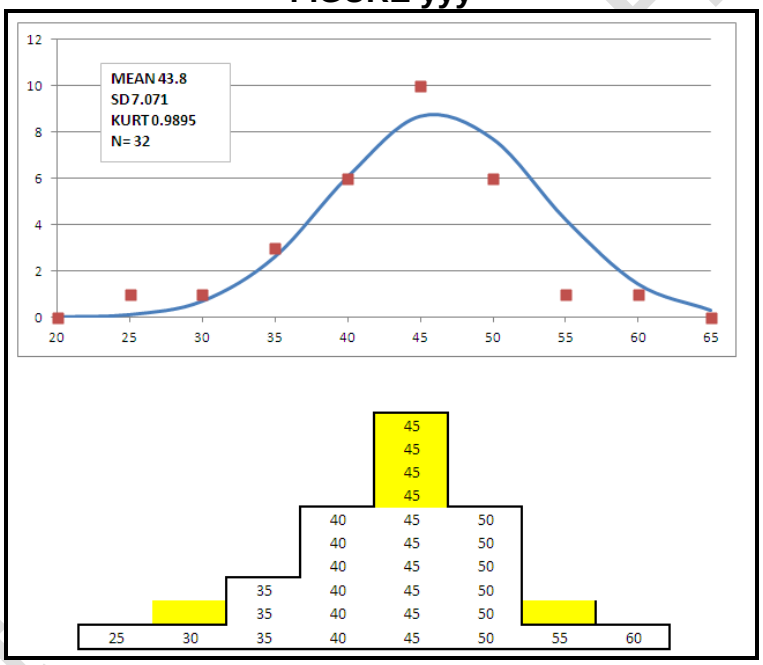
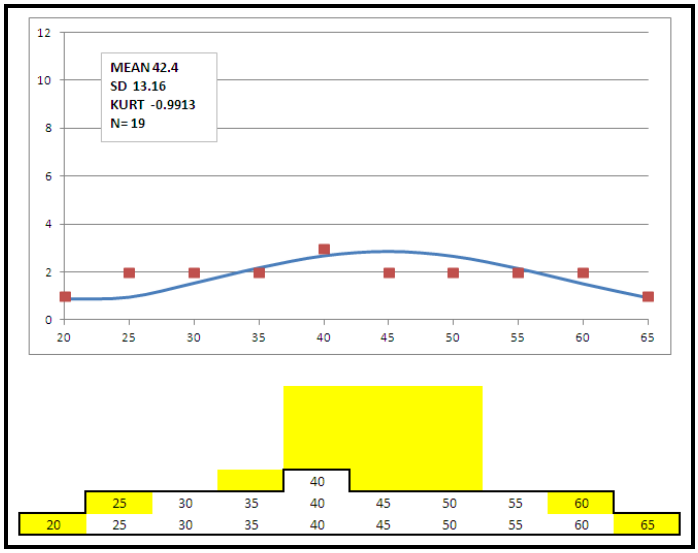


FIGURE zzz



11.3 SKEW

The concept of skew in statistics is identical to way in which we use the word in everyday life just that in statistics the word has an additional association with numeric values that communicates to us that the 'tail' to the skew is to the left if we have negative or the tail is to the right if we have a positive value to the skew. The formula for calculating Skew is very similar to that for Kurtosis – kurtosis worked with deviations from the mean raised to the power 4, skew works with deviations raised to the power 3.

$$\text{skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - x}{s} \right)^3$$

Figures aaa illustrates the skew observed for the same data set as was shown in Figure xxx. Unlike the kurtosis being near the anticipated value of zero the skew of these data is not particularly close to zero. Removing data from the distribution in Figure – signified by the yellow highlighting and adding data – signified by the green highlighting – produced Figures bbb and ccc. Figure bbb clearly has a tail to the left and a negative skew of -1.03 and Figure ccc has a distinct tail to the right and a skew of 0.946. In both scenarios – Figure bbb and Figure ccc – it is quite clear that the normal error curve (shown as the continuous blue line in the Figures) calculated for the corresponding means, SDs and N's does not fit the observed data at all well. This leads to important point that :

When examining Kurtosis and Skew values obtained from your data it is much more important to take note of the SKEW value than the kurtosis value. A skewed dataset is definitely not well described by the equation to the normal distribution curve and you would be very well advised to consider moving over to non-parametric statistics to analyse your data.

FIGURE aaa

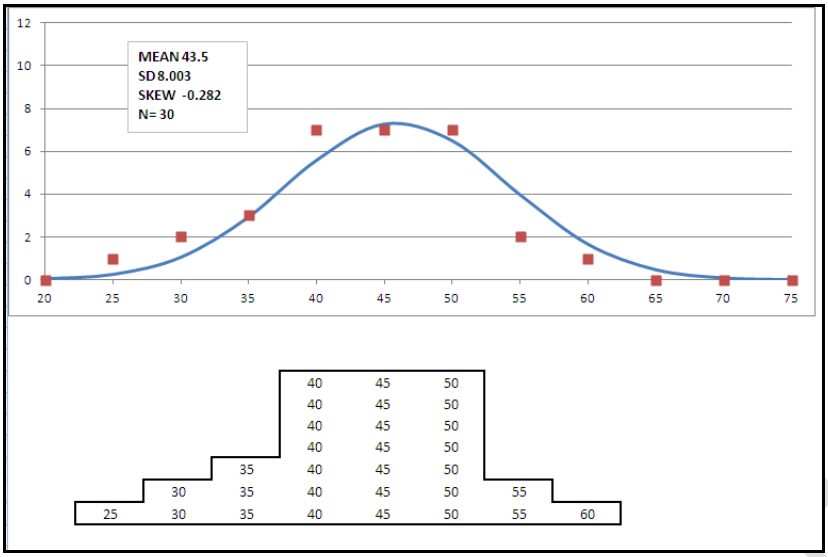


FIGURE bbb

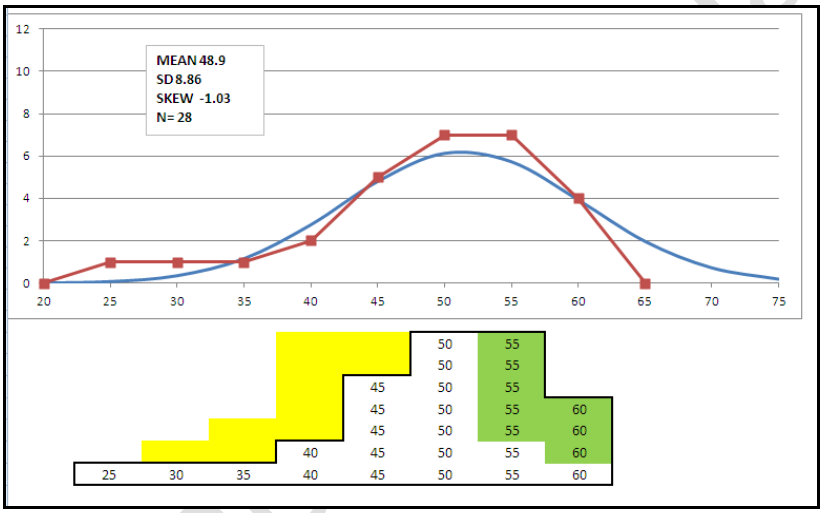
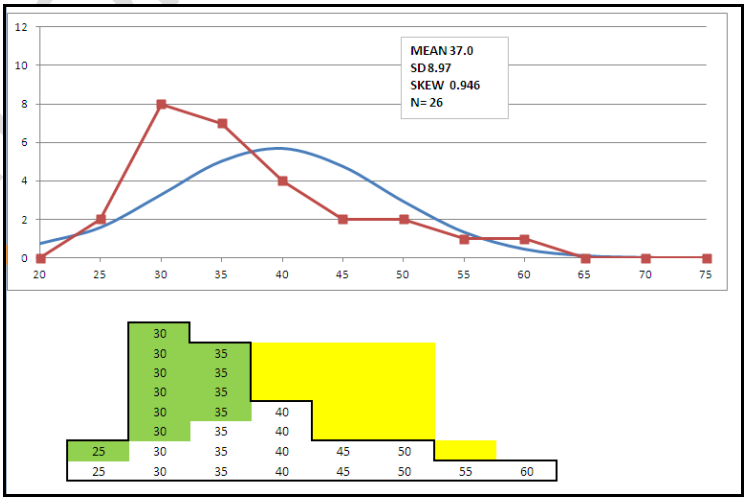


FIGURE CCC



11.4 INTERPRETING VALUES OF KURTOSIS AND SKEW FOR STATISTICAL SIGNIFICANCE

The subjective comment in the previous section about taking more note of the skew value than the kurtosis value can be restated objectively by calculating the Z-score or standadrd normal deviates for the actual values of skew and kurtosis. Both calculations require that the standard error of the skew be evaluated first using the formula :

$$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

You then divide the value of skew by to get the Z-score. If the Z-score for the skew is less than -1.96 or greater than 1.96 then the distribution can be said to be significantly skewed at the p=< 0.05 level of significance.

The formula for calculating the standard error of the kurtosis value follows; you can see that it requires the prior calculation of the standard error of the skew (SES).

$$SEK = 2 * SES \sqrt{\frac{N^2 - 1}{(N-3)(N+5)}}$$

A calculator written in Excel is available from the website :

www.medlabstats.com/alliedhealth

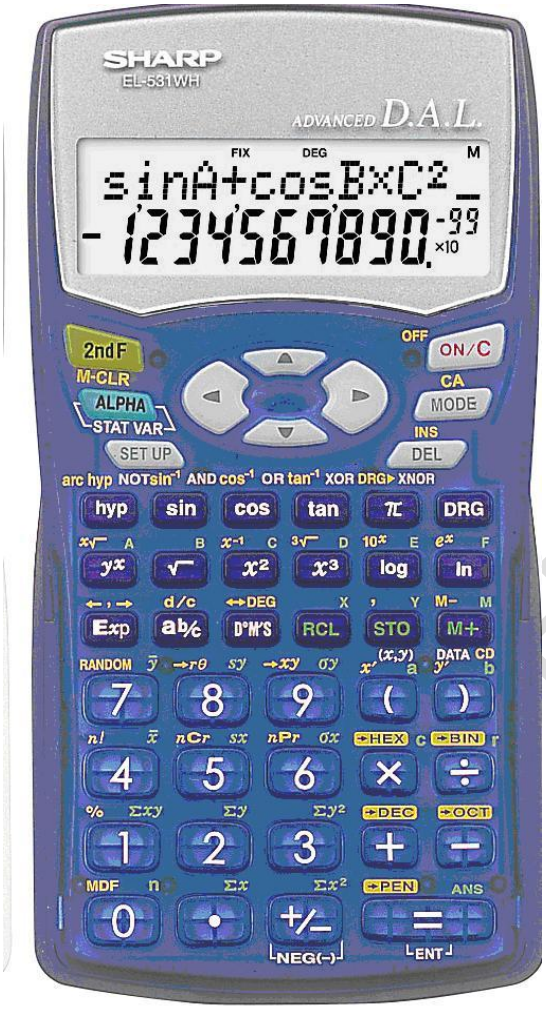
Using that calculator the following results were obtained for the data presented in Figures xxx, yyy, zzz, aaa,bbb and ccc and are shown in Table aaaa. Only the Skews observed in Figures bbb and ccc actually reached statistical significance ie p <= 0.05.

TABLE aaaa

FIGURE	SKEW	Z-score	KURTOSIS	Z-score
xxx	-0.2822	-0.66	-0.0089	-0.01
yyy	-0.4015	-0.97	+0.9895	+1.22
zzz	+0.0331	+0.06	-0.9913	-0.98
aaa	-0.282	-0.66	-0.0089	-0.01
bbb	-1.03	**-2.34**	+0.9772	+1.14
ccc	+0.946	**+2.08**	+0.5359	+0.60

12: Z-SCORE: This is a quantity that indicates how many standard deviations (SDs) an observation is from the mean of Normal Distribution that it belongs to. Strictly speaking the Z-Score can only be calculated if you know the Population Standard Deviation usually donated by the greek lower case letter sigma, σ. This is a good point to look at the common abbreviations used in Parametric Statistics because they will make it clearer what the keys on a typical statistical calculator, as shown in Figure 6, represent.

FIGURE 6



	SAMPLE	POPULATION
Mean of x	$\bar{X}$	**μ_x**
Mean of y	$\bar{y}$	**μ_y**
SD of x	**S_x**	**σ_x**
SD of y	**S_y**	**σ_y**
Slope of Regression Line	**B**	**-**
Intercept of Regression Line	**α**	**-**
Correlation Coefficient	**r**	**-**

The formula for the Z-Score using “Sample” population abbreviations:

$$\frac{x - \bar{X}}{S_x}$$

The formula for the Z-Score using “Population” abbreviations:

$$\frac{X - \mu_x}{\sigma_x}$$

The Z-Score is useful for

(i) Determining the number of times we could expect a particular value in our dataset if it truly is a member of a Normal Distribution.

(ii) The proportion of values in the distribution that is greater than a particular value.

(iii) The proportion of values in the distribution that is less than a particular value.

To answer the first question, (i), use APPENDIX FOUR :

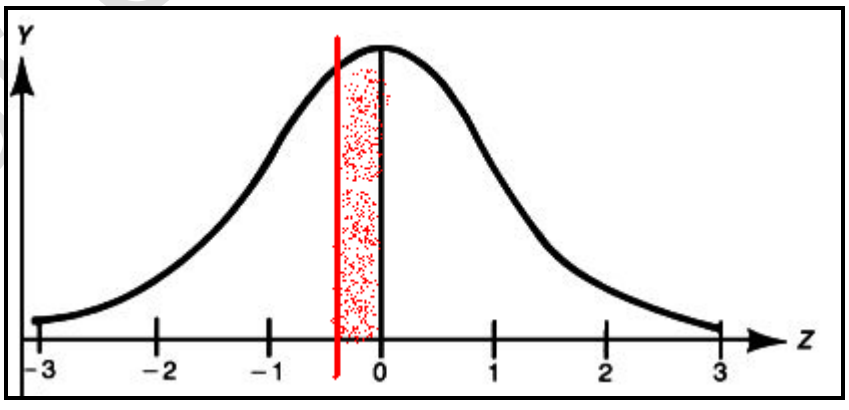
Suppose we want to know the frequency of results that we could expect at 1.25 X the SD of a distribution. The Y value we have for a Z=1.25 is 0.1826. What is not often recognised is that the values in APPENDIX FOUR refer to a distribution with an SD=1. Consequently we have to make a simple calculation to compensate for this. Divide the value we read from APPENDIX FOUR by our actual Sample SD and then multiply it by number of data items, N, in our dataset.

$$\text{Corrected Y} = \frac{\text{Table Value of Y} * \text{N}}{\text{SD of Our Dataset}}$$

To answer (ii) and (iii) we have to use APPENDIX FIVE :

Suppose we have a Z value of -0.31. Using APPENDIX FIVE just go down the left hand column to 0.3 and across to the square under 0.01. That square has the number 0.1217 in it. Because we have a negative value we are looking at the situation shown in Figure 8 – we are looking at the mirror image. And the value 0.1217 means that 12.17% (simply calculated by multiplying 0.1217 by 100) of the area under the distribution lies between -0.31 and zero.

FIGURE 8



Another 50% of the area under the curve lies to the right of zero. So in summary we have

- 12.17% + 50% of the area to the right of the Z value of 0, which equals a total of 62% in round numbers.

- 100% minus 62% of the area under the distribution lies to the left of the Z value of -0.31, which equals 38%.

These are the answers to questions (ii) and (iii) via in our Example Data Set 1 we have a 62% probability of there being values greater than Z = -0.31 and a 38% probability of there being values less than Z = -0.31, on the basis of our CALCULATED Mean and SD of the parent Population Distribution.

13: THE F TEST

The F Test is a **Variance Ratio Test** which must always be performed before you carry out a Student's t Test on the differences between two samples means or a sample mean and a hypothetical target mean. This is because the underlying assumption in the Student's t Test is that we are dealing with distributions which have similar (ideally identical) variances.

The rule is that the LARGER variance of the two datasets under investigation is divided by the by the SMALLER variance of the other dataset. If there is no difference between the two distributions then you would expect this ratio to be ONE. As the distributions become more and more dissimilar the value becomes larger. APPENDIX SIX shows the distribution of F at the P = 0.05 level.

14: THE Z TEST FOR COMPARING TWO SAMPLE MEANS

There are two tests for comparing sample means – this test and the Student's T Test. The Z Test is presented first because it is the easiest to interpret. The formula is

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{x}_1$ and $\bar{x}_2$ are the respective means of the first and second series of x data

$\frac{\sigma_1^2}{n_1}$ and $\frac{\sigma_2^2}{n_2}$ are the corresponding standard deviations squared

and divided by the corresponding numbers of items in the two series of x data ie. n_1 and n_2

There are two provisos for the application of this test – (1) that the size of the two groups must be 30 items or more and (2) the F-Ratio Test must not be significant ie. the variances of the two data sets must be comparable.

An Excel based worksheet for this calculation is on the website.

15: THE STUDENT'S t TEST

Definition (a) ……where you are comparing a single Sample Mean, $\bar{x}$, to a hypothetical Population Mean, **μ**

$$t = \frac{\bar{x} - \mu}{SEM}$$

Where SEM = standard error of the mean of your sample.

Definition (b) ……where you are comparing the means from two samples that HAVE THE SAME VARIANCES

Pooled Variance = $[SD_a^2(N_a-1)+SD_b^2(N_b-1)]/(N_A+N_B-2)$

Pooled SEM = $[(\text{Pooled Variance}/N_a)+(\text{Pooled Variance}/N_b)]^{1/2}$

Students t = ($\bar{x}_a$ - $\bar{x}_b$)/Pooled SEM

Student's t has a distribution very similar to the Z-Score. The main difference is that it departs from "Normality" when the number of data points in your data set is small. Use APPENDIX SEVEN to assess the significance of your Student's t Test analyses

---

16: GOODNESS OF FIT OF A HISTOGRAM OF YOUR DATA TO A NORMAL ERROR CURVE

This procedure brings together two principles - one drawn from parametric statistics and one from non-parametric statistics. First we use parametric statistics to determine the 'expected' frequencies in each cell of our histogram. We do this by using the equation to the area under the normal error curve :

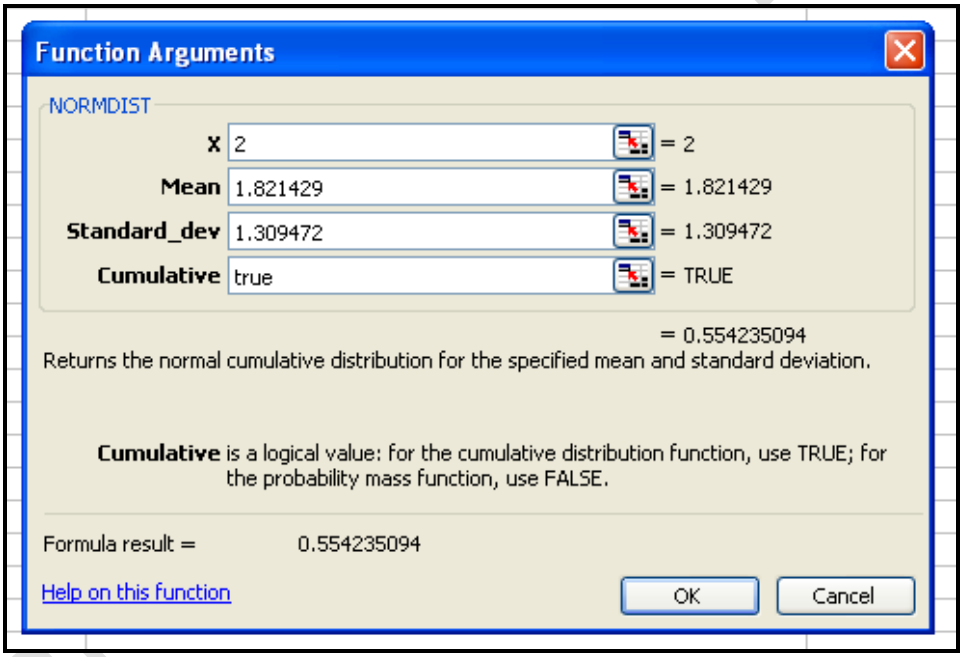
$$\text{Area Under the Curve from } -\infty \text{ to x} = \int_{-\infty}^{x} \frac{N}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}.dx$$

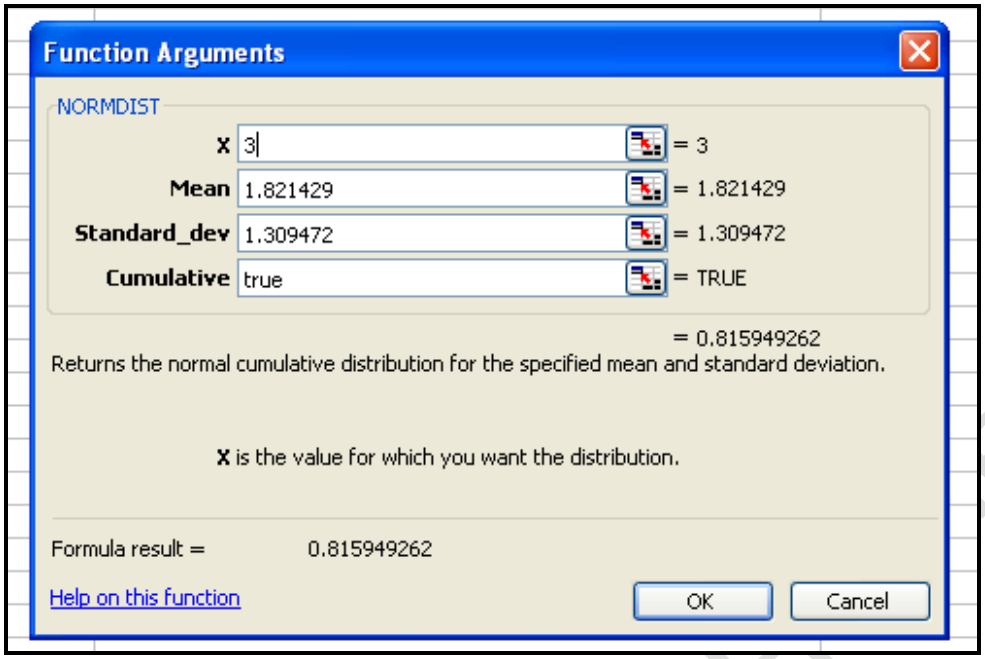
This is not an equation that we would want to solve 'by hand'. Fortunately there is a built in function in Excel called NORMDIST which actually solves the equation for any given mean

and SD. To get the frequency in an actual cell of our histogram we solve it first for the value of x equal to the lower limit of the cell and then for the value of x equal to the upper limit of the cell. By taking the first answer from the second and multiplying it by N we can actually get the expected frequency in that cell.

For example

Number of observations in the histogram	83
MEAN	1.821
SD	1.309
TARGET cell LOWER limit	2
TARGET cell UPPER limit	3
Calculated Cumulative Frequecy up to target cell LOWER limit	46.00
Calculated Cumulative Frequecy up to target cell UPPER limit	67.72
Expected Frequency in TARGET Cell	22





This has been put together online as an Excel worksheet :

www.medlabstats.com/alliedhealth

It gives you the tools to calculate a cell by cell expected frequency and transcribe these answers across to an area of the worksheet where the Chi Squared is calculated. The value of Chi Squared is interpreted as described previously with the **Degrees of Freedom being set to the number of cells in your histogram minus 2.**

17: LINE OF BEST FIT BY LEAST SQUARES LINEAR AND NON-LINEAR REGRESSION

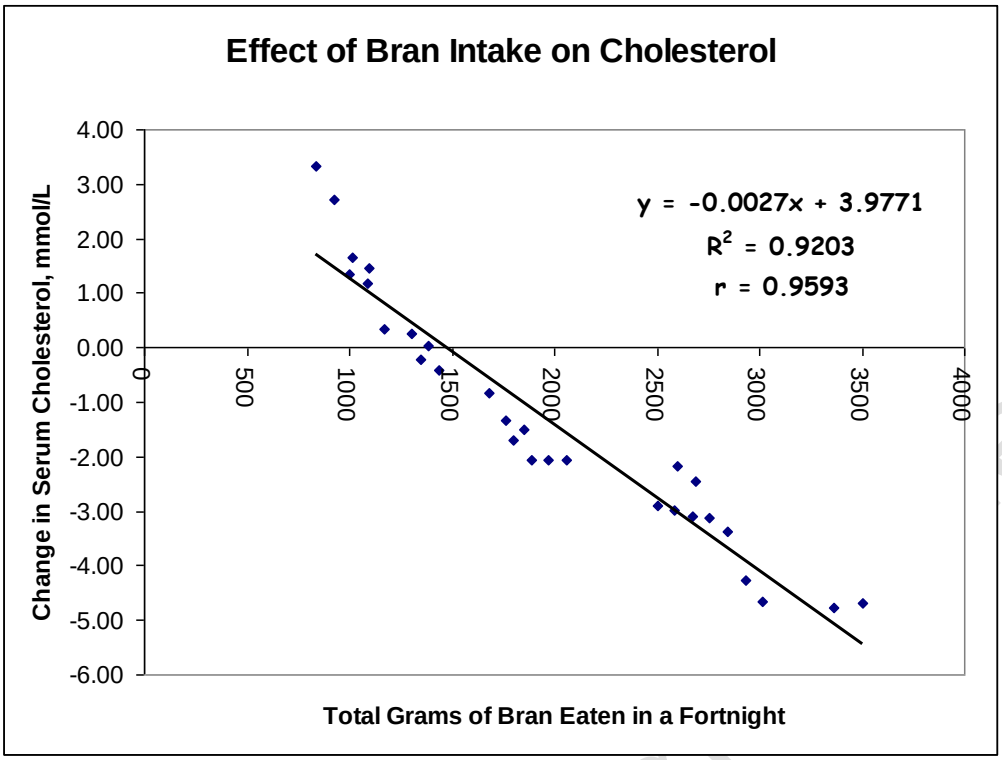
Line of Best Fit Regression is the technique used to investigate the functional relationship shown as graph where the size of a putative Cause is plotted on the x-axis and the putative Effect is plotted on the y-axis. (***putative*** *- purported; commonly put forth or accepted as true on inconclusive grounds*)

Microsoft Excel has both a Regression Tool and Add Trendline Tools which provide all the outputs .

Figure 9 illustrates a Linear Relationship between the Cause : Total Grams of Bran Eaten in a Fortnight and the Effect : Downward Change in the Subjects' Serum Cholesterol.

In a Linear Relationship there are only two functional parameters estimated – the **Slope** of the Straight Line and it's **Intercept** on the y-axis when the value of the Cause is zero. There is quality of goodness of fit parameter provided as well – the **Correlation Coefficient** which will be explained further in the next section**.** These results were obtained by using Microsoft Excel's Add Trendline tool.

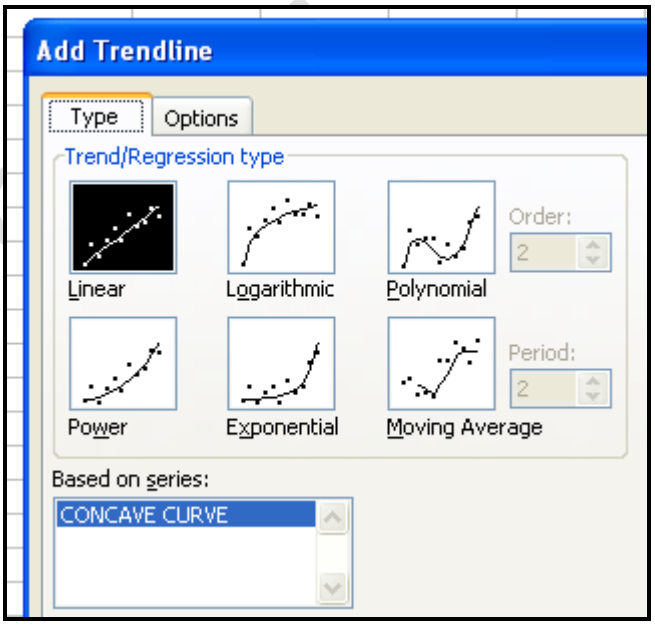
FIGURE 9



Slope = -0.0027
Intercept = 3.9771
Correlation Coefficient = 0.9593

Microsoft Excel also offers least squares fitting to other models :

FIGURE 10



Exponentials of the form $y = A\,e^{Kx}$

Powers of the form $y = Ax^{K}$

Logarithmic of the form y = A Ln(x) – K

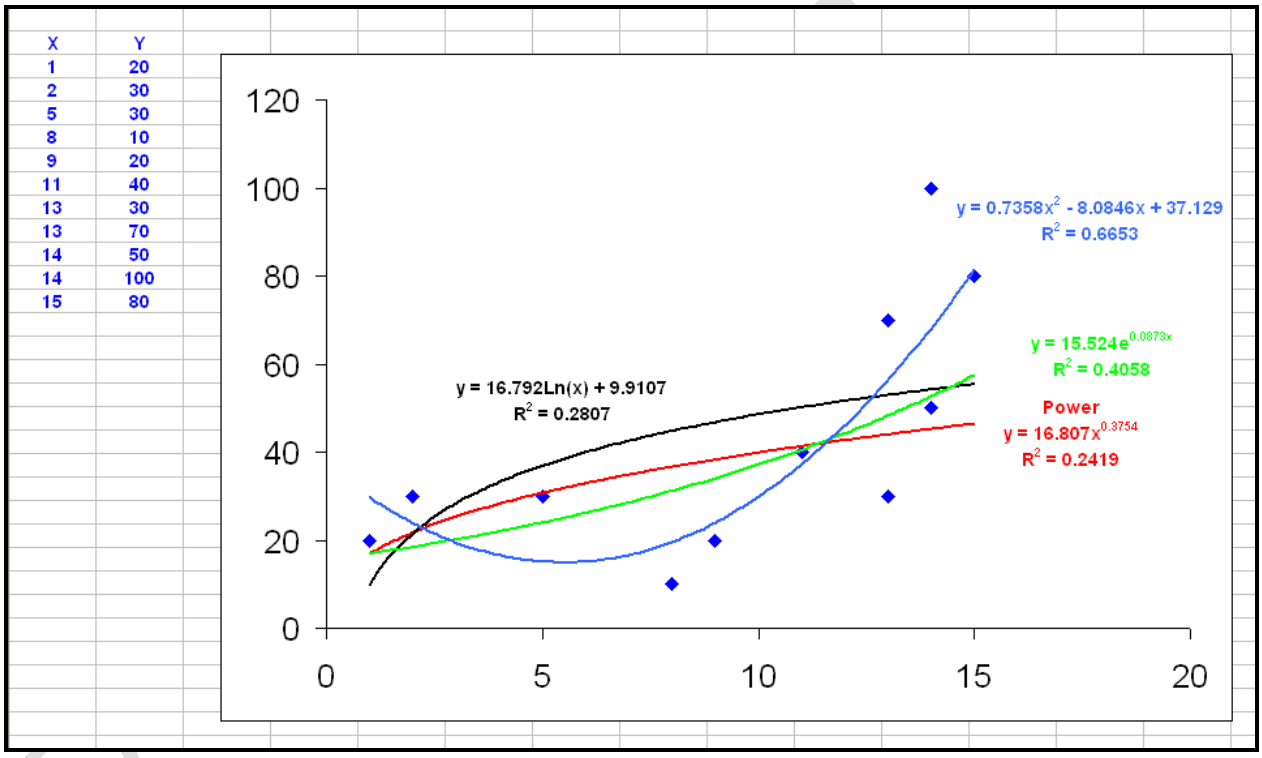
Polynomials of the form y = Ax^2 - Bx + K

Moving Average

(A, K and B are all constants determined during the least squares fitting process.) Note that the logarithmic fitting uses Natural Logarithms not Logarithms to the base 10. Polynomials can be up to the sixth order ie the equation returned has coefficients for the x^5, x^4, x^3, x^2 and x terms.

The 'best fit' equation should not be chosen solely on the basis of the highest associated value of the correlation coefficient. All lines of best fit should be examined 'by eye' on the plot of data points, as shown in Figure 11, and with a knowledge of the science and/or physiology behind the research. In physiological systems exponential responses are more likely than power responses so in Figure 11 the exponential fit is probably 'the fit of choice' even though the power function is associated with the highest value of the correlation coefficient.

FIGURE 11



18: THE CORRELATION COEFFICIENT

The correlation coefficient is defined as

'A measure of the interdependence of two random variables that ranges in value from −1 to +1, indicating perfect negative correlation at −1, absence of correlation at zero, and perfect positive correlation at +1. Also called coefficient of correlation.'

It is usually represented by the letter 'r'. To determine the significance of your value of 'r' you need to look up APPENDIX EIGHT against the appropriate number of **Degrees of Freedom**. **The latter is calculated from the number of data points in the graph minus 2.**

Closely related to the correlation coefficient is the ***coefficient of determination*** which is calculated by squaring the correlation coefficient. Statistically it represents the proportion of the total variation in the y variable that is explained by the regression equation. Multiplying the coefficient of determination 100 gives a percentage. For example if the value of a coefficient of determination is 0.75 then the least squares fitting has accounted for 75% of the variability in your y axis data.

19: ERRORS ON THE ESTIMATES OF THE SLOPE AND INTERCEPT OF A LINEAR REGRESSION

The Excel Regression Tool includes estimators of the error on the slope and intercept; see Figure 12. Much of the details on that report has been blanked out because they do not relate directly to this discussion of the errors on the slope and intercept. The results shown in Figure 12 were obtained for the data shown in Figure 9. They indicate that the Standard Error on the Intercept was 0.3232 and on the Slope (described in the Excel report as *X Variable 1*) was 0.0002. Both of these errors have been subjected to a Student's t Test (refer to column titled *t Stat*) and a p value assigned in the next column.

FIGURE 12

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9593							
R Square	0.9203							
Adjusted R Square								
Standard Error								
Observations	29							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression								
Residual								
Total								
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.9755	0.3232	12.3019	0.0000				
X Variable 1	-0.0027	0.0002	-17.6510	0.0000				

The hypotheses that Excel has tested are :

Is the Intercept significantly different from 0 ?

Is the slope significantly different from 0 ?

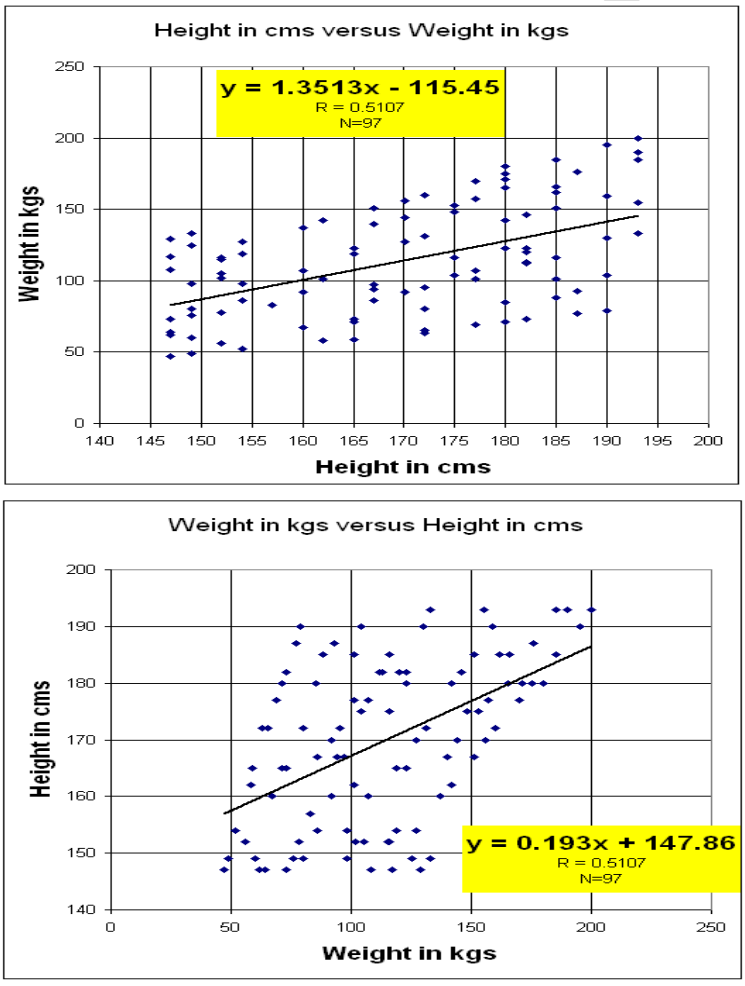
You should always consider whether these are appropriate hypotheses for testing your dataset. For example if your hypothesis is that an intervention should have no systematic effect (deleterious effect) on their blood pressure, then you would want the slope to be close to zero. In another scenario you may have expected the intervention to have simply

lowered everybody's blood pressure by 10mm then you would expect a non zero intercept of -10.

20: GEOMETRIC MEAN REGRESSION

Frequently we are not dealing with straightforward 'cause and effect' but with situations where one parameter is associated with changes in another. the co-association of body weight, height and age is a situation that most of us can identify with; body weight is positively associated with age and height but neither age or height are the sole drivers of body weight. So we are not justified in plotting body weight on the y axis and height on the x axis and then calculating the linear regression equation and then saying that describes the functional relationship between the two. This becomes clear when we do the reverse using the same dataset – plotting body weight on the x axis and and height on the y axis and calculating the regression equation.

FIGURE 13



From the equation for the top graph in Figure 13 the calculated weight for somebody 160 cms tall is 100.8 kgs. In comparison the calculated weight using the equation for the bottom graph is 62.9 kgs.

The solution to this anomaly is to use Geometric Mean (GM) Regression. This fits a line midway between the two lines that we have just calculated and as such corresponds closely to the 'eye's line of best fit'.

The calculations are straight forward

$$\text{GM Slope} = \sqrt{\frac{\text{Slope of y on x}}{\text{Slope of x on y}}}$$

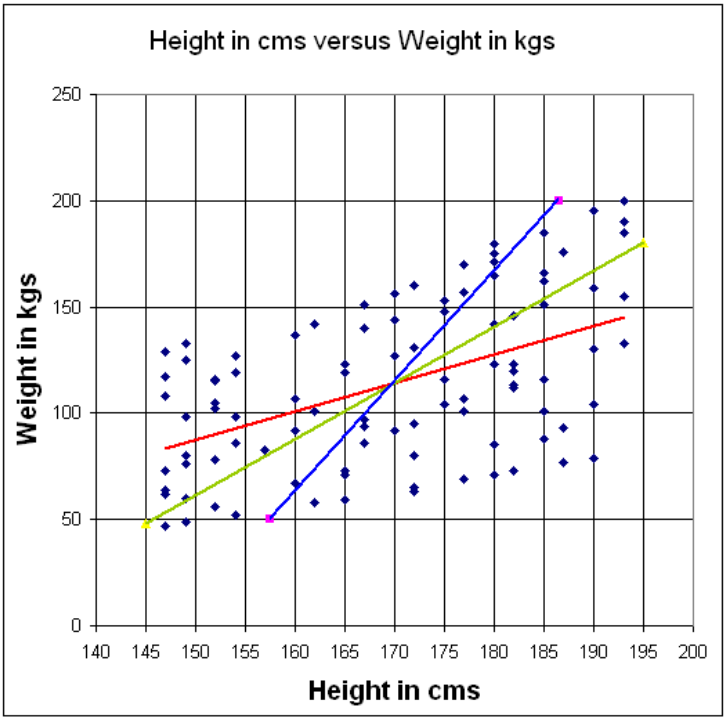
$$\text{GM Intercept} = \text{Mean of the y data} - \text{GM Slope X Mean of the x data}$$

It follows from the examples in Figure 13 that the GM Slope is $\sqrt{}$ (1.3513 / 0.193) = 2.646.
The means of the x data (heights) and the y data (weights) data were 169.9 cms and 114.1 kgs. So the GM Intercept is 114.1 – 2.646 X 169.9 which equals -335.5
The full Geometric Mean Regression Equation for these data is

Y = 2.646 X — 335.5

The corresponding line has been sketched in as a green line on Figure 14

FIGURE 14



A calculator in an Excel Spreadsheet that makes this calculation easier to perform is included on the website at :

www.medlabstats.com/alliedhealth

21: MULTIPLE LINEAR REGRESSION

Multiple Linear Regression is used when we have a 'y' variable and several 'types' of 'x' variables that we suspect all contribute to the changes in 'y'. For example the top speed of a vehicle 'y' is dependent upon the quality of the petrol (x_1), the size of the engine (x_2), the wind resistance of the body (x_3), the tyre pressures (x_4), the weight of the vehicle (x_5) etc. Another example from the healthcare area could be that the length of stay in hospital 'y' could be dependent *in a multidimensional linear model* manner upon the complexity of the surgery (x_1), the number of complications (x_2), the nutritional state of the patient before surgery (x_3), the number of physio sessions provided (x_4), the age of the patient (x_5), the nursing staffing levels in the ward (x_6) etc. The general form of a Multiple Linear Regression Equation is :

$$y = \text{Constant} + m_1 x_1 + m_2 x_2 + m_3 x_3 + m_4 x_4 + m_5 x_5 + m_6 x_6 \ .....$$

Clearly it is not possible to illustrate such relationships as a graph.

Here are some examples from the literature. Valeria Hirschler et al in their paper *'Can Waist Circumference Identify Children With the Metabolic Syndrome?'* produced a simple equation :

Insulin Resistance = 0.050 x Waist circumference
+ 0.033 x Systolic blood pressure
- 5.055

John Harnett et al in their paper '*Risk Factors for the Development of Left Ventricular Hypertrophy in a Prospectively Followed Cohort of Dialysis Patient'* they derived a six parameter multiple linear regression equation that related the left ventricular cardiac mass (LV mass) to the type of dialysis the patient received, their degree of hyperparathyroidism, patient's age, the duration of follow up, their systolic blood pressure and their mean blood haemoglobin level :

LV mass = 2
+ 7.6 x dialysis type
+ 2.2 x degree of hyperparathyroidism
+ 1.12 x age
- 0.05 follow up duration
+ 1.04 x systolic blood pressure
- 0.5 x mean haemoglobin

Paul Duerenberg et al in their paper *'Body mass index (BMI) as a measure of body fatness (BF%) : age- and sex- specific prediction formulas'* derived two MLR formulae, one for children

BF% = 1.4
+ 1.51 x BMI
- 0.70 x age
- 3.6 x sex

and one for adults

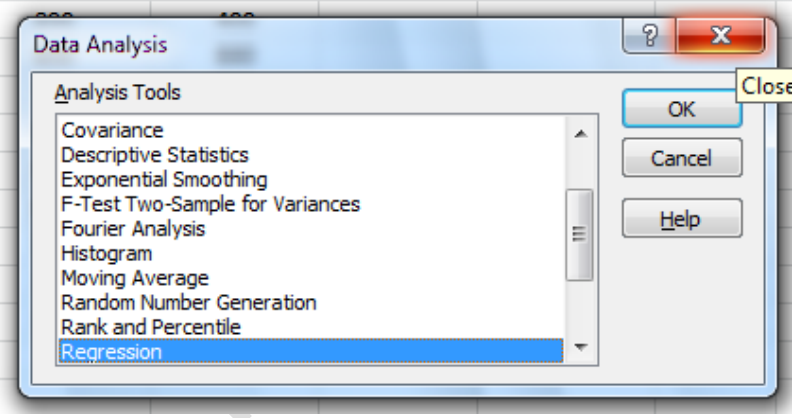
BF% = - 5.4
+ 1.20 x BMI
+ 0.23 x age
- 10.8 x sex

In their equations male =1 and female = 0.

Determining MLR equations can be done using the Regression Tool in Excel, see Figure.

FIGURE

Y value	X Variable 1	X Variable 2	X Variable 3
845.2	110	1000	200
750	200		
748.88	300		
698.6	400		
700	500		
596.5	610		
556.8	700		
560	800		
493	900		
610	990		



This produces an output as shown in Figure x. This analysis indicates that the MLR equation that best fits these data is

Y value = -2522.66
+ 2.64 x Variable 1
+ 3.09 x Variable 2
+ 0.0020 x Variable 3

Other points to note from this output are highlighted with yellow backgrounds. The first to note is the correlation coefficient (shown as *Multiple R*) and the *Significance F.* In this artificial example the significance value is extremely small indicating that the Regression Analysis has come up with a highly significant result.

FIGURE

Y value	X Variable 1	X Variable 2	X Variable 3
845.2	110	1000	200
750	200	890	400
748.88	300	800	880
698.6	400	700	1600
700	500	600	3200
596.5	610	480	6700
556.8	700	400	12800
560	800	300	25600
493	900	180	51200
610	990	100	102400

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9838
R Square	0.9679
Adjusted R Square	0.9519
Standard Error	24.0620
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	104818	34939	60	7.13366E-05
Residual	6	3474	579		
Total	9	108292			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2522.66	1284.91	-1.96	0.10
X Variable 1	2.64	1.18	2.24	0.07
X Variable 2	3.09	1.16	2.66	0.04
X Variable 3	0.0020	0.00	4.89	0.00

The second items to take note of are the *Coefficients* and the *P-value.* It is the latter which indicate whether or not a coefficient in the MLR equation is significant. On thIs first pass analysis it would seem that the *Intercept* and the coefficient of the *X Variable 1* are not significant because they are both greater than 0.05. This observation gives a justification to rerunning the regression analysis excluding the *X Variable 1* data. Figure xxxx shows that on excluding the *X Variable 1* data we do appear to get an MLR equation that has highly significant *Coefficients*.

FIGURE

Y value	X Variable 2	X Variable 3
845.2	1000	200
750	890	400
748.88	800	880
698.6	700	1600
700	600	3200
596.5	480	6700
556.8	400	12800
560	300	25600
493	180	51200
610	100	102400

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9701
R Square	0.9412
Adjusted R Square	0.9244
Standard Error	30.17
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	101921	50960	56	0.00005
Residual	7	6371	910		
Total	9	108292			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	350.9	37	8.85	0.00005
X Variable 2	0.4890	0.0546	8.96	0.00004
X Variable 3	0.001879	0.0005	3.73	0.007

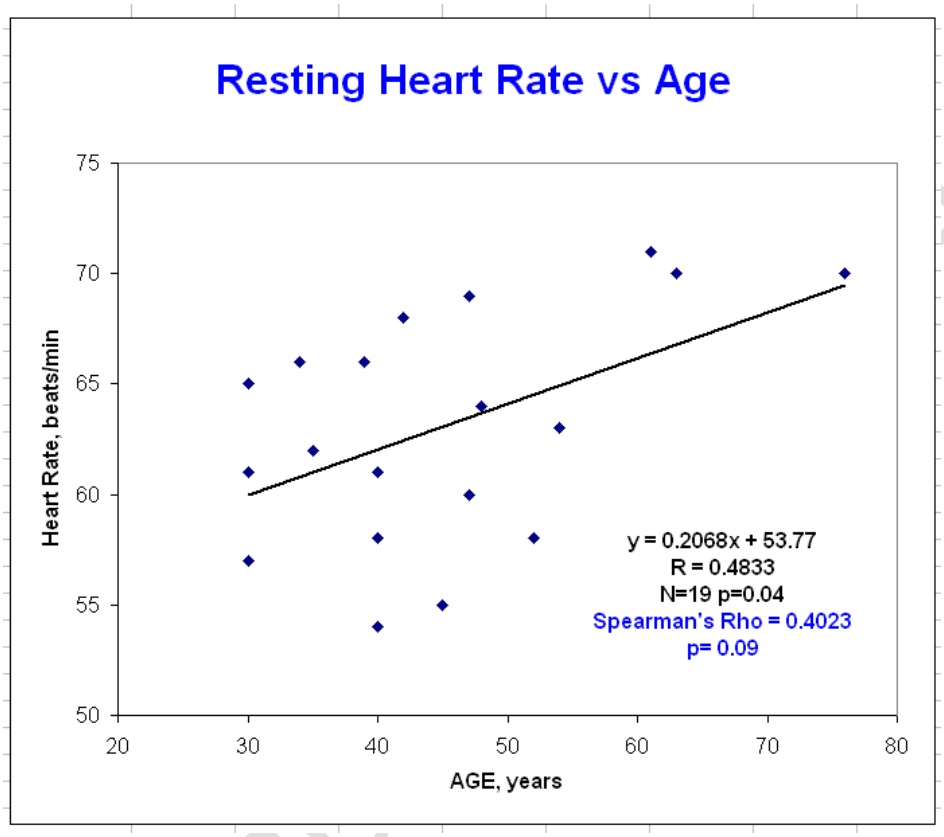
Y value = 350.9
+ 0.4890 x Variable 2
+ 0.0019 x Variable 3

It is this kind of stepwise exclusion of particular 'low significance' columns of data that leads researchers towards identifying the principal determinants of the Y value in their study.

22: RANK ORDER CORRELATION

This must be used whenever our 'y' data are not 'Normally' distributed in the 'y direction'. This is often the case when we have small datasets or are working on projects where we 'incrementally' increase x and so the we get a small number of 'y' values in 'response brackets'. Another situation where Least Squares Linear Regression provides spuriously good Correlation Coefficients is when you have extreme points in your dataset. Again Rank Order Correlation can nullify the undue influence of such points.

FIGURE xxx



The process of the Rank Order Correlation described by Spearman involves the examination of the differences of the ranks of the x data and the y data. There are two formulae in common use. The first formula only applies when there are no tied values within either your x or y datasets.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

The second formula must be used when there are ties in either the x or y data or both :

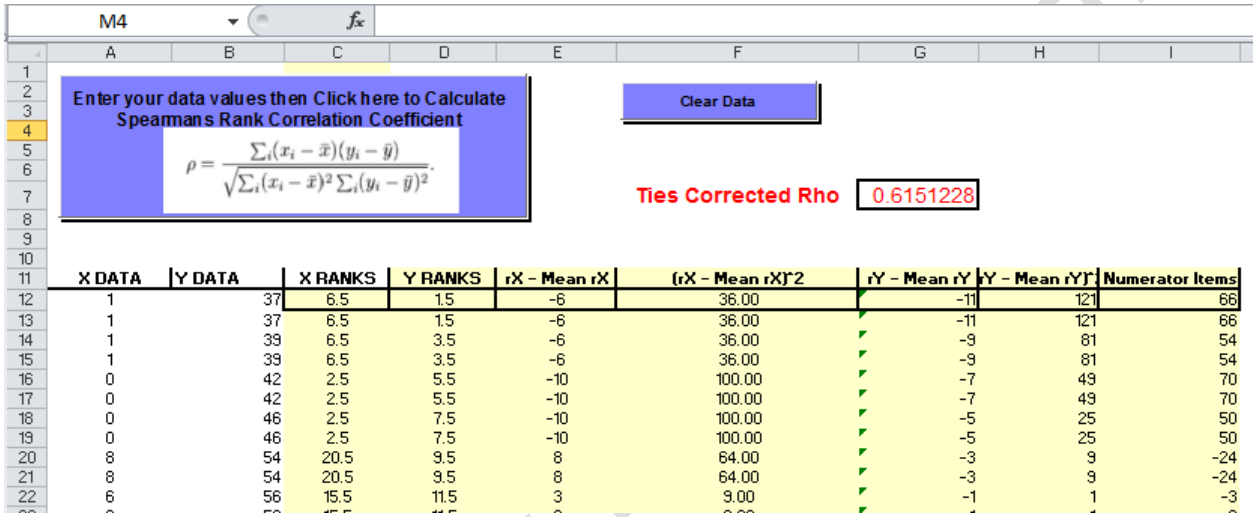
$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Note that in the first formula 'd' represents the differences in the ranks of the x,y pairs and n the number of x,y pairs in the dataset. In the second formula x, y, x bar and y bar are actually the ranks of the x and y data and the mean ranks of the x and y data. As with the correlation coefficient already described under Linear Regression, the Spearman Rank Correlation coefficient can take values between -1 and +1 with 0 indicating no correlation between the x and y data. -1 indicates perfect negative correlation between the x and y data and +1 indicates perfect positive correlation between the x and y data. An Excel spreadsheet based calculator is available on the website :

www.medlabstats.com/alliedhealth

This calculator, Figure xxx, always uses the ties compensated formula.

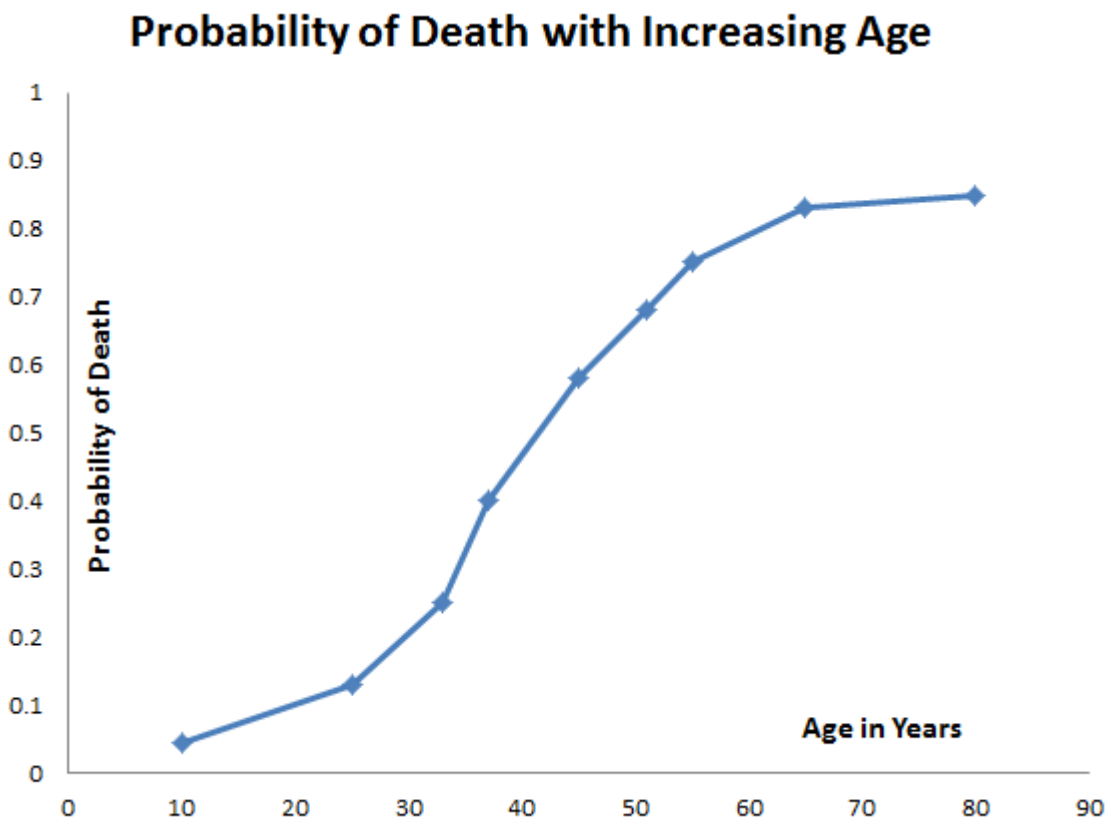
FIGURE xxxx



X DATA	Y DATA	X RANKS	Y RANKS	rX - Mean rX	(rX - Mean rX)^2	rY - Mean rY	(rY - Mean rY)^2	Numerator Items
1	37	6.5	1.5	-6	36.00	-11	121	66
1	37	6.5	1.5	-6	36.00	-11	121	66
1	39	6.5	3.5	-6	36.00	-9	81	54
1	39	6.5	3.5	-6	36.00	-9	81	54
0	42	2.5	5.5	-10	100.00	-7	49	70
0	42	2.5	5.5	-10	100.00	-7	49	70
0	46	2.5	7.5	-10	100.00	-5	25	50
0	46	2.5	7.5	-10	100.00	-5	25	50
8	54	20.5	9.5	8	64.00	-3	9	-24
8	54	20.5	9.5	8	64.00	-3	9	-24
6	56	15.5	11.5	3	9.00	-1	1	-3

23: LOGISTIC REGRESSION

This can be used to back calculate the probability of a particular binary outcome from a statistical model of a dataset. A typical example of a binary outcome is 'survived following treatment' versus 'died during treatment'. In Figure xxxx we can see how such a binary outcome viz. 'death' increases non-linearly with age.

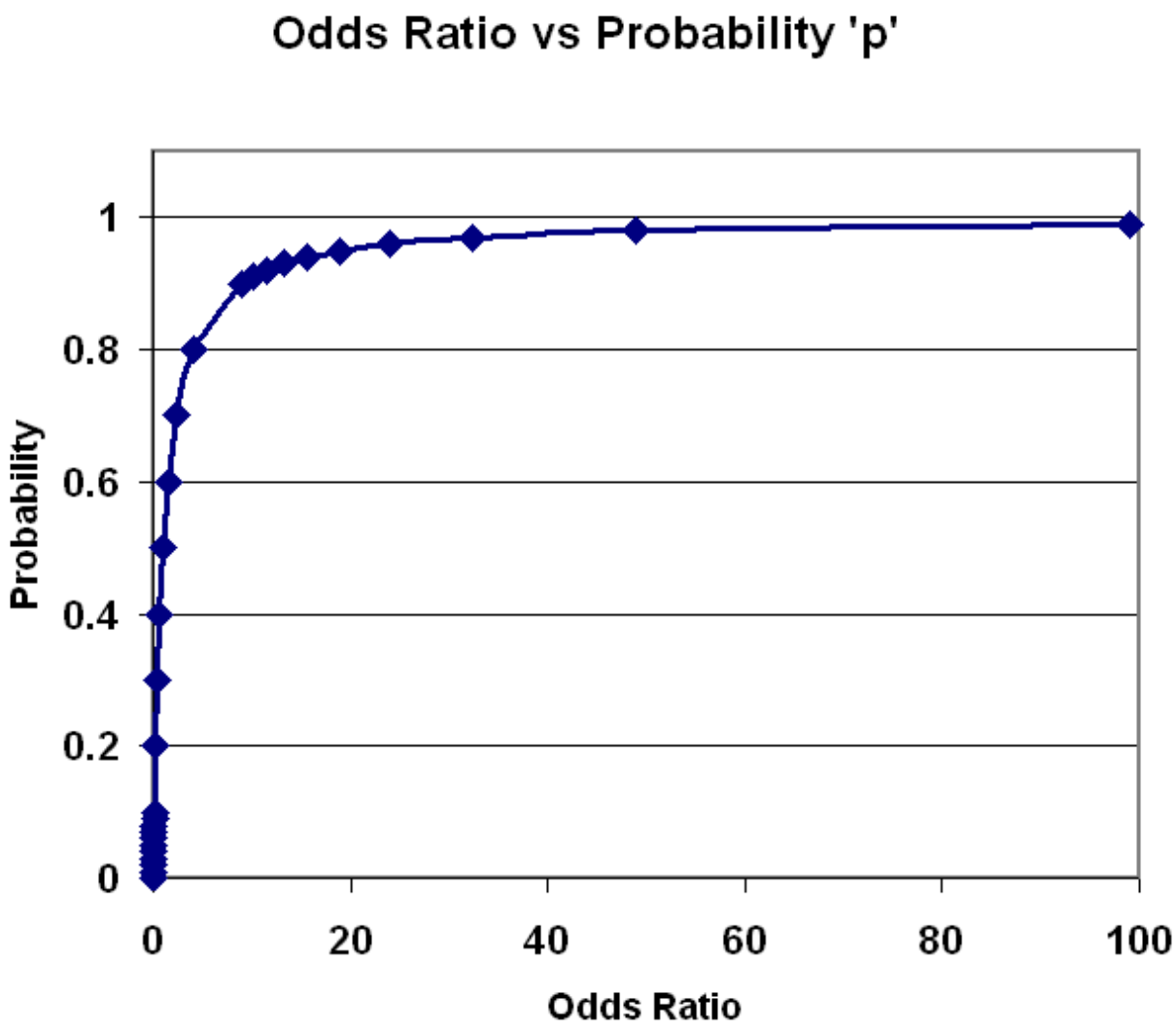


To begin the explanation of how logistic regression can be used to analyze such data and graphs it is best to begin with a discussion of the odds ratio, OR. In the binary outcome scenario we can draw up a table of the probabilities of an event happening, p, or not happening, q, such that p + q is always 1. The Odds Ratio is simply defined as p/q. When we plot the odds ratio in Table xxx versus 'p' then we get a hyperbola as shown in Figure xxxx

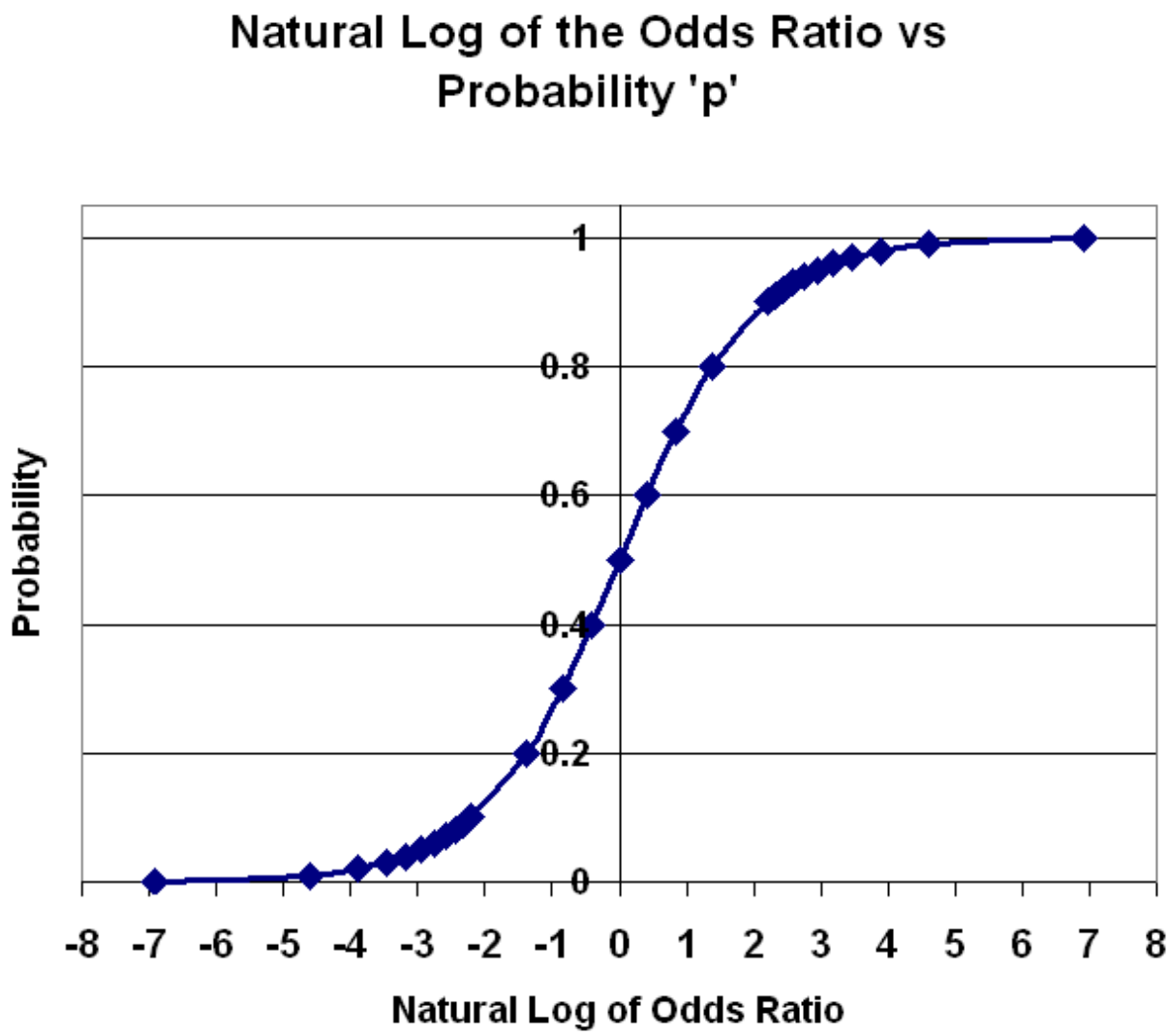
TABLE XXX

p	q	OR
0.001	0.999	0.001001
0.01	0.99	0.010101
0.02	0.98	0.020408
0.03	0.97	0.030928
0.04	0.96	0.041667
0.05	0.95	0.052632
0.06	0.94	0.06383
0.07	0.93	0.075269
0.08	0.92	0.086957
0.09	0.91	0.098901
0.1	0.9	0.111111
0.2	0.8	0.25
0.3	0.7	0.428571
0.4	0.6	0.666667
0.5	0.5	1
0.6	0.4	1.5

0.7	0.3	2.333333
0.8	0.2	4
0.9	0.1	9
0.91	0.09	10.11111
0.92	0.08	11.5
0.93	0.07	13.28571
0.94	0.06	15.66667
0.95	0.05	19
0.96	0.04	24
0.97	0.03	32.33333
0.98	0.02	49
0.99	0.01	99
0.999	0.001	999



The next manoeuvre is to plot the Natural Logarithm of the Odds Ratio versus the Probability 'p' and we get what is known as a sigmoid curve. Also it should be noted that in the process of taking the Natural Logarithm of the Odds Ratio we have actually determined the *'logit'* which you will always find mentioned throughout all discussions of logistic regression. The graph we get is shown in Figure xxxxx



The general equation to a sigmoid curve is of the form

$$Y = A / (1 + B\rho^{X})$$

If we assign 1 to A and ρ, and the exponential function to B we get a sigmoid that is known as the *Logistic* function.

$$Y = 1 / (1 + e^{-X})$$

If we substitute the p for Y and the Odds Ratio for X we get the equation

$$p = 1 / (1 + e^{-(p/q)})$$

Since we are usually only going to be interested in one probability term namely 'p' when using this equation to solve for various values of z, the q term is usually substituted with 1 – p

$$p = 1 / (1 + e^{-(p/(1-p))})$$

Some readers familiar with exponentials will recognise that the term

$$e^{(p/(1-p))}$$

is the Natural Logarithm of the Odds Ratio ! It therefore should come as no surprise that statisticians have latched onto this fact and gone further to model the Natural Logarithm of the Odds Ratio into a linear model equation :

$$\mathbf{Ln(p/(1-p)) = z = \beta_0 + \beta_1 x_1}$$

This implies that the odds ratio is driven by one factor namely $\mathbf{x_1}$ It turns out that this generalised model can be extended out to more than one driving factor in which case it takes on the form of :

$$\mathbf{z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 .....}$$

In both forms β_0 is a constant term and β_0, β_1, β_2, … are coefficients of each x parameter in the model.

The full equation for the multivariate logistic model is :

$$\mathbf{P = \frac{1}{1 + exp - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 ....)}}$$

In order to determine the values of $\mathbf{\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4}$ ..... that best fit your data you need to use a program that uses *maximum likelihood method.* This trials various values for these terms until it gets the minimum deviation between your observed dataset and what the models calculates. This can be done in Excel using the 'Solver' tool. Such an approach it is really only of interest to statisticians who wish to design Excel worksheets that are dedicated to solving the model for a particular experiment. By and large they are not 'user friendly' for practical logistic regression modelling by non-statisticians.

Having obtained values for $\mathbf{\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4}$ ..... that best fit your data you can then back calculate the probability p that would be expected with a particular set of values of $\mathbf{x_1\ x_2\ x_3\ x_4}$ **.....**

In Table xxxx there are data from fictional study of the effects of three dosage levels of aspirin in migraine sufferers. The dose of aspirin is the $\mathbf{x_1}$ term and the Odds Ratios are in the fourth column.

TABLE xxxx

X_1	Number of sufferers where migraine WAS relieved	Number of sufferers where migraine **WAS NOT** relieved	ODDS of getting relief on a particular dose
500 mg aspirin	400	50	(400/50) = 8
300 mg aspirin	360	390	(360/390) = 0.9231
100 mg aspirin	65	435	(65/435) = 0.1494

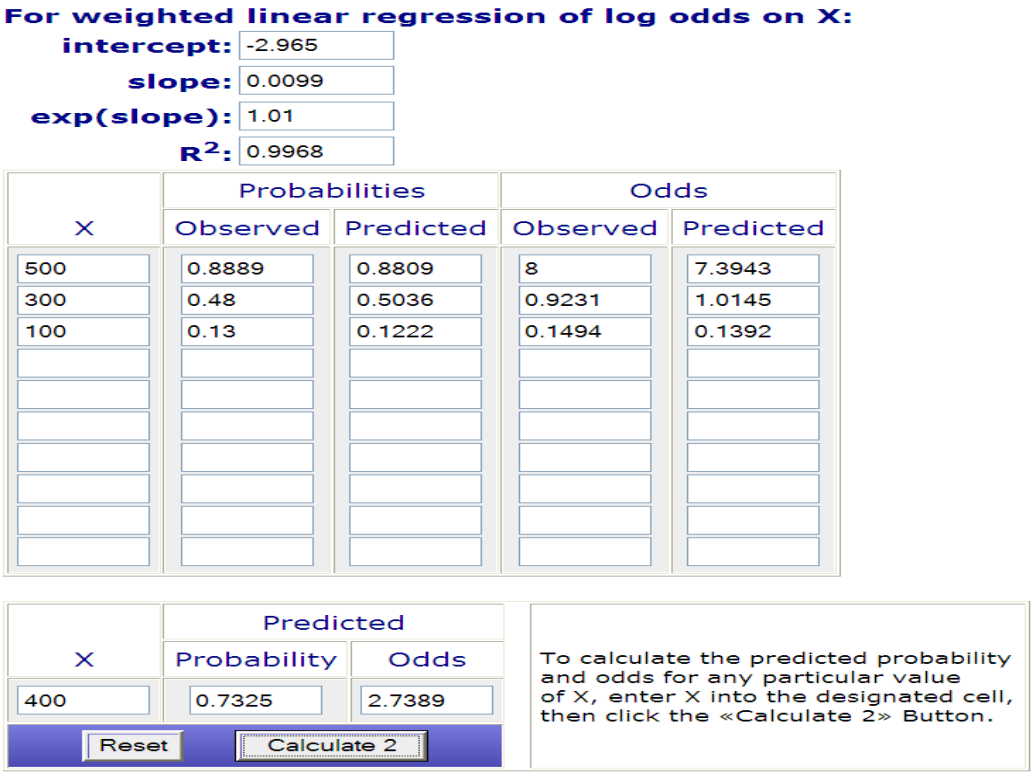
It would be interesting to know what the outcome would have been if we had included a 400 mg aspirin dose. Logistic Regression can provide the answer to such a question see Figure xxxx which is the result calculated using the Excel Solver function mentioned previously.

FIGURE XXXXX

	A	B	C	D	E	F	G	H	I
1	Aspirin Dose	Relieved	Not Relieved	Odds Ratio	Ln(Odds Ratio)	Calc Ln(Odds Ratio)	Diff^2		
2	500	400	50	8	2.079441542	2.022995512	0.00319	B0	-2.95249
3	300	360	390	0.92307692	-0.080042708	0.032802559	0.01273	B1	0.009951
4	100	65	435	0.14942529	-1.900958761	-1.957390393	0.00318		
5						Sum Diffs Squared	0.0191		
6									
7								Dose =	400
8								Calc Ln(Odds Ratio	1.027899
9								Calc Odds Ratio	2.795187
10								Calc p	0.736508

Comparison of the results of the Excel spreadsheet calculation results with those obtained from the Vassar Stats website were identical, see Figure xxxx :

FIGURE xxx



Bahman Tabaei and William Herman in their paper *A Multivariate Logistic Regression Equation to Screen for Diabetes* determined the following equation for z based on their data set :

z = -10.04 + 0.0331 x Age yrs
+ 0.0308 x Random plasma glucose mg/dL
+ 0.25 x Postprandial time, 0 – 8 hrs
+ 0.5620 x Gender
+ 0.0346 x BMI

Postprandial time was defined as the subject's estimate of the number of hours since their last meal at the time they had their blood glucose measured. Gender was scored as 0 for males and 1 for females.

William Cioffi et al in their paper *Prophylactic Use of High-frequency Percussice Ventilation in Patients with Inhalation Injury* derived this logistic regression equation :

Predicted mortality = - 4.8216
+ 0.10299 x PCTB
- 0.18898 x Age in yrs
+ 0.50873 x (Age^2 / 100)
- 0.27915 x (Age^3 / 10000)

where PCTB represented percent of total body surface burned

24: ANOVA (ANALYSIS OF VARIANCE)

In an earlier section we described the Paired and Unpaired Student's t Tests. You will recall that this was a test designed to determine if there was a significant difference between two means and that there was a strict proviso that before you could perform this test legitimately you needed to compare the variances for identity using the F-Test. Well in this section we are going to describe a test for determining whether or not there are significant differences between more than two means but again this test is only legitimate if the variances are sufficiently similar. This test is called the Analysis of Variance and usually is known by the acronym ANOVA. Another condition of the three ANOVA tests described here is that the samples must be completely independent of each other. There is a version of ANOVA called Two Way Repeated Measures ANOVA but this cannot be performed using Excel without an appropriate AddIn so it has been omitted from this text. . (At this point it is worth clarifying what the differences in statistical language between 'replicates' and 'repeated measures' – replicates mean that the same parameter is measured in a cohort of individuals once only. Repeated measures applies to the scenario where a cohort of individuals is subjected to a series of different interventions and during each intervention the same parameter is measured. For example a repeated measures study could have involved measuring maximum exercise capacity on enrolling in a keep fit programme then again at six weeks into the programme and then again six weeks after they have left the programme.)

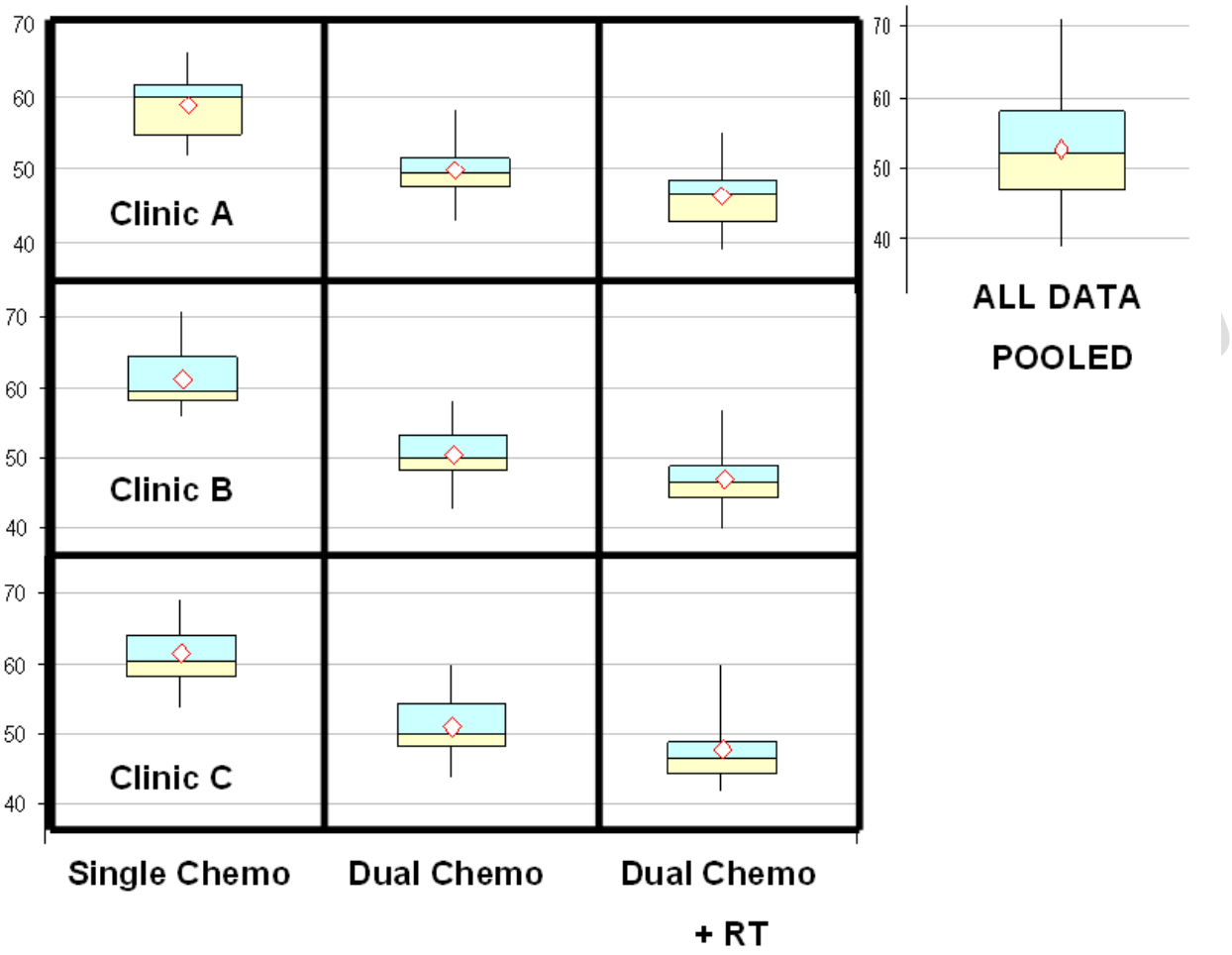
Table xxxx is a typical dataset that is amenable to ANOVA. It describes some data collected during a hypothetical clinical trial of new single drug chemotherapy agent versus a couple of conventional chemotherapy modalities – a dual chemotherapy drugs protocol and the same dual chemotherapy drugs plus sessions of radiotherapy (RT). In this design it is clear that independent observations have been made because we have three groups of 30 different patients attending three different cancer clinics : A, B and C. As each new patient arrived they were allocated by the clinic to only one treatment regimen. Following their treatment they were followed up and the time that it took for their cancer to reappear was measured in weeks ie the table shows how many disease free weeks each patient had before they had a relapse.

TABLE xxxx

	A	B	C	D
1	CLINIC	SINGLE CHEMO	DUAL CHEMO	DUAL CHEMO + RT
2	A	52	43	39
3	A	53	45	42
4	A	54	47	42
5	A	57	49	45
6	A	60	49	46
7	A	60	50	47
8	A	60	50	47
9	A	62	52	49
10	A	64	56	52
11	A	66	58	55
12	B	56	43	40
13	B	57	46	42
14	B	58	48	44
15	B	59	49	45
16	B	59	50	46
17	B	60	50	47
18	B	60	51	48
19	B	66	54	49
20	B	66	56	52
21	B	71	58	57
22	C	54	44	42
23	C	58	47	42
24	C	58	48	44
25	C	59	49	45
26	C	60	50	46
27	C	61	50	47
28	C	62	52	48
29	C	65	55	49
30	C	68	56	54
31	C	69	60	60

Confronted with this moderate amount of data that a relatively simple clinical trial generates it is not easy to decide if the dataset complies with the pre ANOVA test requirement that the variances of the data within each of the nine combinations of clinic and treatment regimen are comparable. With only 10 patients per subset in this example, estimations of each variances is likely to be an overestimate of the 'true variance; (it is not until you get 30 data points per subset can you expect to get variances that are representative enough for you to be able to mentally assess them for inter-subset equivalence). So in situations such as the one shown in Table xxx a common recommendation is to make dot plots or box and whisker plots of each subset of the data collection and perform visual inspections of them. The corresponding box and whisker plots of each of the nine subsets are shown in Figure xxxx along with a box and whisker plot of all the data pooled as one set of 90 points. Visually we can see that the relative spans of each data set is comparable apart from perhaps the subset from Clinic C for the Dual Chemotherapy with Radio Therapy. That group has a reduced first quartile and an extended fourth quartile. Nevertheless the data collection as shown in the 'All data Pooled' box and whisker is gives the impression that overall the study has produced a reasonably symmetrical dataset.

FIGURE xxxxx

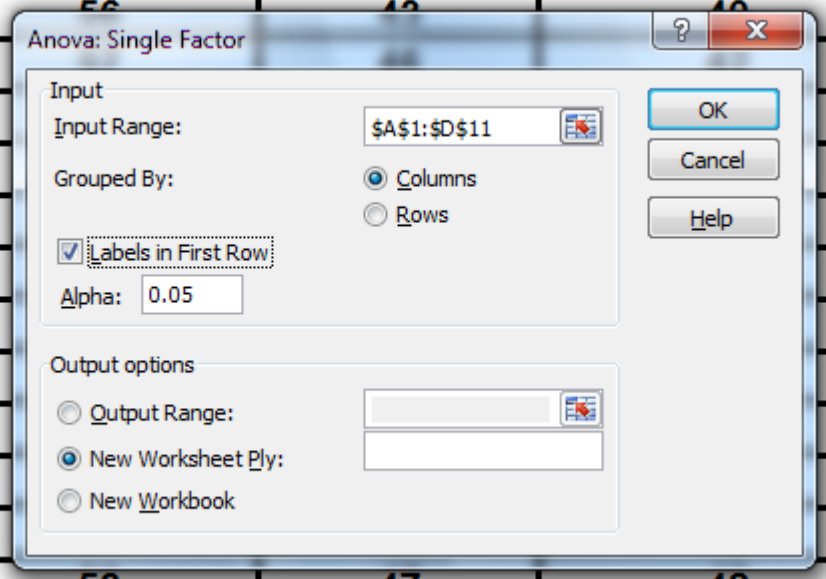


Excel comes with three ANOVA tools built in : ANOVA : Single Factor, ANOVA Two factor with replication and ANOVA Two factor without replication. The words factor are used to describe what is happening in the x and y directions of our data tables. In Table xxxx the factor in the x direction is the treatment regimen and the factor in the y direction is the clinic where the treatment is initiated.

Suppose we work in Clinic A. We can use a ANOVA : Single Factor test to determine for ourselves the answer to the question – Are there any differences in the times to relapse amongst our study group of thirty patients ? The ANOVA tool in Excel is located under the Data : Data Analysis tabs and the dialog box that opens up there is easy to complete; we just select cells D1 to D11 because the contain the data collected by our Clinic A; see Figure xxxxx.

FIGURE xxxxxxx

	A	B	C	D
1	CLINIC	SINGLE CHEMO	DUAL CHEMO	DUAL CHEMO + RT
2	A	52	43	39
3	A	53	45	42
4	A	54	47	42
5	A	57	49	45
6	A	60	49	46
7	A	60	50	47
8	A	60	50	47
9	A	62	52	49
10	A	64	56	52
11	A	66	58	55
12	B			
13	B			
14	B			
15	B			
16	B			
17	B			
18	B			
19	B			
20	B			
21	B			
22	C			
23	C	58	47	42



The results of the ANOVA single factor analysis is shown in Figure xxxxx. In the Summary Table it shows that indeed our varaiances were comparable at 22.18, 20.99 and 23.16 for the three treatments, and in the ANOVA Table it shows that the p value for our differences between the three means was very considerably smaller than 0.05. So we can immediately conclude that yes there was a significant difference(s) between the means of the times to relapse. Unfortunately it gives no indication as which mean was significantly different from the others. The only way to get an answer to this is to go to a website that can apply the Tukey Honest Significant Difference (HSD) test; in this case we have used the one available on *vassarstats.net*. This revealed that there was a significant difference between the mean for the single chemotherapy regimen

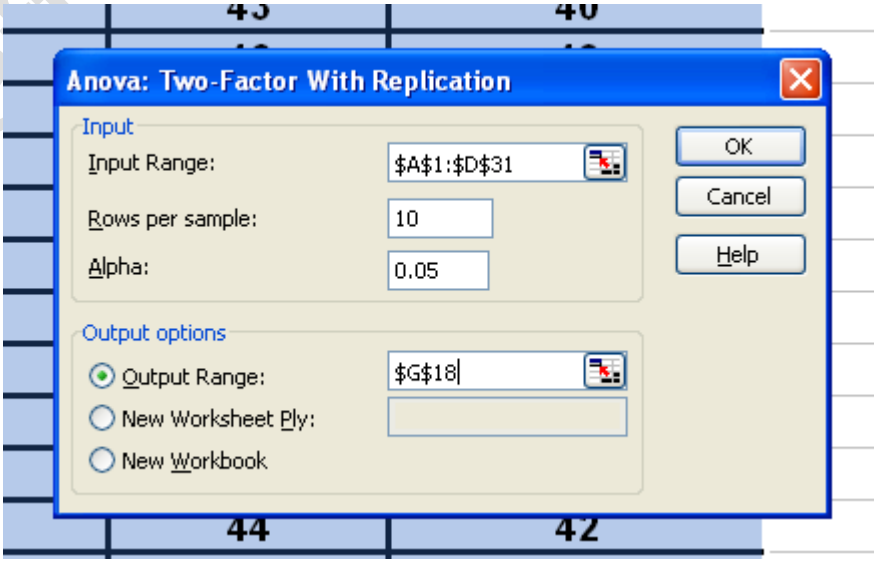
- and the mean for the dual chemotherapy regimen
- and the mean for the dual chemotherapy plus radiotherapy regimen

Anova: Single Factor	CLINIC A					
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
SINGLE CHEMO	10	588	58.8	22.18		
DUAL CHEMO	10	499	49.9	20.99		
DUAL CHEMO + RT	10	464	46.4	23.16		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	817.4	2	408.7	18.49	8.76E-06	3.35
Within Groups	596.9	27	22.11			
Total	1414.3	29				

Tukey HSD Test
HSD[.05]=5.22; HSD[.01]=6.7
M1 vs M2 P<.01
M1 vs M3 P<.01
M2 vs M3 nonsignificant

Suppose now you are the chief investigator in this clinical trial. You are more interested in two intertwined questions – Were there significant differences between the treatments and were these differences due to the treatments themselves or was there a contributory factor from the ways in which the clinics applied the treatment protocol? In this scenario ANOVA two factor with replication is the tool that can go some way towards answering this question. The dialog box for this test in Excel is self-explanatory – we select all column and rows in the table including the headers and row names, see Figure xxxxx

FIGURE XXXXX



The output for this analysis is quite extensive because summary data are presented for each Clinic before the final ANOVA Table; see Figure xxxxx. It is worthwhile scanning the variances obtained for the clinics (ie the blocks of rows) and checking that they do indeed look comparable. Once satisfied that they are acceptable then it is legitimate to examine the p values in the ANOVA table. For 'Sample' in this table we read Clinic for our example and for columns we read 'chemotherapy treatments'. The answer to the second part of our intertwined question *was there a contributory factor from the ways in which the clinics applied the treatment protocol?* Is given on the first line of the ANOVA table …. the p value was 0.38. So there was no inter clinic variability contributing to the differences between the chemotherapy regimens. The second line of the report shows the p value of $4 * 10^{-17}$ which means that there was a highly significant difference between the means of time to remission when compared by chemotherapy regimen. Again there is no indication which pairs or combinations of means are significantly different. Websites such as www.wessa.net do have online calculators which provide the results of the multiway Tukey Honest Significant Difference test. In this example that report runs to fortytwo lines of which twentyone lines specify combinations that are significant at the $p <= 0.05$ level eg. Clinic B's dual chemotherapy mean (50.5) is significantly different from Clinic A's single chemotherapy mean (58.8). Going into such detail is usually not required. Instead, having been assured that there was no significant inter-clinic effect, you can pool the data from the clinics and recalculate an ANOVA : single factor.

FIGURE xxxx

Anova: Two-Factor With Replication

SUMMARY	SINGLE CHEMO	DUAL CHEMO	DUAL CHEMO + RT	Total
CLINIC A				
Count	10	10	10	30
Sum	588	499	464	1551
Average	58.8	49.9	46.4	51.7
Variance	22.18	20.99	23.16	48.77
CLINIC B				
Count	10	10	10	30
Sum	612	505	470	1587
Average	61.2	50.5	47	52.9
Variance	23.29	20.50	24.22	58.85
CLINIC C				
Count	10	10	10	30
Sum	614	511	477	1602
Average	61.4	51.1	47.7	53.4
Variance	22.27	22.54	31.34	58.73
Total				
Count	30	30	30	
Sum	1814	1515	1411	
Average	60.47	50.50	47.03	
Variance	22.46	20.12	24.72	

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	45.8	2	22.9	0.98	0.38	3.11
Columns	2918.07	2	1459.03	62.38	4.00E-17	3.11
Interaction	11.73	4	2.93	0.13	0.97	2.48
Within	1894.4	81	23.39			
Total	4870	89				

Excel has an ANOVA Two factor without replication tool and its use can be demonstrated on a subset of this example viz Figure xxxx which shows the longest times patients in each clinic went before relapse. The dialog box in Excel is self explanatory and the results of the ANOVA showed that there was no significant differences between the row means ie the Clinics (p=0.14) but there was asignificant difference between the means for the therapies] (p=0.00). Again no information is given about which pairs of means were significantly different.

	LONGEST TIME TO REMISSION		
CLINIC	SINGLE CHEMO	DUAL CHEMO	DUAL CHEMO + RT
A	66	58	55
B	71	58	57
C	69	60	60

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
A	3	179	59.67	32.33
B	3	186	62	61
C	3	189	63	27
SINGLE CHEMO	3	206	68.67	6.33
DUAL CHEMO	3	176	58.67	1.33
DUAL CHEMO + RT	3	172	57.33	6.33

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	17.56	2	8.78	3.36	0.14	6.94
Columns	230.22	2	115.11	44.09	0.00	6.94
Error	10.44	4	2.61			
Total	258.22	8				

So in summary ANOVA provides the appropriate tool for comparing means from data that are distributed with comparable variances and that are based on a collection of singlet measurements from different cohorts of patients that are suspected to be different because of their different diagnoses or treatment. The tests in Excel do not assist with the identification of what combinations of means are different – Tukey's HSD test has to be used to answer that question.

25: THE DESIGN OF QUESTIONNAIRES

The purpose of this section is to introduce a discussion of issues around questionnaires that you might want to use to collect data, The tasks involved include

- Composing your questionnaire
- Getting the answers into electronic form
- Analysing the responses
- Turning the results of your analyses into 'statements' and 'rules'

Composing your questionnaire:

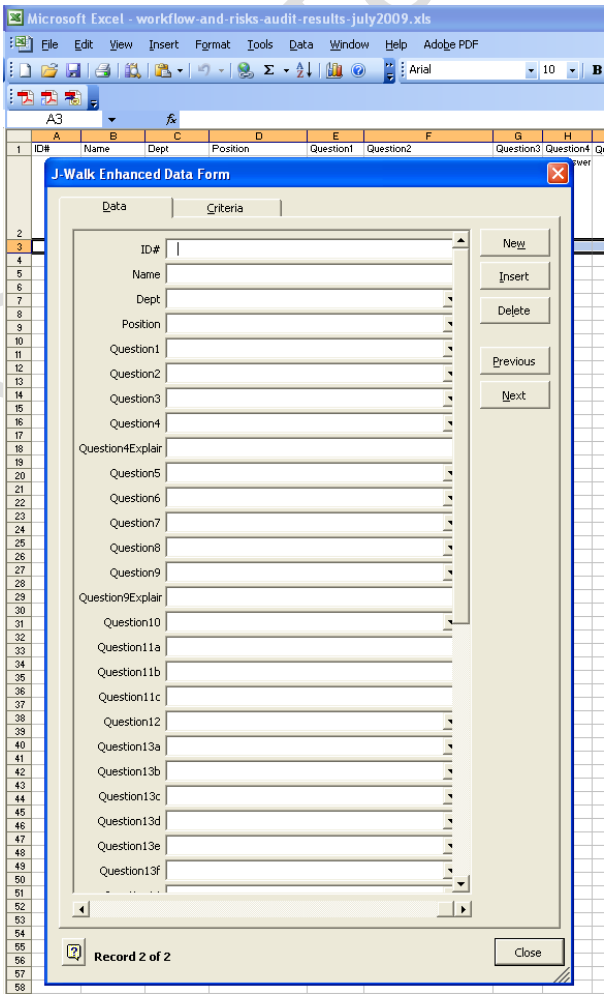
- write it in unambiguous language
- use language and terms that your target participants understand
- always ensure that you give them "not relevant to me" options for every multi-choice question

- when “marking” a questionnaire make sure you have a “no answer” option for every multi-choice question. This will get over the problem that some statistical tests have with fields where they are “missing” data.
- Some answers will be binary eg smoker ? y/n. These are usually easily coded as 0 and 1 for ‘False’ and ‘True’
- Some answers will be categorical eg. gender, age bracket. These are best coded as 1, 2, 3 ….
- Some answers will be quasi quantitative eg Pain relief Poor ----------Excellent where you want the respondent to mark where they think they are along this line (which in reality represents a continuum of sensation). In these situations make sure you provide a line between the two extremes of at least 5 centimetres. You can then ‘quantitate’ the response by measuring the distance from the lower extremity to the respondents mark in millimetres.

Getting the answers into electronic form

There is a freeware Excel Addin called the J-Walk Enhanced Data Form addin which can make the task of transcribing from paper questionnaire to spreadsheet less labour intensive and less prone to error. Essentially you label each column in your spreadsheet with the question titles. When the J-Walk addin is activated it then gathers this information together into the table style shown in Figure xxxxx

FIGURE xxxx

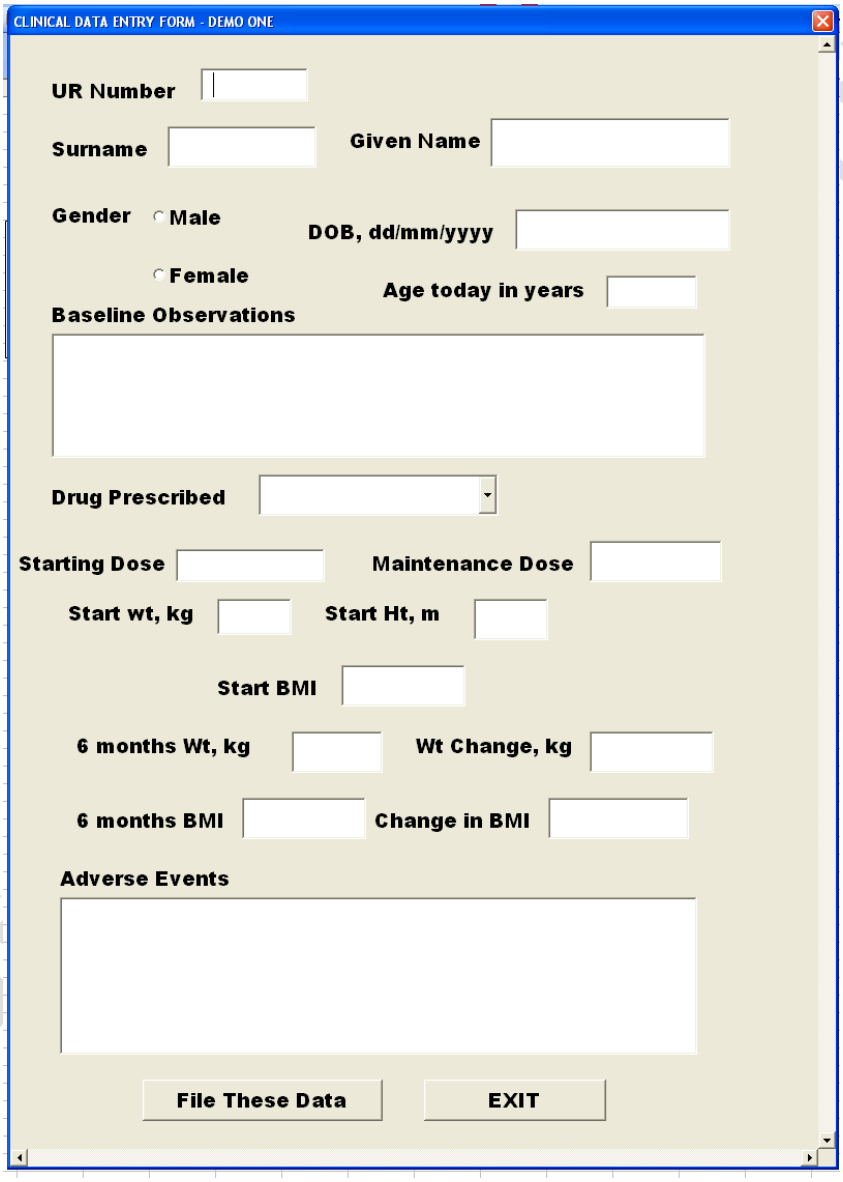


Appendix Nine contains a copy of the instructions on how to use the the J-Walk Enhanced Data Form reproduced from John Walkenbach's website :

spreadsheetpage.com/index.php/dataform/home

Note that there is one version for use with Excel from version 97 to 2003 and another version to use with Excel versions 2007 – 2010.

If this vertical style of data entry form is not to your liking then you have the option of using VBA in Excel to create your own data entry form; (VBA = Visual Basic for Applications). This is versatile enough for you to be able to reproduce the layout of your questionnaire or in fact any style of data collection form, see Figure xxxxxx



Appendix Ten contains the VBA code used to generate this form as well as an article written by Martin Green of www.fontstuff.com which describes the fairly straightforward process of making your own data entry forms for Excel.

Turning the results of your analyses into 'statements' and 'rules' :

- Quantitative data, e.g age, weight, hours worked can be analysed using parametric statistics provided they obey the 'normal distribution' rule.

- Non-normal quantitative data can be analysed using non-parametric statistics.

- The questionnaire should have been designed in such a way that you expect to be able to compose Quantitative Statements from the final analysis eg.

  "Patients who had Type 2 Diabetes were significantly heavier by on average 7.2 kgs". This type of statement is based on parametric statistical analysis of a 'size' questionnaire field.

- The questionnaire should have been designed in such a way that you expect to be able to compose some Underlying Rules from the final analysis eg.

  "75% of Type 2 diabetes were women aged over 48 years, living in communities which were more than 25kms from a G.P and had a family income of less than $32,000 p.a." This type of 'rule' is based on the non-parametric analysis of a number of apparently related (correlated) fields in the questionnaire.

25.1 : ASSESSING THE "QUALITY" OF YOUR QUESTIONNAIRE – CRONBACH'S ALPHA

25.2: EXTRACTING THE 'MEANING' FROM QUESTIONNAIRES

It would be nice to think that our questionnaires are so well designed that we will immediately be able to extract a 'story' from our observations. Unfortunately we are unlikely to have had a very clear idea of the factors which affect our respondents replies to our questions ie. we do not have a highly formulated model of how our respondents are affected by their condition(s) in mind when we designed the questionnaire. So are there some statistical tools to assist ? The two I want to introduce to you are Bayesian Networks and CHAID analysis. But first we need to understand some theory.

Combinations of Probabilites :

If an event happens because of a two contributing factors then the probability of them both occurring is the PRODUCT of the individual probabilities :

Probability of event A occurring = p(A)
Probability of event B occurring = p(B)

Then the probability of A AND B occurring is :

$$p(A) * p(B)$$

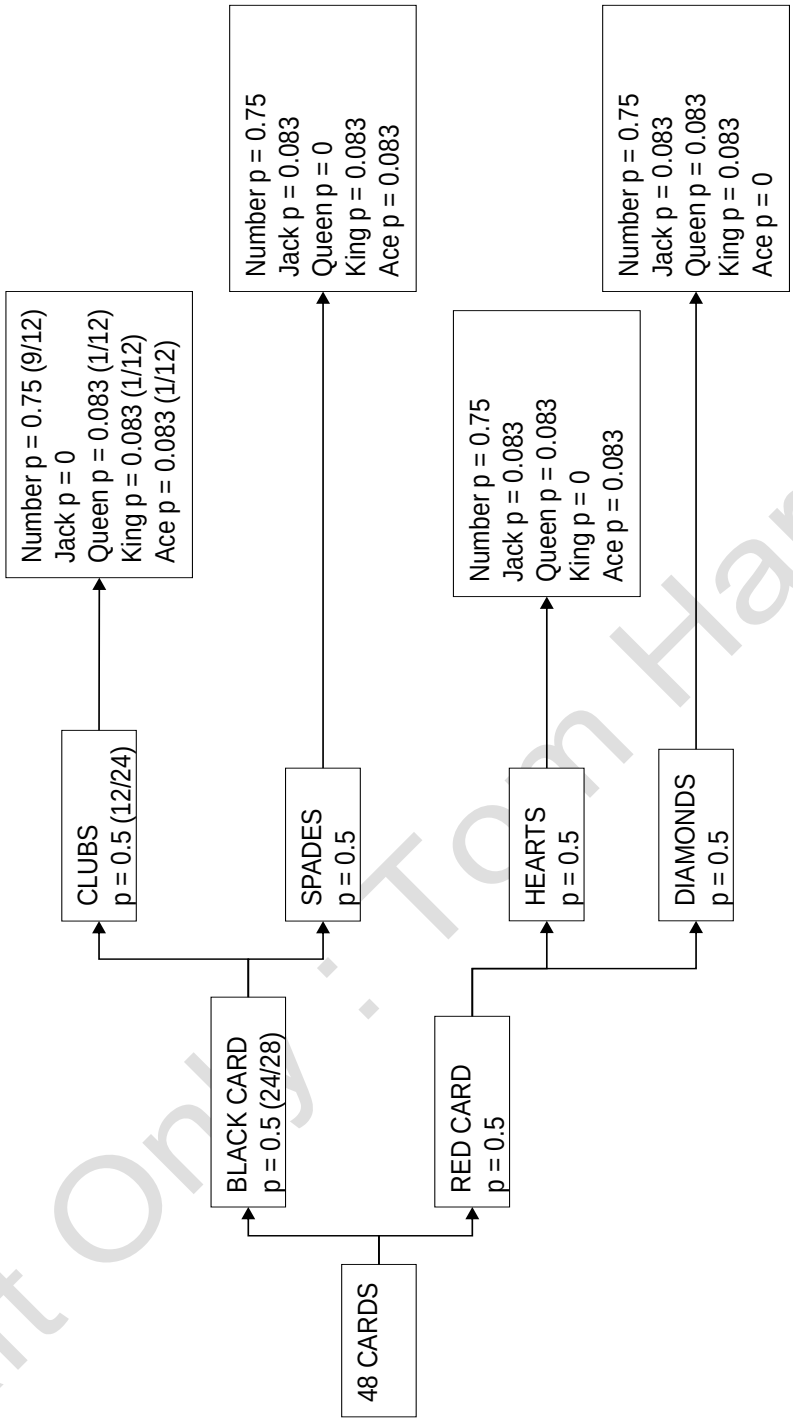
If an event occurs because event A **OR** B **OR** C **OR** D …… occur then the probability of either of occurring is the SUM of the individual probabilities :

$$p(A) + p(B) + p(C) + p(D) + \ldots\ldots = 1$$

These two rules enable us to build DECISION TREES :

26: PROBABILITY (DECISION) TREES

Example : Consider a pack of playing cards which has had the following picture cards removed from it : Jack of Clubs, Queen of Spades, King of Hearts, Ace of Diamonds. The diagram below gives the various sequential probabilities of picking a black, red, club, spade, heart, diamond, jack, queen, king, and ace out of this pack of cards that has been 'tampered with'



This decision tree works like this : Take a card at random – look at it is it Black or Red ? If Black then is it a Club card or a Spade card ? If it is a Club card is it a number card, a queen, king or an ace ?

From this diagram and a knowledge of the AND rule relating to sequences of probabilities we can see that the probability of randomly selecting the Queen of Hearts is :

0.5 * 0.5 * 0.08333 = 0.02083 which when expressed as a fraction is 1 in 48

That most of us do without having drawn up this decision tree. But working backwards is not so intuitive to most of us. For example I have taken a card at random and I want to know what the chances are that it is a Number card in the Club suit ?

This is where Bayes Theorem comes into use.

26.1 BAYES THEOREM

If an event happens because of a two contributing factors then the probability of them both occurring is the PRODUCT of the individual probabilities :

Probability of event A occurring = p(A)
Probability of event B occurring = p(B)

Then the probability of A **AND** B occurring is :

$$p(A) * p(B)$$

Bayes Theorem uses this property to enable us to solve problems of 'sequences of probabilities' **backwards** !

Imagine two events A and B that we intuitively feel seem to be associated with each other. Let the two events be A and B then Bayes Theorem states that :

The Probability that B will occur given that A has occurred

= Probability that A will occur given that B has occurred
multiplied by the Probability of the occurrence of B
divided by the Probability of the occurrence of A

You will usually see this written as :

$p(B \mid A) = p(A \mid B) * p(B) / p(A)$

In practice you need to be quite skilled to formulate your problem so that you correctly identifiy what is A and what is B in your study. My recommendation is that will probably find Bayes Theorem easier to apply if you use DECISION TREES. Under those terms Bayes Theorem reads as :

$$\frac{\text{Probability of arriving at your Target Outcome}}{\text{Sum of all the Probabilities of arriving at the SAME outcome but by all possible routes}}$$

Now return to the card question : I have taken a card at random and I want to know what the chances are that it is a Number card in the Club suit ?

What is the probability along the direct path to a Number card in Clubs :

0.5 * 0.5 * 0.75 = 0.1875

There are three other sets of number cards in the pack which I could also have picked out and the probabilities to those are also the same along their direct paths :

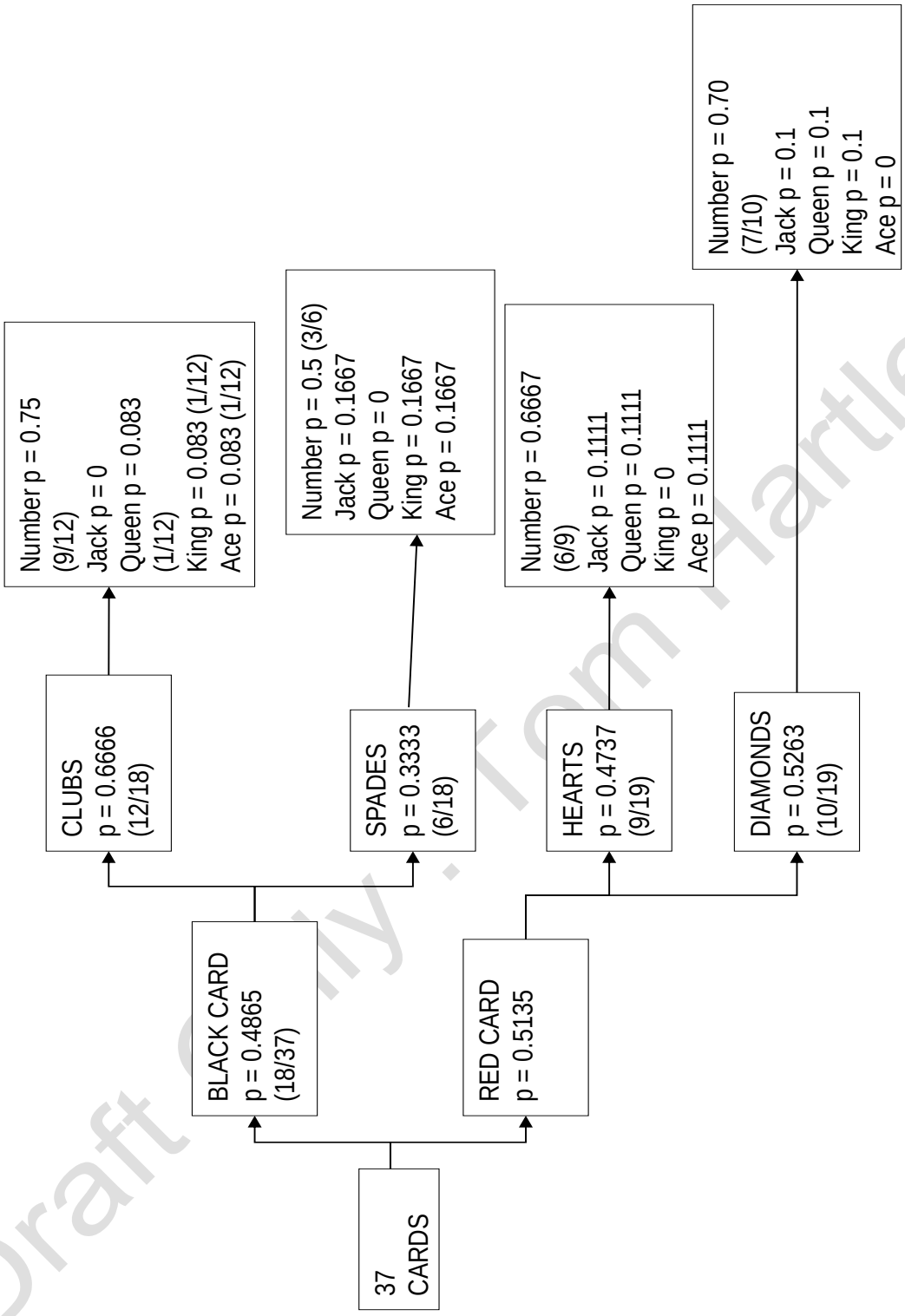
Number card in Spades = 0.5 * 0.5 * 0.75 = 0.1875
Number card in Spades = 0.5 * 0.5 * 0.75 = 0.1875
Number card in Spades = 0.5 * 0.5 * 0.75 = 0.1875

Bayes Rule can now be used to calculate the probability of having picked out a Number card in Clubs

0.1875 / (0.1875 + 0.1875 + 0.1875 + 0.1875)

which when expressed as a fraction is 1 in 4

Actually this answer could have been arrived at by most of us without Bayes formula because the decision tree is highly symmetrical. Consider the same question if I had removed 6 of the Spades number cards, 3 of the Hearts number cards and 2 of the Diamond number cards. The Decision Tree changes to this :



Direct route probability to a Number Card in Clubs is now 0.4865 * 0.6666 * 0.75 = 0.2432

The route to a Number Card in Spades is now 0.4865 * 0.3333 * 0.5 = 0.0811
The route to a Number Card in Hearts is now 0.5135 * 0.4737 * 0.6667 = 0.1622
The route to a Number Card in Diamonds is now 0.5135 * 0.5263 * 0.7 = 0.1892

Solving Bayes Theorem equation : 0.2432 / (0.2432 + 0.0811 + 0.1622 + 0.1892) = 0.3599 which when expressed as a fraction is 9 in 25
For completeness check this against the formal Bayes Theorem equation -

p(B | A) = probability that it is a number card given that it is a Club

p(A | B) = probability that it is a number card given that it is a Club = 9/12

p(A) = probability that it is a number card = 25 / 37

p(B) = probability that it is a Club = 12 / 37

p(B | A) = (9/12) * (12/37) * (37/24) = 9/25

TURNING TABULATED DATA INTO DECISION NETWORKS

Suppose we have conducted a survey of 250 people buying cars. We have asked them if they bought a second hand or a new car and we have also asked them what income bracket they thought they came from – Low, Middle or High

The tabulated data looked like this :
Table A

INCOME GROUP	**LOW**	**MEDIUM**	**HIGH**	**TOTAL**
Number of Customers	**60**	**110**	**80**	**250**
Proportion of Customers	**0.24**	**0.44**	**0.32**	**1.00**

Table B

INCOME GROUP	**LOW**	**MEDIUM**	**HIGH**	**TOTAL**
Number of **S/Hand** Cars purchased	**45**	**65**	**30**	**140**
Number of **New** Cars purchased	**15**	**45**	**50**	**110**
TOTALS	**60**	**110**	**80**	**250**

Table C : Proportions by Column of Table B

INCOME GROUP	**LOW**	**MEDIUM**	**HIGH**
Number of **S/Hand** Cars purchased	**0.75**	**0.59**	**0.38**
Number of **New** Cars purchased	**0.25**	**0.41**	**0.62**
TOTALS	**1.00**	**1.00**	**1.00**

With these data we can build a Decision Network diagram and ask questions that are soluble via Bayes Theorem.

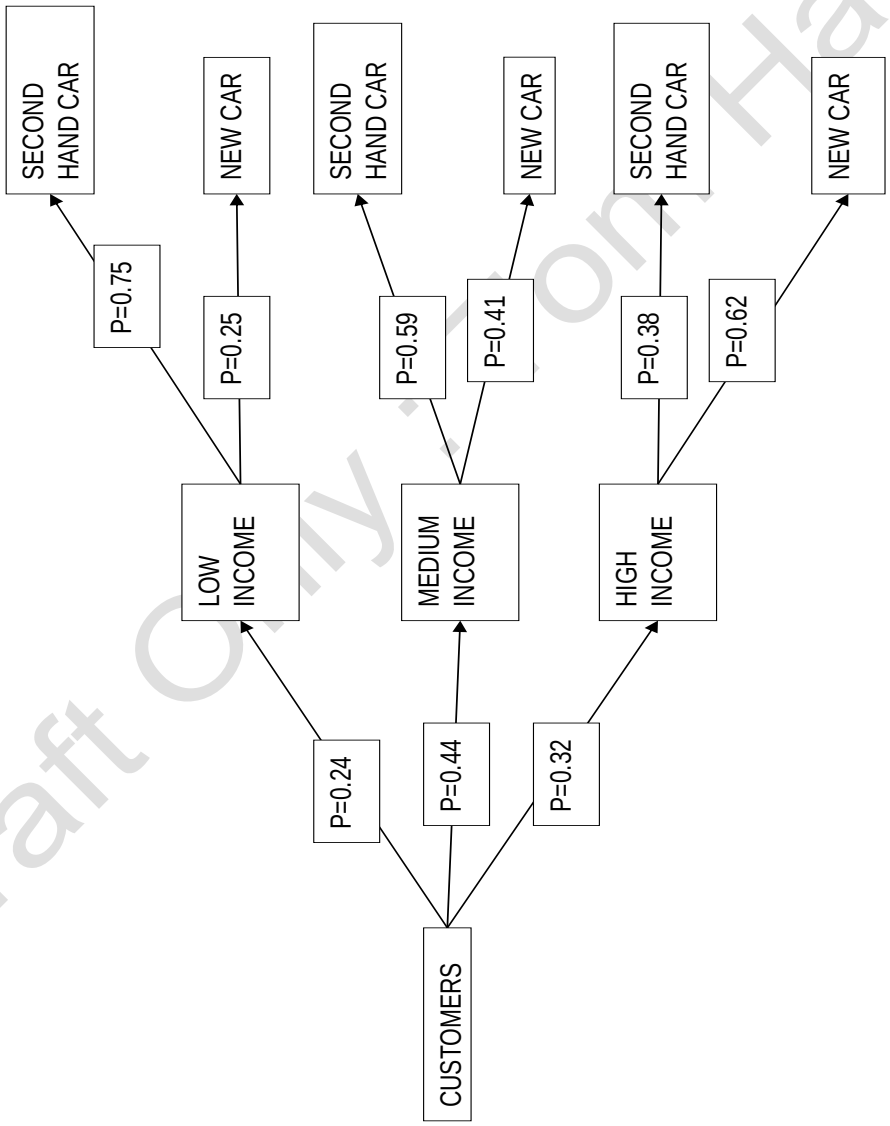
What is the probability when selling a Second Hand car it is purchased by a Middle Income customer ?

Direct Route probability = 0.44 * 0.59 = 0.2596

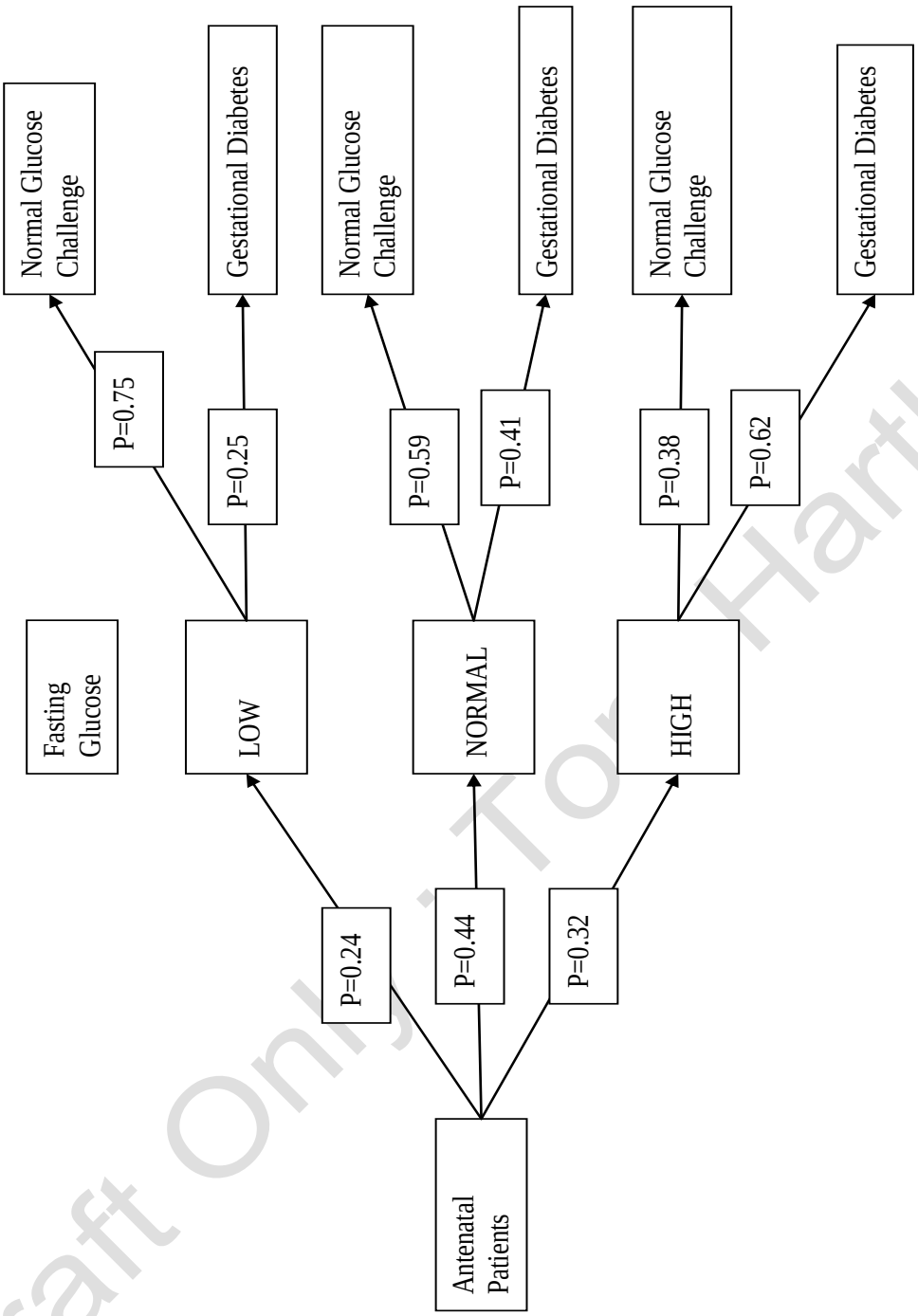
Probability of selling a Second Hand car to Low, Middle and High income customers =

(0.24 * 0.75) + (0.44 * 0.59) + (0.32 * 0.38) = 0.1800 + 0.2596 + 0.1216 = 0.5612

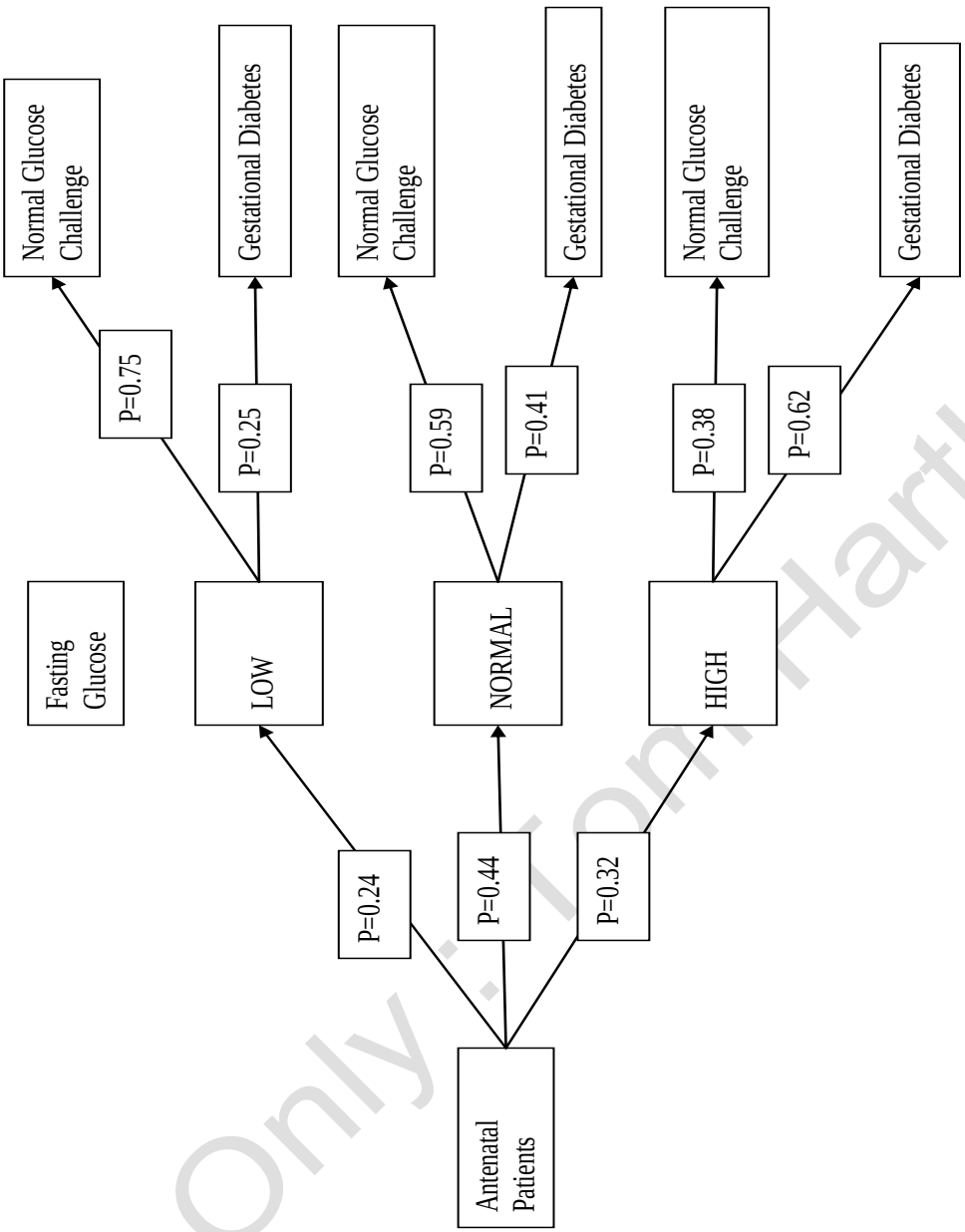
The probability when selling a Second Hand car it is bought by a Middle Income customer = 0.2596 / 0.5612
= 0.4626



If we rename all of these boxes it will look more clinical :



And if we now feed in a hypothetical Antenatal Clinic Size of 1534 patients we can see how many cases of Gestational Diabetes they are going to have to manage : 92 + 277 + 304 = 673. This forecasting property of our analyses is perhaps one of the most useful in these times of critical balances between the size of clinical caseloads and the size of the clinical facilities available to service that caseload.



26.2 DATA MINING BASED UPON BAYES THEOREM

A full explanation of data mining based upon Bayes Theorem is beyond the scope of this text. Instead a demonstration dataset was created with the underlying 'key' pattern:

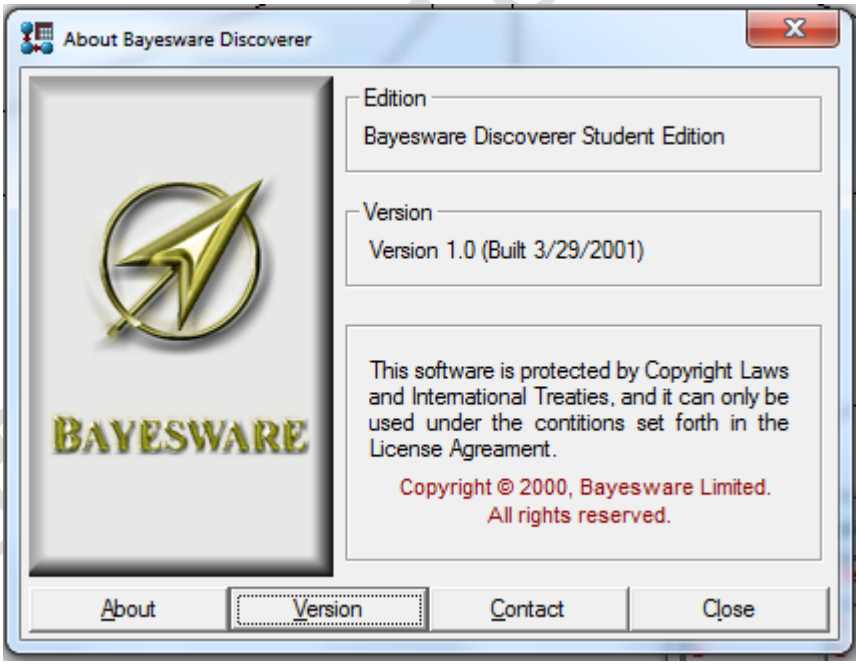
Test 1	Test 2	Test 3	Test 4	Biopsy Diagnosis
L	L	N	N	A
N	L	L	L	B
N	N	L	N	C

Within the dataset were seeded results that did not exactly fit the diagnostic pattern. The dataset consisted of 120 lines structured as shown in Table xxx.

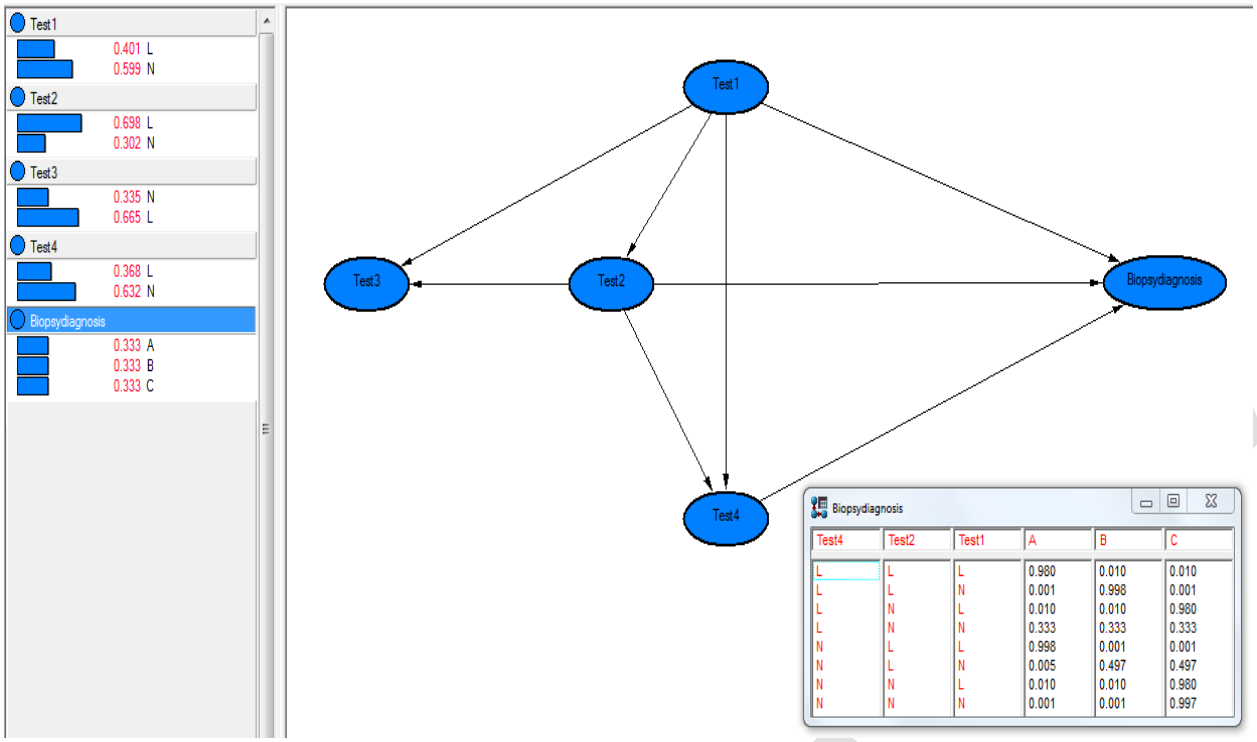
TABLE xxx

Unique Pattern	Count with this pattern	Proportion of Matching Biopsy Diagnosis with this pattern
L L N L A	4	0.1
L L L N A	4	0.1
L L N N A	32	0.8
L N L L C	4	0.1
L N L N C	4	0.1
N L L L B	32	0.8
N L N L B	4	0.1
N L L N B	4	0.1
N L L N C	4	0.1
N N L N C	28	0.7
	120	3

It was fed into Bayesware Discoverer

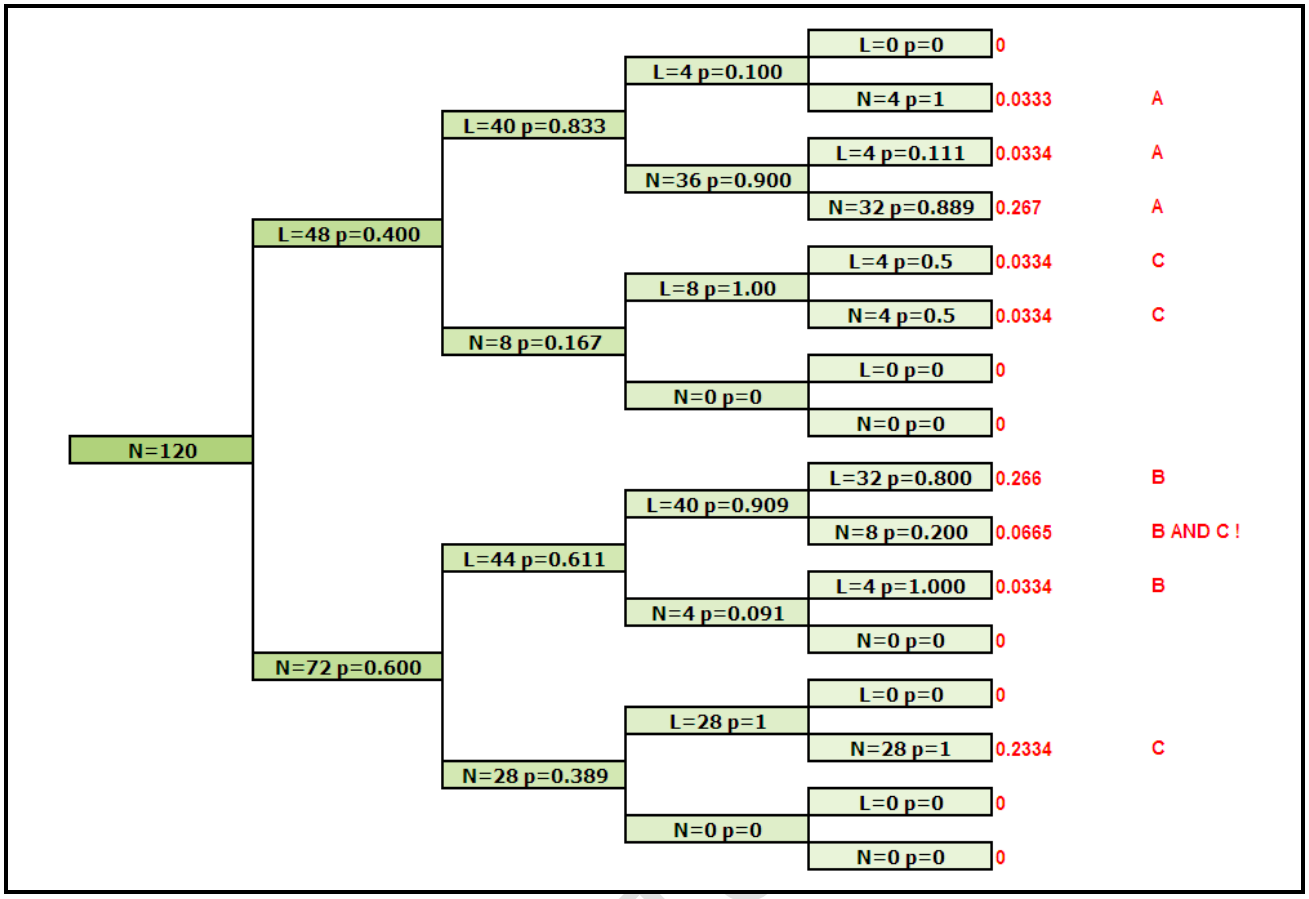


The Bayesian Network produced by this program illustrated that outcomes of Tests 1, 2 and 4 were the most useful determinants of the correct diagnosis ie one that agreed the best with the definitive biopsy diagnosis.



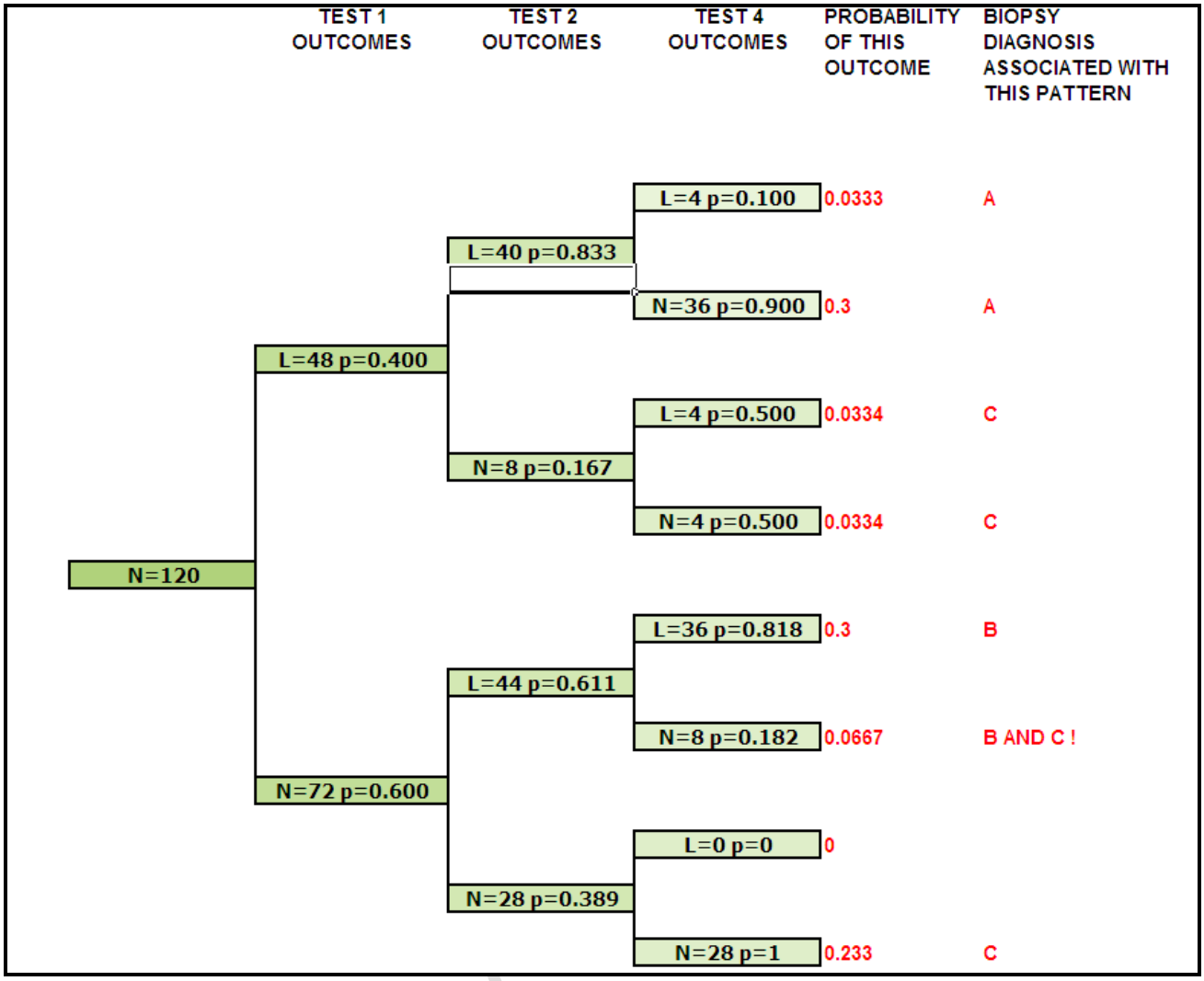
How does Bayesware Discover achieve this ? Firstly it used a Naïve Bayes Classifier and then an iterative search of the Log of the Maximum Likelihood Ratios to build the network. Unfortunately the Help file with the software does not go into the details of these processes. However, if we take a view of these data as a Decision Tree we can appreciate what the software has achieved in a fraction of the time it takes us to build the two Decision Trees necessary to reach a similar conclusion. The first Decision Tree is based upon all the data in Table xxx and is shown in Figure xxxx

FIGURE xxx



On comparing the full decision network with the Bayesware Discoverer network we immediately realise that the software is suggesting that information from Test 3 is not required to get unique patterns for each biopsy diagnosis. In order to compare the decision network approach with the Bayesware approach we need to recalculate the decision network without Test 3. This is shown in Figure xxx.

FIGURE XXX



We can now compare this diagram with the 'truth table' produced by Bayesware in Figure xxxx

FIGURE XXXX

Biopsydiagnosis

Test4	Test2	Test1	A	B	C
L	L	L	0.980	0.010	0.010
L	L	N	0.001	0.998	0.001
L	N	L	0.010	0.010	0.980
L	N	N	0.333	0.333	0.333
N	L	L	0.998	0.001	0.001
N	L	N	0.005	0.497	0.497
N	N	L	0.010	0.010	0.980
N	N	N	0.001	0.001	0.997

Bayesware indicates that the Tests Results combination of LLL has a p=0.980 association with the biopsy diagnosis of A. Our network agrees with this but notice the p value on the network is only 0.0333. This is because our network computes p values across the whole dataset; there were only 4 cases where LLL was associated with a biopsy diagnosis of A and 4/120 = 0.0333. Bayesware is saying that the Tests Results combination of LLL was 98% associated with a biopsy diagnosis of A; (Why did Bayesware not get it equal to 1 ? Probably because of the loss of precision in the Bayesware calculator related to the conversion to and from logarithms.) So the network and the Bayes Network agree on this conclusion.

Bayesware reports that the combination LLN (which on our network is NLL – the order of interpreting the Test Results is not an issue in this experiment) is associated with 99.8% correct identification of patients with biopsy diagnosis B. The decision network has a p value of 0.3 and an association with the biopsy diagnosis of B – again this p value is lower because it is the p value across the whole network; within the group B biopsy diagnosis group it is seen in 36/40 = 0.9. The network is showing that the combination NLL is 100% associated with biopsy diagnosis B and this agrees with Bayesware, within the limits of Bayesware's calculator's precision.

Bayesware reports that NLL detects biopsy diagnosis of A with 99.8% accuracy - which in our network is LLN and which also corresponds to biopsy diagnosis of A in 36 patients.

Bayesware reports that NNL biopsy diagnosis of C with 98% accuracy - which in our network is LNN and which also corresponds to biopsy diagnosis of C in 4 patients.

Bayesware reports that NNN biopsy diagnosis of C with 98% accuracy - which is the same in our network and which also corresponds to biopsy diagnosis of C in 28 patients.

Bayesware also picks up on the difficulty of the pattern NLN being distributed 50:50 between the biopsy diagnoses of B and C, Bayesware reports 49.7:49.7 again probably because of the precision limitations of their calculations.

Finally we need to look at the seemingly anomalous row in Bayesware for their pattern LNN which in our network corresponds to NNL. Bayesware reports that biopsy diagnoses of A,B or C could be associated with this with 33.3 5 accuracy, in other words a 1 in 3 chance of the right diagnosis – so not a diagnostic pattern. The decision network diagram for NNL 'crashes' or more exactly terminates after NN and the row has a p=0 which also means that the pattern is not diagnostic.

So in summary Bayesware Discoverer provides a fast tool for determining dependencies in a data set. It's speed makes it much more attractive than going through the processes of careful sorting of your data using Excel and then working out a variety of network diagrams that include and exclude one or a combinations of the columns of data.

The example worked through here was on a simple set of binary outcomes – test result was normal or low. Bayesware discovery is best suited to categorical data – so it would have worked equally well with "test result low, normal, or high". When it comes to continuous data ie if we had put in numerical values for the results of Tests 1, 2, 3 and 4, it would have quantised them into quartiles before doing the analysis. If you have only a small dataset such as this one where we had only 120 lines this would have given you many more possible combinations to test – 4 tests with 4 possible outcomes gives you 256

permutations in comparison to what we have just described where 4 tests each with 2 possible outcomes gave us the possibility of 16 outcomes. In reality the system under investigation only went out to 9. The alternative to using Bayesware's internal 'quartile' quantiser is to do the quantisation first yourself particularly if you have no low values in the system you are testing; you could have three quantiles - normal, high and extreme.

A recent article that employed Bayes Network Analysis in a healthcare scenario can be viewed at : http://www.e-epih.org/DOIx.php?id=10.4178/epih/e2011006

27: THE STATISTICAL DESIGN OF RESEARCH PROJECTS:

27.1: SAMPLE SIZE REQUIRED FOR A STUDY TO REACH A SPECIFIED POWER

The three cornerstones of contemporary design of medical research projects are:

- The CONSORT statement
- The STARD statement
- The estimation of the 'Power of the Study'

CONSORT is the acronym for Consolidated Standards of Reporting Trials and is the 'pro-forma' for the design of and data collection from 'randomised controlled trials'. These are trials that test the effectiveness of a clinical intervention in a defined group of patients or a defined combined group of patients and normal controls. An article that describes the 2010 version of CONSORT is in Appendix 2 to these notes.

STARD is the acronym for Standards for Reporting of Diagnostic Accuracy and is the 'pro forma' for the design and data collection of trials where the accuracy of new diagnostic investigation e.g. a laboratory test, is compared to that of either 'current usual practice' or an 'index diagnostic investigation'. By 'index' is meant the best available diagnostic investigation for the clinical condition being considered. A copy of the STARD statement is in Appendix 3.

27.2: POWER OF A STUDY

Regardless of whether a medical research project falls under the umbrella of CONSORT or STARD, all studies at the outset have to define what the 'Power of the Study' is inherent within the intended size of the study. In general terms, if the expected differences in the critical measures are anticipated to be small, then the investigators will have to recruit a large number, 'N', of participants. Conversely if the expected differences in the critical measures are expected to be large then usually only a relatively small group of participants will need to be recruited. A spin off of carrying a 'Power of the Study' estimate is that it focuses the investigators on what type of statistical tests they are going to use on their expected dataset before they start collecting data. This reduces the risk that at the end of the study the investigators suddenly realise that they have collected the wrong kind of data for the statistical tests they were planning to use and even worse, have collected an inadequate amount of data to detect the differences they were interested in. They then resort to applying a whole range of statistical tests on what becomes a 'fishing trip' for statistically significant differences in their dataset.

The Power of a Study that is anticipated to involve two data sets that are normally distributed is straightforward to calculate; it essentially involves rearranging the formula for a Student's t Test to solve for 'N' rather than 't', (see Figures 15 and 16).

Sample size for paired data:

$$n = \frac{\sigma_d{}^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

where :

$n = \text{sample size}$

$\sigma = \text{standard deviation of the within - pair difference}$

$\text{diffference} = \text{clinically meaningful difference}$

$Z_\beta = \text{corresponds to power (.84 = 80\% power)}$

$Z_{\alpha/2} = \text{corresponds to two - tailed significance level (1.96 for } \alpha = .05)$

Sample size for unpaired data:

$$n_1 = \frac{(r+1)}{r} \frac{\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

where :

$n_1 = \text{size of smaller group}$

$r = \text{ratio of larger group to smaller group}$

$\sigma = \text{standard deviation of the characteristic}$

$\text{diffference} = \text{clinically meaningful difference in means of the outcome}$

$Z_\beta = \text{corresponds to power (.84 = 80\% power)}$

$Z_{\alpha/2} = \text{corresponds to two - tailed significance level (1.96 for } \alpha = .05)$

Fortunately there are calculators available for solving these and other experimental designs. I recommend :

PS : Power and Sample Size Calculator

which can be downloaded from

The Power of a Study that is anticipated to involve two data sets that are not normally distributed is not as straightforward. In these notes I have encouraged the use of the Mann Whitney U Test for such datasets. Without knowing the likely distribution of your datasets under this scenario it is difficult to obtain an objective estimate of 'N'. There are sophisticated data simulation approaches to this problem which are beyond the scope of this discussion. Nevertheless you can probably make a reasonable estimate the distribution of your 'untreated' subjects from prior experience or reports. You could then increment these by an expected amount to simulate the treatment effect and then calculate 'U'. Then by successively removing data points at random from each dataset and recalculating 'U' you could arrive at a point where 'N' best matches the practicalities of recruiting sufficient cases to your desired p =< 0.05. If you are unsure of what shape your data distribution is going to be then you can use the formulae for the estimation of the errors above and below your expected median; (these are taken from the book by Snedecor).

$$Upper\ Limit = \frac{N+1}{2} + \sqrt{N}$$

$$Lower\ Limit = \frac{N+1}{2} - \sqrt{N}$$

You round down the lower limit and round up the upper limit. These two equations actually give you the item number of the data point that corresponds to the limit – so to use them you need to have your data sorted into ascending order. Snedecor gives the example of blood pressures measured in 11 men.

110, 112, 116, 120, 126, 132, 142, 147, 154, 156, 179.

So the lower limit is

$$\frac{11+1}{2} - \sqrt{11} = 6 - 3.3 = 2.7 \approx 2nd\ item$$

which is 112.

The upper limit is

$$\frac{11+1}{2} + \sqrt{11} = 6 + 3.3 = 9.3 \approx 10th\ item$$

which is 156.

I have used these formulae and random data generator features in Excel to calculate the limits on a simulated dataset of Vitamin D concentrations which I wanted to go from 10 to 100, (see Figure 17). On the day the simulator actually gave me a set that went

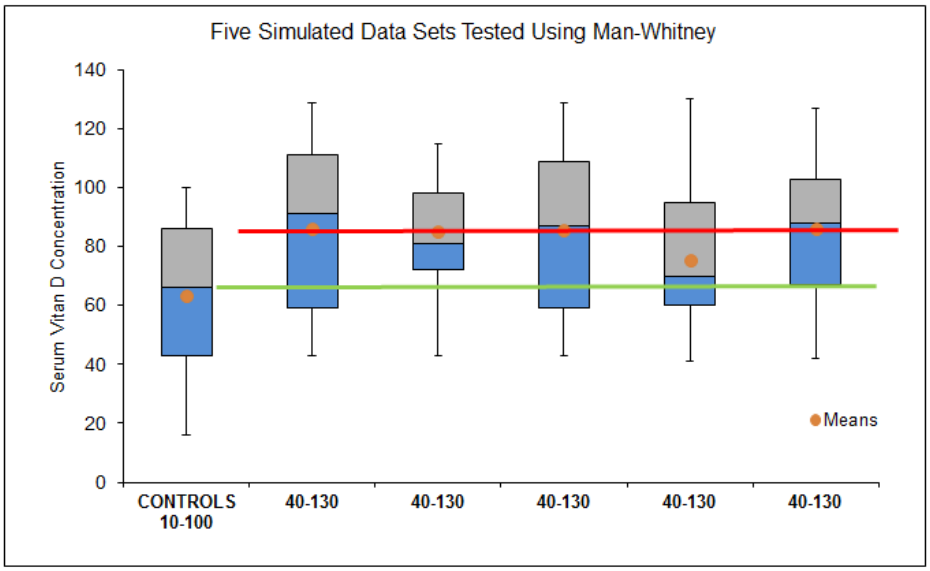
from 16 to 100 and had a median of 66; see the column with the grey background. (I had expected the median to have been closer to the theoretical one of 55 but in simulations you must adhere to what the simulator gives you and not go 'hunting' for a dataset that you prefer !) The lower and upper limits around this median occur at the 9th and 20th items respectively which are concentrations of 43 and 85. The scenario is that I want to give these people a Vitamin D supplement … What is the minimum difference must I achieve between the before and after medians to be 95% confident that I have caused an improvement ? The answer is that the post supplement median must be equal to or greater than the upper limit of the control median, which is 85. I then used the simulator to produce five random datasets within the range 40 to 130 because this has a theoretical median of 85. The Box and Whisker Plot (Figure 18) shows those five datasets, the green line is at the Control Dataset Median of 66 and the red line is at the target median of 85. As you can see only three of the datasets are above the target. Nevertheless when you put the Control Dataset up against the five simulated datasets and analyze them using the Mann-Whitney U Test then four of them, (1st, 2nd, 3rd and 5th) come out with highly significant 'z' values; a 'z' value of 1.96 or greater means that the two datasets are significantly different at the p=<0.05 level.

So in summary to achieve a significant result in our study we would need to have 29 subjects in the group to have an 80% chance of detecting a difference with a significance level of p=<0.05.

FIGURE 17

	CONTROLS 10-100	40-130	40-130	40-130	40-130	40-130
	16	43	43	43	41	42
	19	44	51	44	46	43
	23	45	55	48	50	45
	34	49	57	54	52	51
	35	50	70	55	55	53
	42	52	71	56	58	59
	42	56	71	57	58	66
	43	59	72	59	60	67
	43	67	73	63	62	76
	52	73	74	69	62	76
	54	73	78	70	62	77
	56	77	79	70	63	77
	57	84	80	72	63	78
	61	86	80	86	65	81
	66	91	81	87	70	88
	68	92	86	90	71	91
	69	94	89	94	72	95
	72	97	93	95	75	95
	82	101	95	102	76	97
	85	108	97	106	82	98
	85	110	97	108	89	99
	86	111	98	109	95	103
	86	111	107	111	96	111
	89	114	109	114	99	112
	92	118	110	119	102	117
	94	118	112	122	103	123
	94	119	113	124	107	124
	99	129	115	128	117	127
	100	129	115	129	130	127
Median	66	91	81	87	70	88
Mean	63.58621	86.2069	85.2069	85.6552	75.2069	86.1379
U		232.5	224.5	233.5	318.5	234
z		-2.92	-3.05	-2.91	-1.59	-2.9
Lower L #	9	9	9	9	9	9
Upper L #	20	20	20	20	20	20
Median LL	43	67	73	63	62	76
Median UL	85	108	97	106	82	98
MIN	16	43	43	43	41	42
MAX	100	129	115	129	130	127

FIGURE 18



	CONTROLS 10-100	40-130	40-130	40-130	40-130	40-130
Max	100	129	115	129	130	127
75th Percentile	86	111	98	109	95	103
Median	66	91	81	87	70	88
25th Percentile	43	59	72	59	60	67
Min	16	43	43	43	41	42
Mean	64	86	85	86	75	86

27.3: RESOLVING THE PROBLEM OF APPROPRIATE PLACEBO/CONTROL GROUP DESIGN

28: APPLIED STATISTICS

28.1: THE SHEWHART OR LEVEY JENNINGS QC PLOT

Whenever a device is used for making a measurement in healthcare it should be checked that it is in fact reading correctly. In the pathology laboratory this is readily achieved by including at least two but usually three quality control specimens. These are not 'real' specimens but aliquots taken from a large pool of say blood serum about which we know its biochemical composition. For example suppose we know from expert analysis of the pool that it contains 5.5 mmol/L of cholesterol. Now when an aliquot of that pool is used today then it becomes a quality control specimen in today's batch of patient samples and I would expect it to return a value close to 5.5 mmol/L. How close ? Well within +/- 2SDs on 95% occasions that I analyze it. That way I can be assured that my cholesterol analyses of the patients sera in the same batch are very probably being analyzed to the same level of accuracy (closeness its true value) and precision (margin of error) as was obtained by the expert laboratory. If I chart my results for the QC material against the time and date on which I performed the analysis and I also indicate on that chart the position of the target mean and the positions of the +/- 2SDs limits about than mean then I have constructed a Shewhart Process Control chart; so named after Walter A Shewhart who published the idea in 1924. In 1950 the concept was latched onto by Stanley Levey and Elmer Jennings who are credited with promoting the concept amongst chemical pathology laboratories. In

1981 James Westgard published his multi-rule Shewhart approach to statistical quality control in the clinical laboratory. An implementation of this is available on the website :

www.medlabstats.com/alliedhealth

It has been realized in Excel and flags events where the current measurement of and analyte in a quality control material

(a) falls outside the target mean plus or minus two SDs; the value is highlighted with an orange background eg. Cell B7 in Figure xxx
(b) it contravenes Westgard's 'delta is equal to or greater than four times the SD' rule; this is indicated by red asterisks appearing in that column of the spreadsheet eg. Cell E8 in Figure xxx
(c) It contravenes Westgard's 'run of four results all on the same side of the mean' rule; eg Cell F16 in Figure xxx

The probabilities of these rules being broken are as follows
(a) p = 0.028 or 2.28%,
(b) p = 0.028 * 0.028 = 0.00052 or 0.05%
and (c) 0.3413 * 0.3413 * 0.3413 * 0.3413 = 0.0136 or 1.36%

Westgard has proposed other rules but these three a probably the most widely applied on a to day to day basis in most pathology laboratories.

So if you practice involves the use of point of care biochemical and haematology analyzers then you should perform similar three point QC samples each day you use them and have levy-Jennings plots for all of them..

What can users of other clinical measuring instruments do to quality control their use. For example how can weighing scales, tape measures, blood pressure meters, goniometers, dynamometers, skinfold thickness calipers etc be quality controlled.

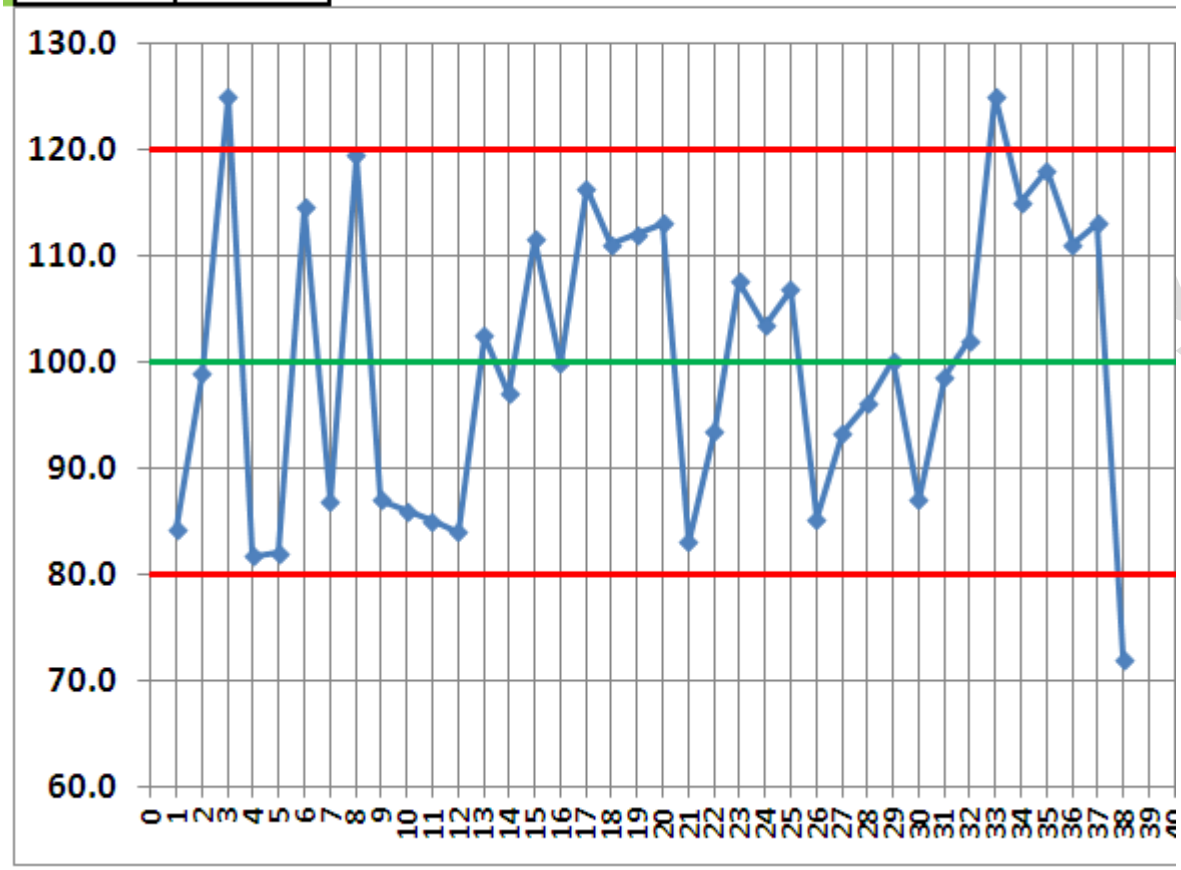
Weighing scales are best checked periodically by putting on standard weights once a month and checking that the scale is reading the specified value. The standard weights should be at the low and high end of what those scales are normally used for. So for example if adults are being weighed on them then standard weights of say 50kg and 120kg should be available for making monthly checks. Results should be recorded on a form; there is no need to plot a Levy-Jennings style chart. Calculating the running means and SDs would be advisable because then you can also monitor whether or not a set of scales is reading significantly lower or higher (using the simple paired Student's t Test) than the others in your clinical examination rooms.

Blood pressure meters are more of a problem. If in doubt about a unit then it should be referred back to the manufacturer for testing. However, it is not uncommon for practitioners to notice that say in a clinic with three consulting rooms each with its own blood pressure meter that when they consult in consulting room one the patient's blood pressures tend to read 'lower'. How could you ascertain if this was happening. Well you can borrow a techinique from laboratory haematology called Bull's Mean. It requires that in each consulting room the clinicians always record the systolic and diastolic pressures on a record chart. The theory around Bull's Mean is that the running means of the last twenty patients should not vary by more than +/- 3%. A computer simulation has also brought up a second rule : that the average of three consecutive Bull's means should not vary by more

than +/- 2%. To chart a Bull's Mean involves very similar process as for the Levy-Jennings plot.

(The multirule quality control procedure for interpreting Bull's moving averages, proposed by Levy and colleagues (Am J Clin Pathol 1986; 85: 719-721.), has been evaluated by computer simulation with the use of the approach of Cembrowski and Westgard (Am J Clin Pathol 1985; 83: 337-345.). With this procedure, a batch of patient specimens is rejected if either of two criteria are satisfied: (1) the Bull's mean of one of the red blood cell indices is outside its 3% limits, or (2) the average of three consecutive Bull's means is outside its 2% limits. Power function curves were used to summarize the performance of the multirule approach and demonstrated error-detection capabilities that are superior to the more common implementation of Bull's algorithm using 3% limits for single Bull's means. The increased error detection achieved by the multirule procedure allows shifts in hemoglobin and mean corpuscular volume to be more readily detected but also results in the detection of small shifts in red blood cell count. A modified multirule procedure was also tested and was found to be ineffective. The authors recommend the multirule of Levy and colleagues but caution that its use may result in the detection of small shifts in the red blood cell count.)

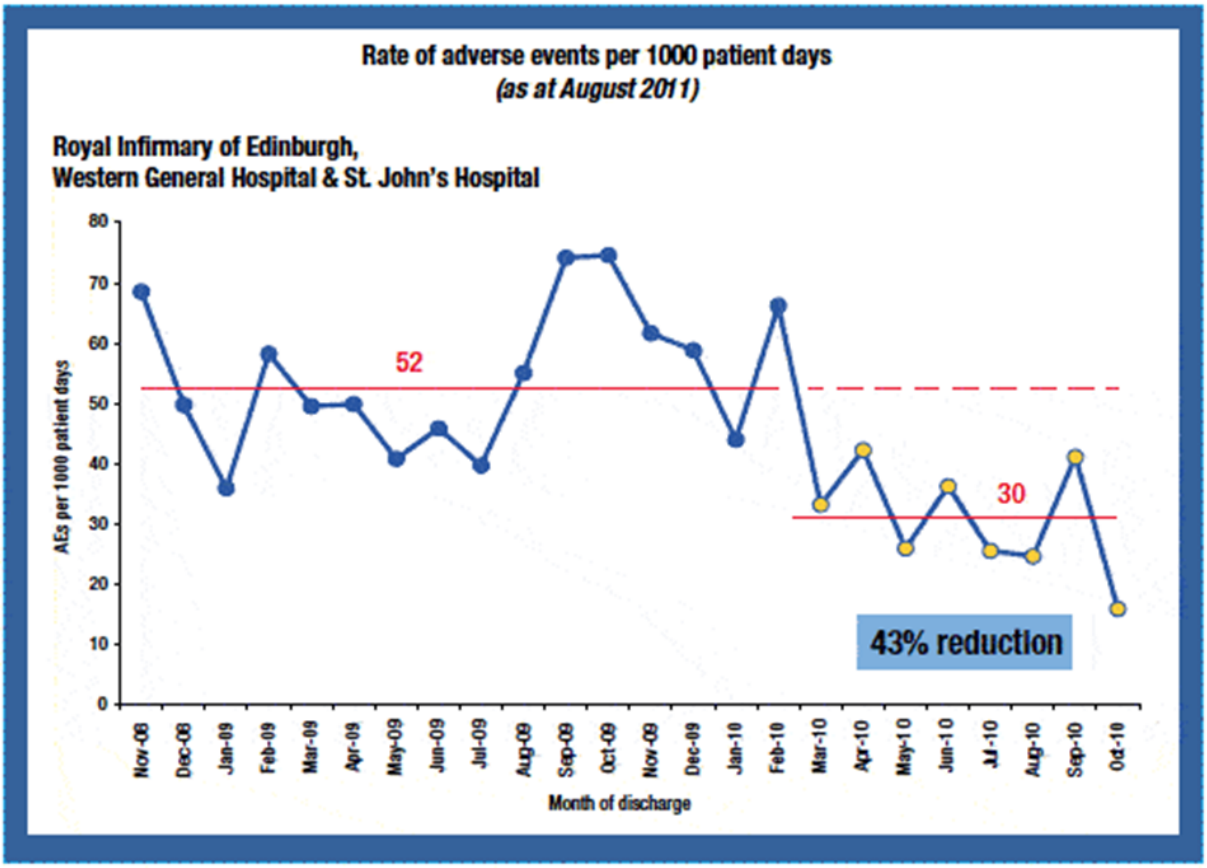
TARGET UCL	TARGET LCL
120	80



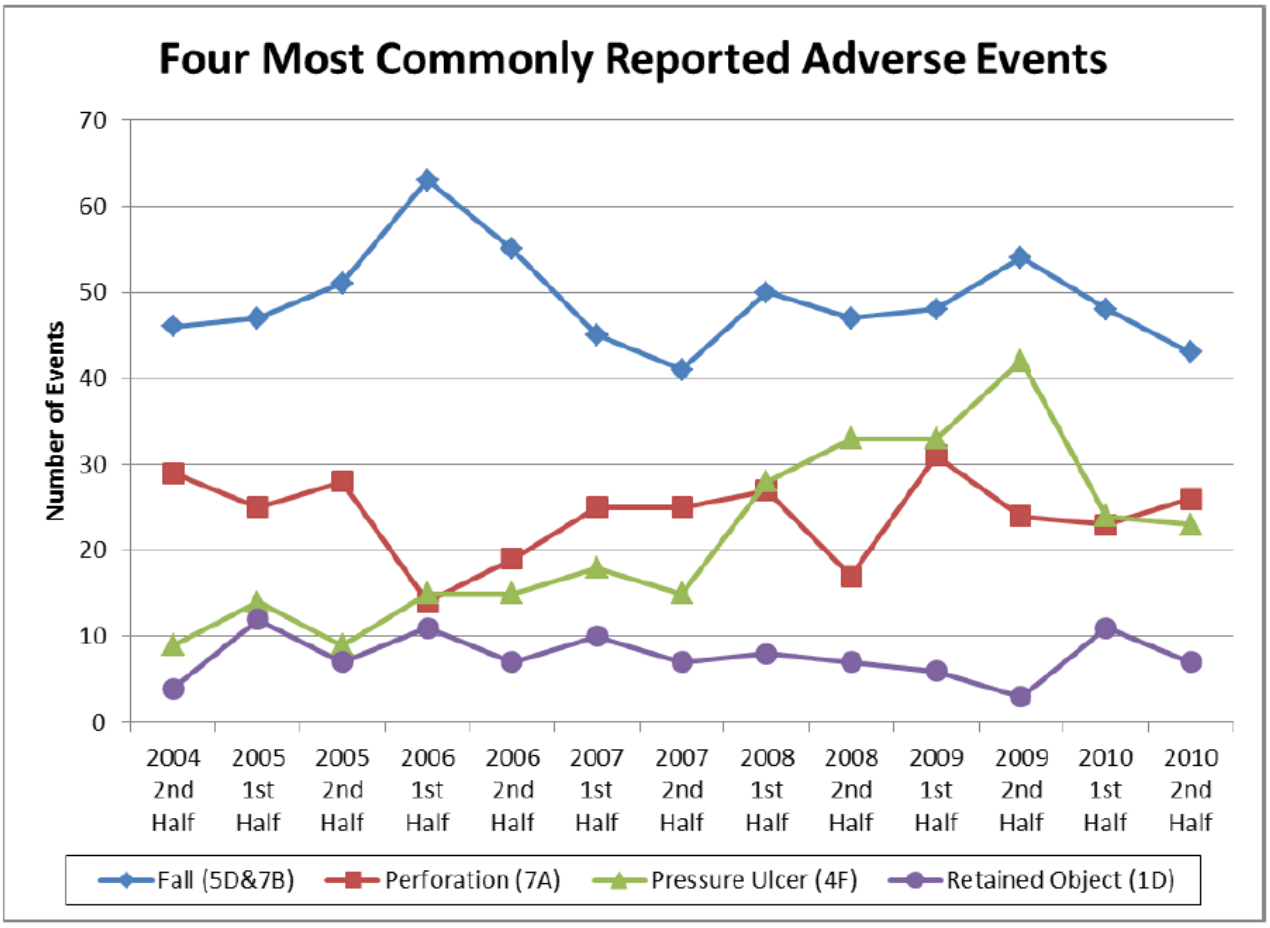
	A	B	C	D	E	F	G
1							
2					TARGET MEAN	TARGET SD	
3	Batch Number	QC RESULT	RUNNING MEAN	RUNNING SD	100	10	
4					Delta 4 SD Rule Broken	Run of 4 > 1SD or 4 < 1SD	SNDs
5	1	84.3					-1.57
6	2	98.9	91.6	10.324			-0.11
7	3	125	102.7	20.619			2.50
8	4	81.7	97.5	19.850	**********	-	-1.83
9	5	82	94.4	18.532		-	-1.80
10	6	114.5	97.7	18.499		-	1.45
11	7	86.8	96.2	17.385		-	-1.32
12	8	119.4	99.1	18.070		-	1.94
13	9	87	97.7	17.375		-	-1.30
14	10	86	96.6	16.797		-	-1.40
15	11	85	95.5	16.311		-	-1.50
16	12	84	94.6	15.903		4 < MINUS 1SD	-1.60
17	13	102.5	95.2	15.385		-	0.25
18	14	97	95.3	14.790		-	-0.30
19	15	111.5	96.4	14.853		-	1.15
20	16	99.8	96.6	14.375		-	-0.02
21	17	116.3	97.7	14.717		-	1.63
22	18	111	98.5	14.615		-	1.10
23	19	112	99.2	14.538		-	1.20
24	20	113	99.9	14.483		4 > PLUS 1SD	1.30
25	21	83.1	99.1	14.584		-	-1.69
26	22	93.5	98.8	14.282		-	-0.65

Control charts, Cusum techniques and funnel plots. A review of methods for monitoring performance in healthcare : Luc Noyez : *Interact CardioVasc Thorac Surg (2009) 9 (3): 494-499.*

28.2 TRIGG TRACKING SIGNAL



NHS Scotland Chief Executive's Annual Report 2010/11
http://www.scotland.gov.uk/Publications/2011/11/10140644/4



http://www.ct.gov/dph/lib/dph/hisr/hcqsar/healthcare/pdf/adverseeventreport2011.pdf

28.3 YOUDEN PLOTS

Youden plots

1. The original Youden plot (MedCalc)

For the original Youden plot (Youden, 1959) (see Figure 1) the two samples must be similar and reasonably close in the magnitude of the property evaluated.

The axes in this plot are drawn on the same scale: one unit on the x-axis has the same length as one unit on the y-axis.

Each point in the plot corresponds to the results of one laboratory and is defined by a first response variable on the horizontal axis (i.e. run 1 or product 1 response value) and a second response variable 2 (i.e., run 2 or product 2 response value) on the vertical axis.

A horizontal median line is drawn parallel to the x-axis so that there are as many points above the line as there are below it. A second median line is drawn parallel to the y-axis so that there are as many points on

the left as there are on the right of this line. Outliers are not used in determining the position of the median lines. The intersection of the two median lines is called the Manhattan median.

A circle is drawn that should include 95 % of the laboratories if individual constant errors could be eliminated.

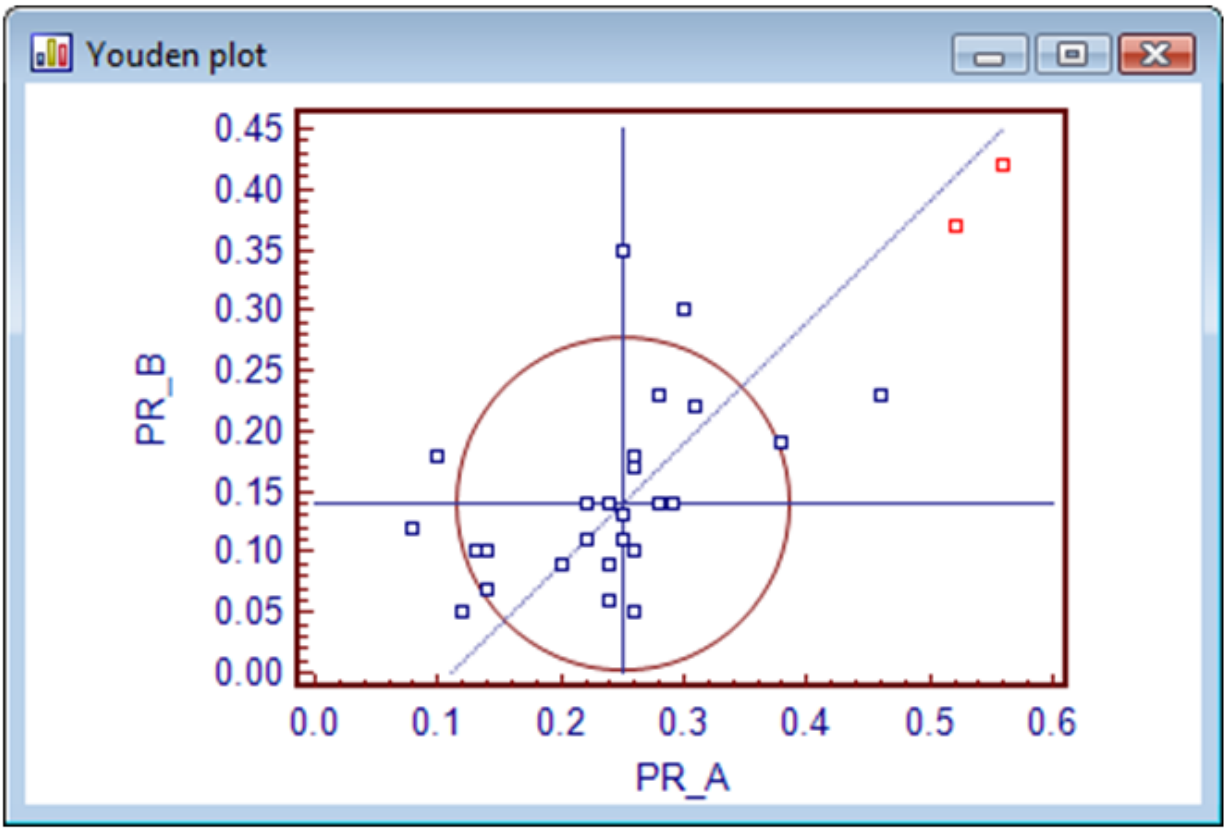
A 45-degree reference line is drawn through the Manhattan median.

Interpretation

Points that lie near the 45-degree reference line but far from the Manhattan median, indicate large systematic error.

Points that lie far from the 45-degree line indicate large random error.

Points outside the circle indicate large total error.



28.3 THE BLAND AND ALTMAN PLOT FOR COMPARING TWO METHODS OF MAKING THE SAME MEASUREMENT

There is a natural – and statistically incorrect tendency – that when a scientist is presented with two sets of data measured on the same samples but by two different techniques, to arbitrarily decide that say the measurements by technique ‘A’ go on the x axis and those measured by technique ‘B’ go onto the y axis. Bland and Altman pointed that this was

statistically incorrect out in a paper in the Lancet in 1986. They said that if you want to be completely unbiased in your assessment of the agreement between two sets of measurements made on the same samples by two different techniques then you must :

Plot the Means of the Pairs of Measurements on the 'x' axis
and
Plot the Differences between the Pairs of Measurements on the 'y' axis

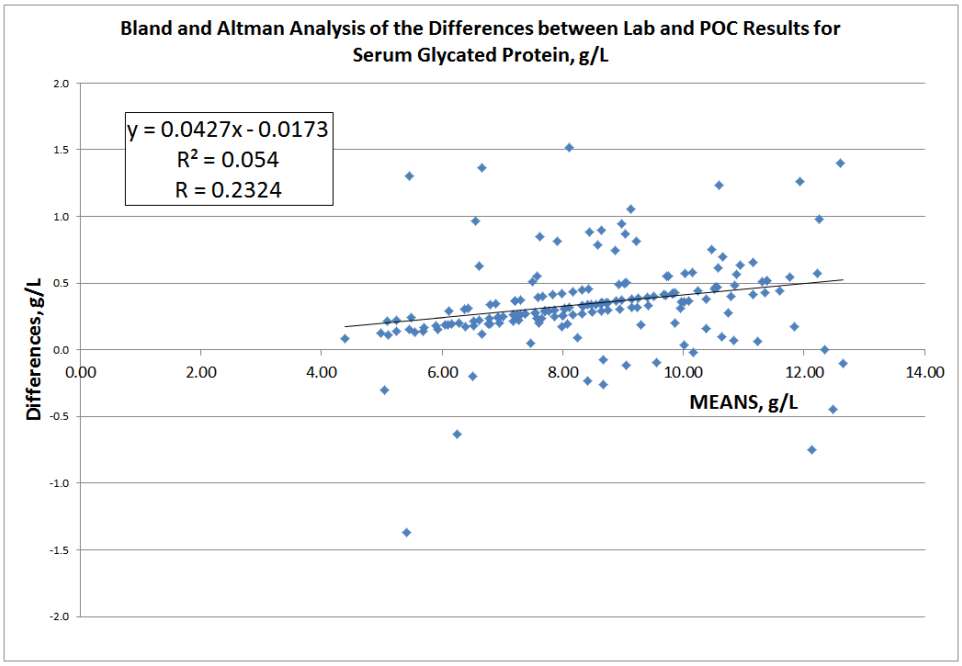
Why ? Because clearly the measurement of say cholesterol in the sample by technique 'A' can in no way determine what technique 'B' is going to get for the cholesterol concentration in the same sample. There is no cause and effect relationship here. However by looking at the means and differences then an independent – dependent relationship is created and that is statistically acceptable.

Once constructed the Bland and Altman plot can then be used to estimated the Systematic Bias and the Fixed Bias between the two techniques. The actual values for these are obtained from least squares linear regression analysis such that :

Systematic Bias = the slope of the line
and the
Fixed Bias = the intercept of the line on the 'y' axis.

The technique for drawing the graph and getting the equation to the line of best fit using Excel Figure 20 will be demonstrated in the lecture.

FIGURE 20



Because the x,y scatter plot drawing tool in Excel gives only the minimum information about the line of best fit we also have to use the Regression Data Analysis tool. This will be demonstrated in the lecture.

The output from the Regression Data Analysis tool (see Figure 21) shows that there is a significant relationship between the Means and the Differences because the Regression

Coefficient of 0.234 is significant at the p = 0.002 level of significance. It also shows us that only the slope of the line, 0.0427, is significant at the p = 0.002 level of significance.

FIGURE 21

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.2324
R Square	0.0540
Adjusted R Square	0.0486
Standard Error	0.3346
Observations	178

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.124917089	1.124917089	10.0492022	0.002
Residual	176	19.70160457	0.111940935		
Total	177	20.82652166			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.0173	0.117165275	-0.14749308	0.883	-0.248510765	0.21394863	-0.24851077	0.213948632
MEANS	0.0427	0.013484294	3.170047603	0.002	0.016134137	0.06935757	0.016134137	0.069357574

The overall conclusion we can make about this comparison of two analytical methods of measuring glycated proteins are :

(i) Because the SLOPE of the Bland and Altman plot IS SIGNIFICANT then there is A SIGNIFICANT SYSTEMATIC BIAS between the two methods of about 4.3%

(ii) Because the INTERCEPT of the Bland and Altman plot is NOT SIGNIFICANT then there is NO SIGNIFICANT FIXED BIAS between the two methods.

28.4 CURVE FITTENG USING CUBIC SPLINES

There are situations where it is not possible to fit data on a graph to a mathematical model. Nevertheless it is still desirable to do because with a 'best fit' curve through data it is possible to use that graph to read off values that should occur between values that were not included in the original study. In the author's own work cubic splines have been used to fit smooth curves to radioimmunoassay calibration curves.

The cubic spline is the mathematical analogue of the engineer's spline, which consists of a long narrow flexible strip of wood or plastic which can be made to pass over or under a series of fixed pegs by bending it and/or loading it with strategically placed lead weights. This process can be expressed as a cubic of the form :

$$y_i = A_i + B_i x_i + C_i x_i^2 + D_i x_i^3$$

where x_i and y_i are the values at the *i*th peg.

For intermediate points between the pegs the equation takes the form

$$y = A_i + B_i (x_j - x_i) + C_i (x_j - x_i)^2 + D_i (x_j - x_i)^3$$

where $x_i < x_j < x_{i+1}$

In other words, the values of the coefficients A, B, C and D are entirely local and apply only along the interval of the curve between the *i*th and the *(i + 1)*th pegs. The cubic spline procedure is essentially a minimisation procedure aimed at obtaining the values of those

coefficients which give a smooth curve over each segment x_i to x_{i+1} but with the added proviso that these must give a smooth progression from and to the curvatures in both the previous segment x_{i-1} to x_i and the following segment x_{i+1} to x_{i+2} respectively. This is achieved mathematically by ensuring that the first and second derivatives of the equation that applies at each x,y data point or knot (as it is referred to in spline fitting) are both equal. A program for cubic spline fitting of data is available on the website :

www.medlabstats.com/alliedhealth

The program produces two statistics – the standard error of the estimate of y from x and a value for the *Reduced Chi Squared.* The latter is calculated using the equation

See page 32 of 1st Edition QC book.

A good fit is indicated by a *Reduced Chi Squared* equal to or less than 3.84. A perfect fit is indicated by a *Reduced Chi Squared* equal to one. You should note that if a smoothing factor of zero is used then the *Reduced Chi Squared* will always be zero.

Nakagawa, Iwao, Ishida, et al

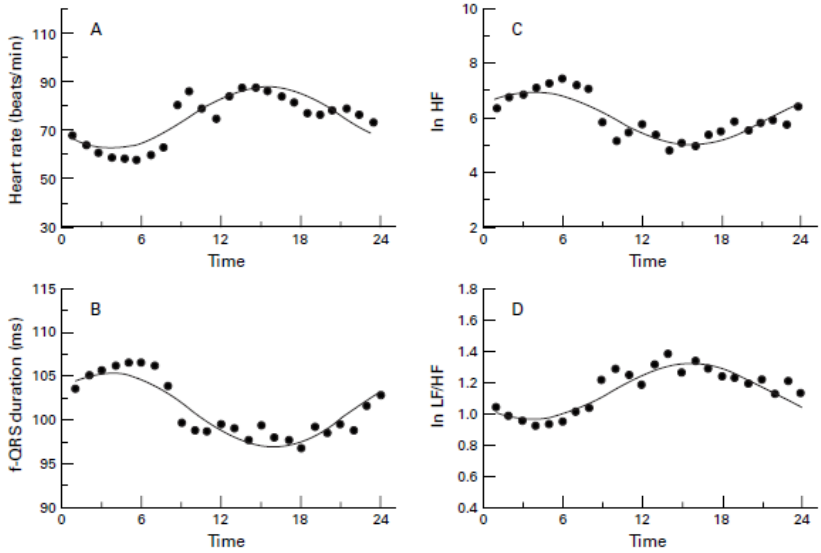


Figure 1 Circadian rhythm of the hourly means of heart rate (A), f-QRS duration (B), ln HF (C), and ln LF/HF (D) in 20 subjects. Solid lines represent curves fitted to the data by the single cosinor method. LF and HF, low and high frequency components, respectively, of heart rate variability; LF/HF, LF to HF ratio.

To be completed

APPENDIX ONE : CRITICAL VALUES OF CHI SQUARES

Degrees of Freedom	P Value		
	0.10	**0.05**	**0.01**
1	2.71	3.84	6.64
2	4.60	5.99	9.21
3	6.25	7.82	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.73
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.80
19	27.20	30.14	36.19
20	28.41	31.41	37.57

APPENDIX TWO : WILCOXON RANK SUMS

NUMBER OF SUBJECTS IN GROUP WITH MOST MEMBERS	NUMBER OF SUBJECTS IN GROUP WITH FEWEST MEMBERS														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	36-52	45-63	55-75	66-88	79-101	92-116	106-132	121-149
3	-	-	-	-	15-30	22-38	29-48	38-58	47-70	58-82	69-96	82-110	95-126	110-142	125-160
4	-	-	-	10-26	16-34	23-43	31-53	40-64	49-77	60-90	72-104	85-119	99-135	114-152	130-170
5	-	-	6-21	11-29	17-38	24-48	33-58	42-70	52-83	63-97	75-112	89-127	103-144	118-162	134-181
6	-	-	7-23	12-32	18-42	26-52	34-64	44-76	55-89	64-104	79-119	92-136	107-153	122-172	139-191
7	-	-	7-26	13-35	20-45	27-57	36-69	46-82	57-96	69-111	82-127	96-144	111-162	127-181	144-201
8	-	3-19	8-28	14-38	21-49	29-61	38-74	49-87	60-102	72-118	85-135	100-152	115-171	131-191	149-211
9	-	3-21	8-31	14-42	22-53	31-65	40-79	51-93	62-109	75-125	89-142	104-160	119-180	136-200	154-221
10	-	3-23	9-33	15-45	23-57	32-70	42-84	53-99	65-115	78-132	92-150	107-169	124-188	141-209	159-231
11	-	3-25	9-36	16-48	24-61	34-74	44-89	55-105	68-121	81-139	96-157	111-177	128-197	145-219	164-241
12	-	4-26	10-38	17-51	26-64	35-79	46-94	58-110	71-127	84-146	99-165	115-185	132-206	150-228	169-251
13	-	4-28	10-41	18-54	27-68	37-83	48-99	60-116	73-134	88-152	103-172	119-193	136-215	155-237	174-261
14	-	4-30	11-43	19-57	28-72	38-88	50-104	62-122	76-140	91-159	106-180	123-201	141-223	160-246	179-271
15	-	4-32	11-46	20-60	29-76	40-92	52-109	65-127	79-146	94-166	110-187	127-209	145-232	164-256	184-281

APPENDIX THREE : TABLE FOR THE MANN-WITNEY U TEST

Critical Values for the Wilcoxon/Mann-Whitney Test (U)

Nondirectional α=0.05 (Directional α=0.025)

n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	0	0	0	0	1	1	1	1	1	2	2	2	2
3	-	-	-	-	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	-	-	-	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	-	-	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	-	-	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	-	-	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	-	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	-	0	2	4	7	10	12	15	17	21	23	26	28	31	34	37	39	42	45	48
10	-	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	-	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	-	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	-	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	-	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	-	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	-	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	-	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	-	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	-	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	-	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Nondirectional α=0.01 (Directional α=0.005)

n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
3	-	-	-	-	-	-	-	-	0	0	0	1	1	1	2	2	2	2	3	3
4	-	-	-			0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	-	-	-		0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	-	-	-	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	-	-	-	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	-	-	-	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	-	-	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	-	-	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	-	-	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	46
12	-	-	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	-	-	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	-	-	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	-	-	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	-	-	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	-	-	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	-	-	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	-	0	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	-	0	3	8	13	18	24	30	36	42	46	54	60	67	73	79	86	92	99	105

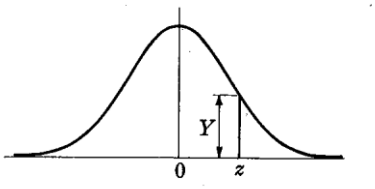
U_{test} is the lesser of the two test statistics (U_1 & U_2) that you have just calculated.

If your $U_{test} \leq U_{crit}$, in the above Table then reject the 'null hypothesis', (H_0).

Dashes (-) indicate that the sample size is too small to reject the Null Hypothesis at the chosen α level of significance.

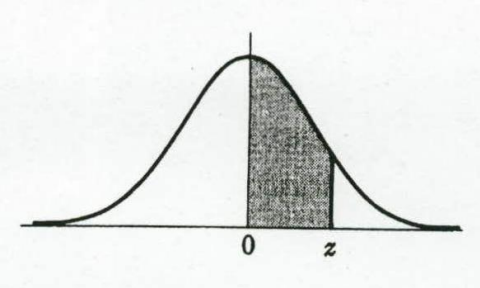
APPENDIX FOUR : ORDINATES (y values or Frequencies) OF THE STANDARD NORMAL CURVE

ORDINATES (Y) of the STANDARD NORMAL CURVE at z



	Second Decimal Place in Z			Second Decimal Place in Z	
Z	0.00	0.05	**Z**	0.00	0.05
0.0	0.3989	0.3984	2.0	0.0540	0.0488
0.2	0.3910	0.3867	2.2	0.0355	0.0317
0.4	0.3683	0.3605	2.4	0.0224	0.0198
0.6	0.3332	0.3230	2.6	0.0136	0.0119
0.8	0.2897	0.2780	2.8	0.0079	0.0069
1.0	0.2420	0.2299	3.0	0.0044	0.0038
1.2	0.1942	0.1826	3.2	0.0024	0.0022
1.4	0.1497	0.1394	3.4	0.0012	0.0010
1.6	0.1109	0.1023	3.6	0.0006	0.0005
1.8	0.0790	0.0721	3.8	0.0003	0.0002

APPENDIX FIVE : STANDARD NORMAL (z) DISTRIBUTION : AREA UNDER THE CURVE FROM THE MEAN TO YOUR SPECIFIED VALUE OF Z



Second Decimal Point in Z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952

2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

APPENDIX SIX : THE F DISTRIBUTION : 95th PERCENTILE VALUES FOR THE F DISTRIBUTION

N2 = the degrees of freedom in the numerator, N1 = the degrees of freedom in the denominator.

	N_2																		
N_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4,67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2,19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.77	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64

30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

APPENDIX SEVEN : STUDENTS'S T TEST TABLE OF t

- Select the column with probability that you want.
  - eg. 0.05 means '95% chance'
- Select the row for degrees of freedom.
  - For two values, number of degrees of freedom is $(n_1 + n_2)-2$
- Compare the value in the cell with your t-value.
- The results are significant if the t-value is <u>greater</u> than the value in the cell.

Degrees of Freedom	Probability, p			
	0.1	0.05	0.01	0.001
1	6.31	12.71	63.66	636.62
2	2.92	4.30	9.93	31.60
3	2.35	3.18	5.84	12.92
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.87
6	1.94	2.45	3.71	5.96
7	1.89	2.37	3.50	5.41
8	1.86	2.31	3.36	5.04
9	1.83	2.26	3.25	4.78
10	1.81	2.23	3.17	4.59
11	1.80	2.20	3.11	4.44
12	1.78	2.18	3.06	4.32
13	1.77	2.16	3.01	4.22
14	1.76	2.14	2.98	4.14
15	1.75	2.13	2.95	4.07
16	1.75	2.12	2.92	4.02
17	1.74	2.11	2.90	3.97
18	1.73	2.10	2.88	3.92
19	1.73	2.09	2.86	3.88
20	1.72	2.09	2.85	3.85
21	1.72	2.08	2.83	3.82
22	1.72	2.07	2.82	3.79
23	1.71	2.07	2.82	3.77
24	1.71	2.06	2.80	3.75
25	1.71	2.06	2.79	3.73
26	1.71	2.06	2.78	3.71

27	1.70	2.05	2.77	3.69
28	1.70	2.05	2.76	3.67
29	1.70	2.05	2.76	3.66
30	1.70	2.04	2.75	3.65
40	1.68	2.02	2.70	3.55
60	1.67	2.00	2.66	3.46
120	1.66	1.98	2.62	3.37
infinity	1.65	1.96	2.58	3.29

APPENDIX EIGHT : TABLE OF CRITICAL VALUES OF THE LEAST SQUARES REGRESSION CORRELATION COEFFICIENT

Table of Critical Values

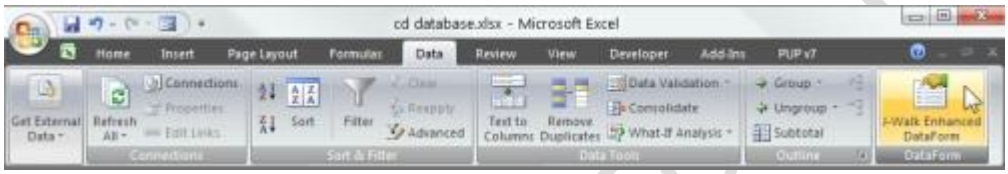
df = N-2 N = number of pairs of data	Level of significance for two-tailed test			
	.10	.05	.02	.01
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.628
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
25	.323	.381	.445	.487
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.284
90	.173	.205	.242	.267
100	.164	.195	.230	.254

APPENDIX NINE : J-WALK ENHANCED DATA ENTRY FORM FOR EXCEL : HOW TO USE IT

The J-Walk Enhanced Data Form is a general-purpose data entry form that works with any worksheet database. A worksheet database is a range of cells that contain fields (columns) and records (rows). The cells in the database can contain values, text, dates, logical values, or formulas. The first row of the database should contain field names.

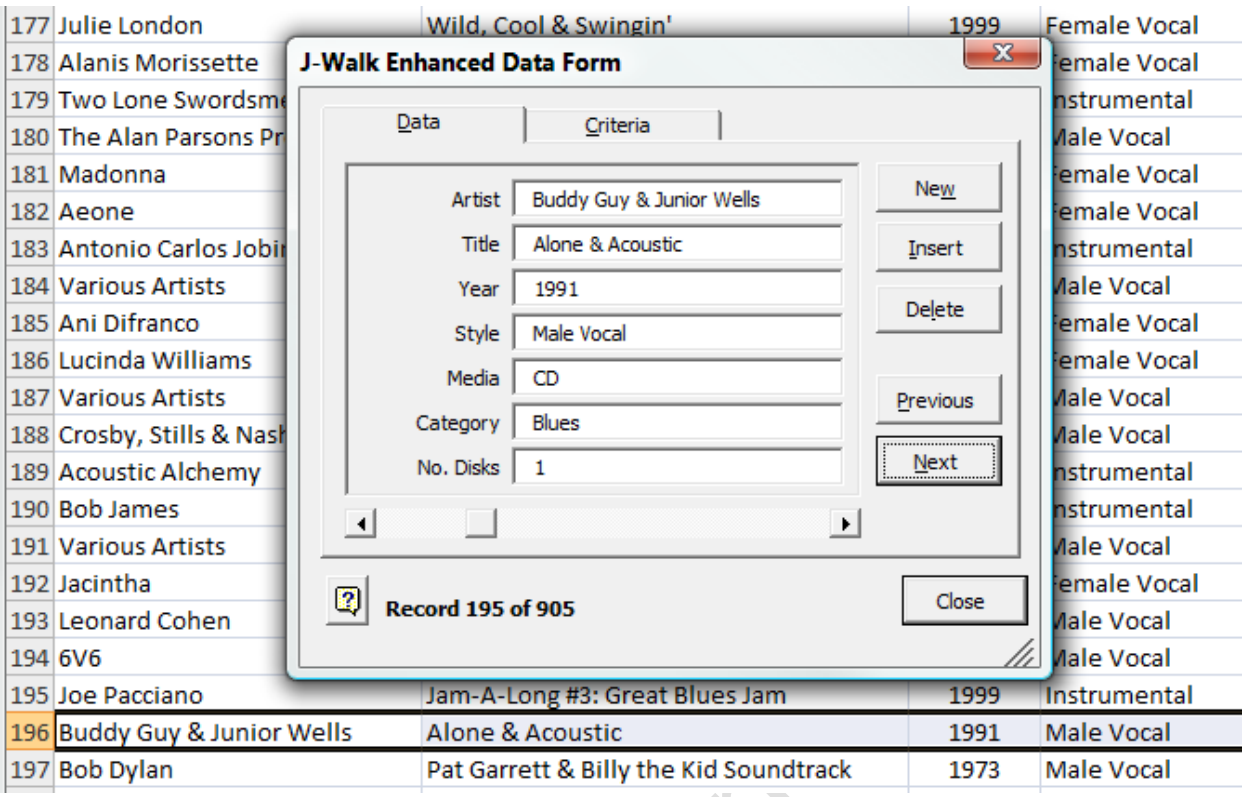
Displaying the Enhanced Data Form

- **For Excel 2007 Users:**
  When the Enhanced Data Form add-in is installed, the Excel 2007 **Data** tab displays a new group DataForm, and this group contains one icon: **J-Walk Enhanced Data Form**. Select any cell in your worksheet database table, and then select the Ribbon command.



- **For Users of Excel 97 - 2003:**
  When the Enhanced Data Form add-in is installed, Excel's **Data** tab displays a new menu item: **J-Walk Enhanced Data Form**. Select any cell in your worksheet database table, and then select the Ribbon command.

The figure below shows the Enhanced Data Form. The exact configuration depends on the number of fields, the field names, and whether the Data Form has been customized. When the dialog box is displayed, it will show the record that corresponds to the active cell. In addition, the database row will be highlighted in the worksheet.



Notice that the dialog box has two tabs: **Data** and **Criteria**. The **Data** tab is used for viewing, editing, and entering data. The **Criteria** tab is to specify search criteria that will enable you to locate specific records.

Changing the size of the form

You can easily change the size of the Enhanced Data Form. Click the lower right corner of the form and drag. You can make the form taller, shorter, wider, or narrower. The width of the data entry fields change accordingly (but the width of the descriptive field names remain fixed).

To adjust the width of the field labels, see Customizing the Data Form.

Viewing data

The horizontal scroll bar (at the bottom of the Enhanced Data Form) is used to quickly activate a particular record. The label below the scroll bar displays the current record number and the total number of records (for example, **Record 195 of 905**). You can change the current record (row) by using the horizontal scroll bar, or by using the **Previous** or **Next** buttons. The data is displayed in the dialog box, and is also highlighted in the worksheet.

Editing data

To change the data displayed in the Enhanced Data Form, activate the appropriate field and use standard editing techniques. Note that you can use the Tab key (and Shift+Tab) to cycle among the fields. If all of the fields are not visible, use the vertical scroll bar. This scroll bar is not present if all fields are displayed. After you've edited the field(s), click **Next** or **Previous** to store the changes in the worksheet. Press Enter (or click **Close**) to close the dialog box.

Adding new data

To add new data to your worksheet database, click the **Insert** button or the **New** button. Clicking the **Insert** button inserts a new row above the current row. Clicking the **New** button adds the data to the end of the database table. After you enter the data in the dialog box, add it to the worksheet by clicking **Next**, **Previous**, **Insert**, or **New**.

When a new record is added, the text [NEW] is entered into the first field. This is done in order to maintain the integrity of the database table.

If your database is an Excel 2007 table (created by using the **Insert - Table** command), the table is automatically expanded to include new data.

Deleting data

Click the **Delete** button to delete the current record. Subsequent records will be shifted upward to eliminate the gap caused by the deleted row.

Undoing operations

After you've made a change to your database, you can undo the change by clicking the **Undo xxxx** button. This button will display the operation that will be undone. For example, it may display **Undo Delete**. There is only one level of undo. The following operation can be undone:

- Insert
- New
- Delete

Entering search criteria

In some cases, you may want to locate records that meet certain criteria. To enter search criteria, click the **Criteria** tab of the Enhanced Data Form dialog box. You'll see the same field names listed in the **Data** tab -- but the background color is different to remind you that you're in the Criteria tab.

Enter the data to find in the appropriate fields. For additional information about searching, click the **Tips** button.

After you've entered your search criteria, click the **Data** tab. You'll find that the dialog box has changed in three ways:

- The **Next** button now displays **Find Next**. Use this button to display the next record that matches your search criteria.
- The **Previous** button now displays **Find Prev**. Use this button to display the previous record that matches your search criteria.
- The **Criteria** tab now displays **<<Criteria>>**. This is just a reminder that search criteria are in effect.

To cancel this search mode and return to normal, click the **Criteria** tab and click the **Clear** button.

APPENDIX TEN : VBA CODE FOR DATA ENTRY FORMS AND ARTICLE BY ….. ON HOW TO BUILD YOUR OWN DATA ENTRY FORM IN EXCEL

VBA code for a simple Data Entry Form as shown in Figure xxxx :

```
Private Sub CmdFileThem_Click()
Dim iRow As Long
Dim ws As Worksheet
Set ws = Worksheets("Demo-One")
'Find first empty row in database
iRow = ws.Cells(Rows.Count, 1).End(xlUp).Offset(1, 0).Row

'Copy data into the database
ws.Cells(iRow, 1).Value = TextBox1.Value
ws.Cells(iRow, 2).Value = TextBox2.Value
ws.Cells(iRow, 3).Value = TextBox3.Value
If OptionButton1 Then ws.Cells(iRow, 4).Value = "Male" Else ws.Cells(iRow,
4).Value = "Female"
ws.Cells(iRow, 5).Value = TextBox4.Value
ws.Cells(iRow, 6).Value = TextBox5.Value
ws.Cells(iRow, 7).Value = TextBox6.Value
ws.Cells(iRow, 8).Value = TextBox7.Value
ws.Cells(iRow, 9).Value = TextBox8.Value
ws.Cells(iRow, 10).Value = TextBox9.Value
ws.Cells(iRow, 11).Value = TextBox10.Value
ws.Cells(iRow, 12).Value = TextBox11.Value
ws.Cells(iRow, 13).Value = TextBox12.Value
ws.Cells(iRow, 14).Value = TextBox13.Value
ws.Cells(iRow, 15).Value = ComboBox1.Value
ws.Cells(iRow, 16).Value = TextBox14.Value

End Sub

Private Sub ComboBox1_Change()
End Sub

Private Sub CommandButton2_Click()
Unload Me
End Sub

Private Sub TextBox11_()
End Sub

Private Sub TextBox11_AfterUpdate()
wt2 = Val(TextBox11)
ht = Val(TextBox9)
If wt2 > 0 Then
BMI2 = wt2 / (ht * ht)
TextBox12 = Str(BMI2)
wtchange = Val(TextBox8) - wt2
```

```
TextBox15 = Str(wtchange)
wt1 = Val(TextBox8)
BMI1 = wt1 / (ht * ht)
changeBMI = BMI1 - BMI2
TextBox16 = Str(changeBMI)
End If
End Sub

Private Sub TextBox4_AfterUpdate()
birthday = TextBox4.Value
today = Format(Date, "dd/mm/yyyy")
age = DateDiff("yyyy", birthday, today)
TextBox14 = Str(age)
End Sub

Private Sub TextBox4_Change()
End Sub

Private Sub TextBox9_AfterUpdate()
wt1 = Val(TextBox8)
ht = Val(TextBox9)
If wt1 > 0 Then
BMI1 = wt1 / (ht * ht)
TextBox10 = Str(BMI1)
End If
End Sub

Private Sub UserForm_Activate()
ComboBox1.AddItem "Placebo"
ComboBox1.AddItem "Xantrax"
ComboBox1.AddItem "Phentermine"
End Sub
```

********** END OF VBA CODE EXAMPLE **********

Build a UserForm for Excel

Introduction

A UserForm is a custom-built dialog box that you build using the Visual Basic Editor. Whilst this example works in Excel you can use the same techniques to create a UserForm in any of the Microsoft Office programs that support VBA.
With a UserForm you can create a user-friendly interface for your workbook or document, making data entry more controllable for you and easier for the user.

About the Project

This document shows you how to build a simple UserForm for entering personal expenses data on to a worksheet in Excel. The work is divided into two main sections: building the form itself and then writing the VBA code to make it work. The finished UserForm will look something like this (*Fig.1*).

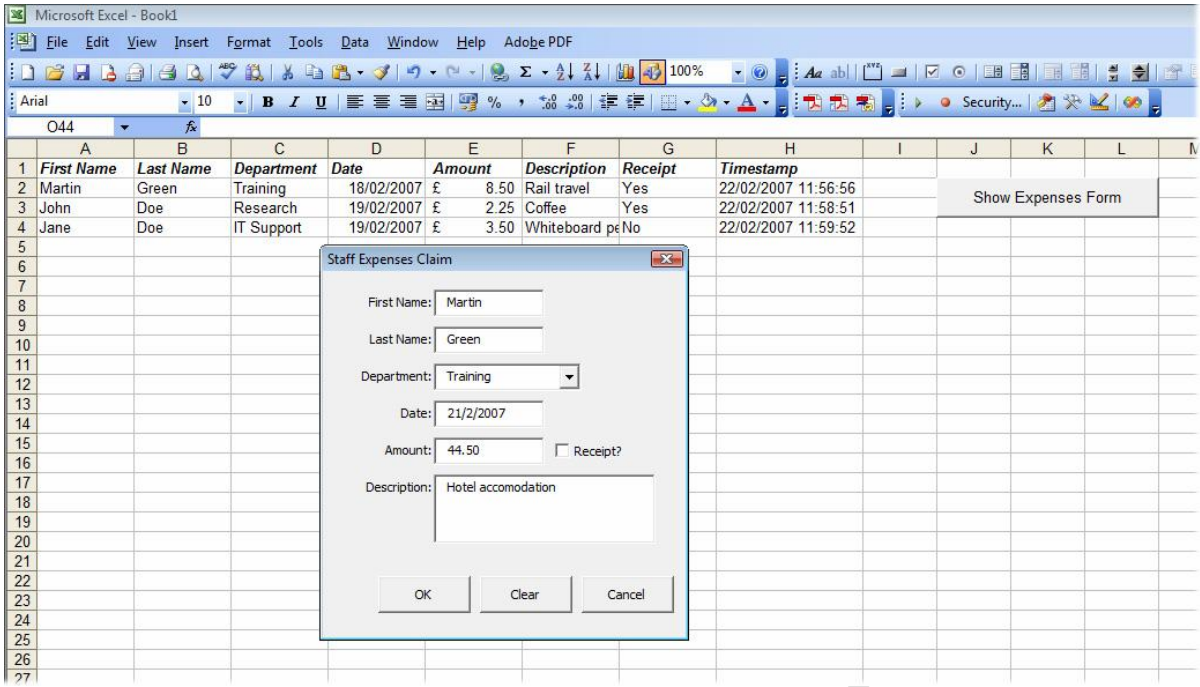


Fig. 1 The finished UserForm project.

NOTE: The screenshots here show how things look in Excel 2003 running on Windows Vista. If you are working in an different version of Excel or Windows the appearance will be slightly different but
all the techniques explained here will be exactly the same.

Build the Form

Start Excel and open the Visual Basic Editor (keyboard shortcut: **[Alt]+[F11]**). You will need to use both the **Project Explorer** and the **Properties Window** so if they are not visible open them from the **View** menu.

HINT: When building your UserForm try to avoid double-clicking on anything (unless the instructions tell you to do so) because this sometimes opens the form's code window. If you do this accidentally simply close the code window by clicking its Close button, or switch back to the UserForm design window with the keyboard shortcut **[Control]+[Tab]**.

Insert a New UserForm

Make sure that the current workbook (e.g. *VBAProject (Book1)*) is selected in the Project Explorer then open the **Insert Menu** and choose **UserForm**. When you do this a new, blank UserForm appears in the code window of the Visual Basic Editor and a corresponding entry appears in the Project Explorer (*Fig. 2*). The Project Explorer shows a new folder named *Forms* containing the new UserForm which has been given the name *UserForm1*.

You should also see the **Toolbox** (*Fig. 3*). If it is not visible click anywhere on the new UserForm (the Visual Basic Editor hides the toolbox when it thinks you are working elsewhere) and if it still does not appear open it from the **View** menu.

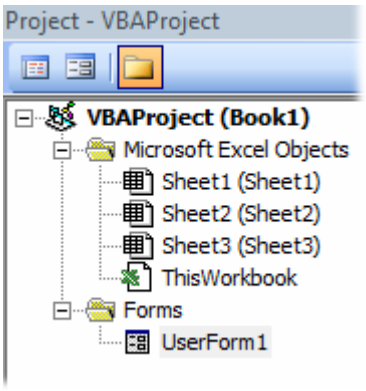


Fig. 2 The Project Explorer shows the UserForm.

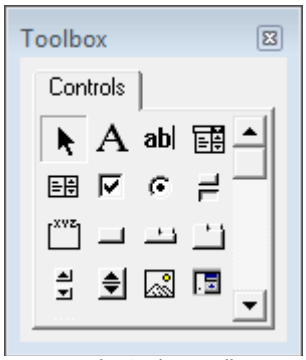


Fig. 3 The Toolbox

The UserForm has a dotted border around it. On the border, in the lower-right corner and halfway along the bottom and right sides of the form, are small white squares. These are the resizing handles. You can use the mouse to drag these handles to make the UserForm the required size. The grid of dots on the form is to help you easily align items you place there.

Rename the UserForm and Add a Caption

A single project can include many UserForms so it is a good idea to give each one a meaningful name. With the UserForm selected find the **Name** property in the Properties Window (it is normally the first item in the list) and change it *frmExpenses*. Then change the **Caption** property to *Staff Expenses Claim*. The Project Explorer now displays the UserForm's new name and the Title Bar of the form immediately changes to show the new caption (*Fig. 4*).

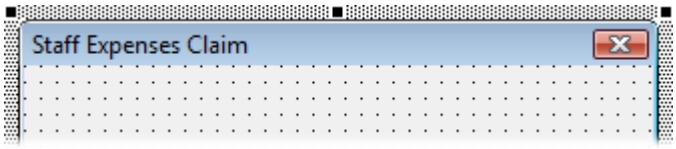


Fig. 4 The form's Title Bar shows the new caption.

When naming forms and their controls remember that you must not include spaces in the name, or
use any of the VBA "reserved words" (i.e. those keywords that are part of the VBA language such as "Date").

Add a TextBox Control and a Label

The *controls* are the items, such as textboxes, comboboxes and command buttons, that will be placed on the form. The standard selection of controls is represented by buttons on the Toolbox. Point at a Toolbox button to see a tooltip showing the name of the control.

Add a TextBox control to the form by clicking on the **TextBox** button in the Toolbox then clicking somewhere near the centre of the form. As with the UserForm itself any control that you place on the form shows a dotted border and resizing handles when the item is selected (click on any item to select it).

You can change the size and shape of a control either by dragging the resizing handles (the mouse pointer will change to a double-headed arrow (*Fig. 5*)) or by changing the values of its *Height* and *Width* properties in the Properties Window. To move a control drag the dotted border at a point between resizing handles (the mouse pointer will show a four-headed arrow (*Fig. 6*)) or change the values of its *Top* and *Left* properties.

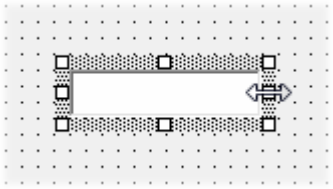


Fig. 5 Resizing a control.

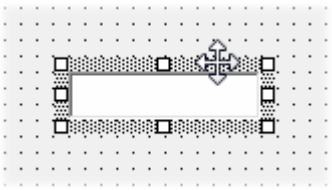


Fig. 6 Moving a control.

Drag the textbox to a point near the top of the UserForm and about halfway across the form. Each control should have a meaningful name so that when you write the code you can easily identify it. This one currently has the name *TextBox1*. Use the Properties Window to change its name to *txtFirstName*.

HINT: It is helpful when naming controls to add a prefix describing the type of control ("txt" for textbox, "cbo" for combobox etc.). This reminds you what type of control it is when you are working in the code. It forces the names to appear together when displayed in a list. It also lets you use words that are otherwise reserved (e.g. txtDate *instead of* Date*).*

Now use the toolbox to place a **Label** control on the form. To change the caption of the label you can either type directly on to the label or you can change its *Caption* property in the Properties Window. Change the label's caption to *First Name:*.

Change the *TextAlign* property of the label to *3-fmTextAlignRight* then double-click the lower-right corner resizing handle to snap the label to fit the text (*Fig. 7*). Drag the label to a position just to the left of the *FirstName* textbox.

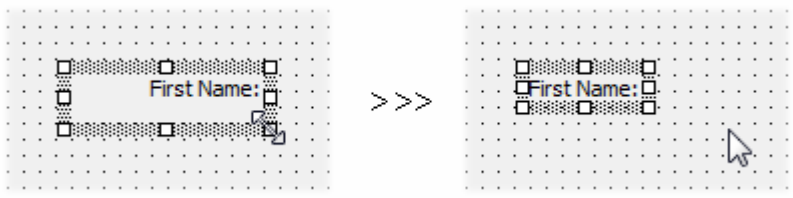


Fig. 7 Double-click the lower-right corner handle to snap the label to size

.

When you move controls around by dragging them they snap to points on the grid. Whilst this is a useful feature, sometimes you need to position objects more accurately. You will notice that you can not place the label exactly level with the centre of the textbox. The grid forces it to be too high
or too low. Use the Properties Window to subtract (or add as necessary) 3 units from the *Top* property of the label so that it is correctly positioned in relation to the textbox (*Fig. 8*).

Fig. 8 Use the Properties Window to finely adjust the position of a control.

It isn't necessary to give the label a different name because, in this project, you will not be referring to it in the code, although in other circumstances you might want to do so.

Add the Remaining Controls

Use the same techniques to add the remaining controls to the form. You need to add four textboxes, a combobox (a textbox with a drop-down list), a checkbox and three command buttons.

Here is a list of the remaining controls you need to add and their properties:

TextBox ..... Name: *txtLastName*
Label..... Caption: *Last Name:*
ComboBox ..... Name: *cboDepartment*
Label..... Caption: *Department:*
TextBox ..... Name: *txtDate*
Label..... Caption: *Date:*
TextBox ..... Name: *txtAmount*
Label..... Caption: *Amount:*
CheckBox..... Name: *chkReceipt*, Caption: *Receipt?*
TextBox ..... Name: *txtDescription*, Height: *45*, Width: *132*, Scrollbars: *2-fmScrollbarsVertical*
Label..... Caption: *Description:*
CommandButton ... Name: *cmdOK*, Caption: *OK*
CommandButton ... Name: *cmdClear*, Caption: *Clear*
CommandButton ... Name: *cmdCancel*, Caption: *Cancel*

*HINT: At any time you can check out exactly how the UserForm will look in use by pressing the **[F5]** key on your keyboard or clicking the **Run** button on the Visual Basic Editor toolbar. Doing this will open the UserForm in its host program (in this case Excel). To return to the Visual Basic Editor, close the UserForm by clicking the close button **[x]** in its upper-right corner.*

The finished UserForm should look something like this (*Fig. 9*):

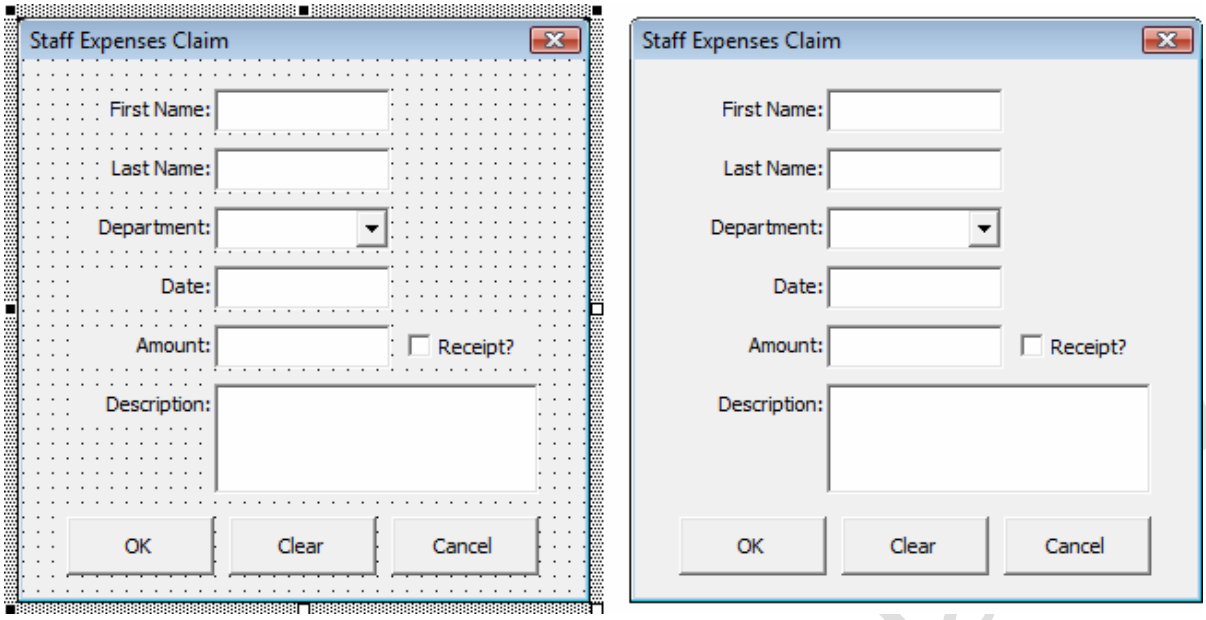


Fig. 9 The finished UserForm in design view (left) and in use (right).

Create the ComboBox List

The *Department* combobox now needs to be told where to get the information to build its list. There are two ways to do this. It can be done with code (this method is described in the next section) or it can refer to a named range of cells in the workbook. The latter method is often preferred because you can easily edit the list without having to rewrite any code.
Switch to Excel and open a worksheet in the same workbook. Type a column of items representing the entries you want to appear in the combobox list. Put one item in each cell. If you want the items to appear in alphabetical order sort the list in the worksheet.
Now select the cells containing the list items and name the range of cells. The easiest way is to click in the *Name* box (the box just above and to the left of cell A1), type the name *Departments* (*Fig. 10*) then press **[Enter]**. Click somewhere else on the worksheet then check that you have correctly named the range by clicking the down-facing arrow to the right of the Name box. You should see your range name in the list. Choose it and check that Excel selects the cells that contain your list.

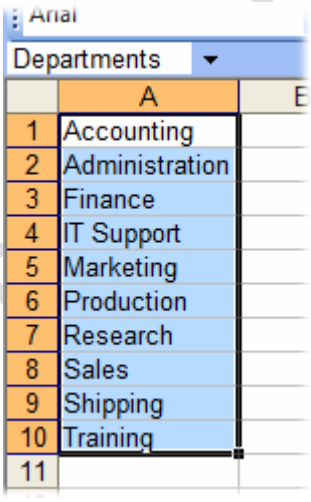


Fig. 10 Name a range of cells containing the list.

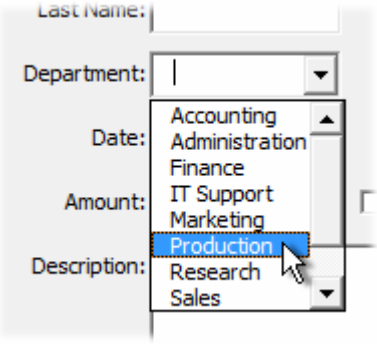


Fig. 11 The combobox displays the list.

If you add items to the list at a later date you may need to redefine the list. You can do this by opening Excel's **Insert** menu and choosing **Name** then **Define**.
Return to the Visual Basic Editor and click on the *Department* combobox to select it then go to the Properties Window and find the *RowSource* property. Enter the same name as you used for the range containing your list (in this example *Departments*).

Test the form (press the **[F5]** key) and see that the combobox now displays your list (*Fig. 11*).
Check the Tab Order

Many people when working in a form like to move around from control to control by clicking the **[Tab]** key on their keyboard. The order in which the tab key moves you around a form is initially defined by the order in which you placed the controls on the form.
Run the form (open it in Excel) and, starting from the *FirstName* textbox, press the **[Tab]** key repeatedly and check that it takes you through the form in a logical order. If you want to change the order close the form and in the Visual Basic Editor open the **View** menu and choose **Tab Order**. Here you can move items up and down the list to control the behaviour of the **[Tab]** key in the form (*Fig. 12*).

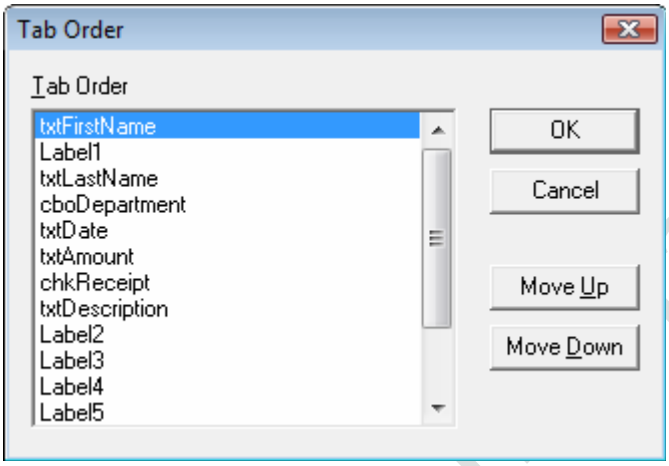


Fig. 12 The Tab Order dialog box.

Write the VBA Code

The design of the form is now finished. The next job is to write the VBA code to power it. The code is needed to power the three command buttons…
Coding the Cancel Button
The *Cancel* button is the simplest one to code. It needs to do the same job as the built-in close button (**[x]**) in the upper right corner of the form.
Double-click the *cmdClose* command button (or right-click it and choose **View Code**) to open the UserForm's code module. The Visual Basic Editor will have written the *Sub* and *End Sub* lines of the button's *Click* event for you. Place your cursor in the empty line between these lines, press your **[Tab]** key then enter the line:

```
Unload Me
```

Your code should look like this (*Listing 1*):

Listing 1

```
Private Sub cmdCancel_Click()
Unload Me
End Sub
```

Test the code. Open the **Debug** menu and choose **Compile VBAProject**. If you get an error message check your typing and compile again. Then switch to the form design window (press **[Control]+[Tab]** or double click the form's name in the Project Explorer). Press the **[F5]** key to run the form in Excel. Click the form's **Cancel** button and the form should close and return you to the Visual Basic Editor.
Coding the OK Button
The *OK* button has three jobs to do. It has to:
1. Check the user's input so that all the required information has been supplied (this is called "validation".
2. Write the data on to the worksheet in Excel.
3. Clear the form ready for the next entry.
In the design view of the form double-click the *cmdOK* button and enter the following lines into the *cmdOK_Click* event procedure (*Listing 2*):

Listing 2

```
Private Sub cmdOK_Click()
If Me.txtFirstName.Value = "" Then
MsgBox "Please enter a First Name.", vbExclamation, "Staff Expenses"
Me.txtFirstName.SetFocus
Exit Sub
End If
End Sub
```

This procedure uses an *If Statement* to check the contents of the *txtFirstName* textbox. If the textbox is empty (i.e. its contents are "" – the two quote marks with nothing between them represents "nothing") a message is displayed, the focus is set to that textbox (the user's cursor is taken there), and the procedure is cancelled.
As before, compile and test the code. Open the form and, without entering anything in the *First Name* text box, click the **OK** button. You should see the error message (*Fig. 13*). Dismiss the message box then type an entry in the *First Name* textbox and try again. No message box should be displayed when you click the **OK** button.

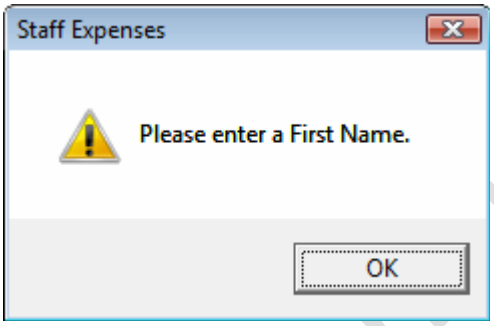


Fig. 13 The error message reminds the user to enter a First Name.

Make a similar entry for each textbox and for the combobox. You can view the complete code listing at the end of this document (*Listing 8*).
In addition to checking that an entry is present, it is also sometimes necessary to check that the entry is correct. Enter the following statements (*Listing 3*). They use the *IsNumeric()* function to check that the value in the *txtAmount* textbox is a number (and not something else such as text); and the *IsDate()* function to check that the value in the *txtDate* textbox is a date.

Listing 3

```
If Not IsNumeric(Me.txtAmount.Value) Then
MsgBox "The Amount box must contain a number.", vbExclamation, "Staff Expenses"
Me.txtAmount.SetFocus
Exit Sub
End If
If Not IsDate(Me.txtDate.Value) Then
MsgBox "The Date box must contain a date.", vbExclamation, "Staff Expenses"
Me.txtDate.SetFocus
Exit Sub
End If
```

Now, having reached a point where all the required entries are present and correct, it's time to
write the entries on to the worksheet. This code involves using a *variable* to hold the number of
rows of data on the worksheet.
Make an empty line at the top of the current procedure, immediately after the statement *Private*
Sub cmdOK_Click() and enter the line:

```
Dim RowCount As Long
```

I have assumed that the entries are going to be made on *Sheet1* of the current workbook, starting
in cell *A1*. You might like to prepare the worksheet by typing a row of headings in the top row. The
code will work the same way if there are headings or not.
Return to the end of your code and enter the new line:

```
RowCount = Worksheets("Sheet1").Range("A1").CurrentRegion.Rows.Count
```

This statement counts how many rows of data are included in the region that includes cell *A1* and
stores that number in the *RowCount* variable. Now enter the lines that write the date on to the
worksheet (*Listing 4*):

Listing 4

```
With Worksheets("Sheet1").Range("A1")
.Offset(RowCount, 0).Value = Me.txtFirstName.Value
.Offset(RowCount, 1).Value = Me.txtLastName.Value
.Offset(RowCount, 2).Value = Me.cboDepartment.Value
.Offset(RowCount, 3).Value = DateValue(Me.txtDate.Value)
.Offset(RowCount, 4).Value = Me.txtAmount.Value
.Offset(RowCount, 5).Value = Me.txtDescription.Value
If Me.chkReceipt.Value = True Then
.Offset(RowCount, 6).Value = "Yes"
Else
.Offset(RowCount, 6).Value = "No"
End If
.Offset(RowCount, 7).Value = Format(Now, "dd/mm/yyyy hh:nn:ss")
End With
```

The code uses a number of similar statements to write the value of each control into a cell. Each cell is identified by its position relative to cell *A1* by using the VBA *Offset* property. This requires two numbers, the first representing the number of rows away from cell *A1* (which is held in the *RowCount* variable), the second representing the number of columns away from cell *A1* (which is written into the code as a number).
Note that the *DateValue()* function is used to change the date entry into a real date (rather than a date represented as text) before passing it to Excel.
The value of a checkbox is expressed as "TRUE" or "FALSE" so, because I wanted to see "Yes" or "No" on the worksheet, the code uses an *If Statement* to make the required entry.
Finally a timestamp is written into the last column using the *Format()* function to specify how the exact time, as supplied by the *Now()* function, is displayed. Now is a good time to compile and test the code again.
To complete the procedure, after the data has been written to the worksheet, the form needs to be
emptied. This requires another variable to be entered at the top of the procedure:

```
Dim ctl As Control
```

This variable represents the controls on the worksheet and will be used in the following loop which visits each control, checks to see if it is a textbox or a combobox, and if it is it sets the control's value to an empty string (""). If the control is a checkbox it sets its value to *False* Enter the following lines (*Listing 5*):

Listing 5

```
For Each ctl In Me.Controls
If TypeName(ctl) = "TextBox" Or TypeName(ctl) = "ComboBox" Then
ctl.Value = ""
ElseIf TypeName(ctl) = "CheckBox" Then
ctl.Value = False
End If
Next ctl
```

Compile and test the code again. If any errors occur then check that your typing is exactly as shown here.

Coding the Clear Button

The function of this button is to clear the form manually if the user wishes to do so. It uses exactly the same procedure as the last part of the *OK* button's procedure, so double-click the *cmdClear* button to create its click event procedure and enter the code shown in *Listing 5*. You can copy the code from the *OK* button's procedure to save time.

Compile, Test and Save the Finished UserForm

That completes the coding of the form. Compile and test the code. If you are satisfied that the form is working correctly save the file. The job is almost finished. All that remains is to create a

macro to open the form.

A Macro to Open the UserForm

As you have seen, it is easy to open the UserForm from the Visual Basic Editor, but the person who
is going to use this tool needs an easy way to open the form from Excel. There are several ways to do this, all involving a macro containing the very simple statement:

```
frmExpenses.Show
```

Manually Opening the Form

This statement can be included in a macro that the user can call from the usual menu. To create this macro first go to the Visual Basic Editor's **Insert** menu and choose **Module** to add a standard module to the workbook containing the UserForm. Enter the following code into the new module's code window (*Listing 6*):

Listing 6

```
Sub OpenExpensesForm()
Worksheets("Sheet1").Activate
frmExpenses.Show
End Sub
```

The first line is optional. It tells Excel to switch to *Sheet1*. But, since the code that writes the data on to the worksheet specifies the worksheet by name, it could be omitted and the data would still be written in the correct place.

The user can run this macro from the menu in the usual way (**Tools > Macro > Macros**) or you could assign it to a custom menu item, toolbar button, a button on the worksheet or a drawing object.

Opening the Form Automatically

You can make use of one of Excel's built-in event procedures to open the UserForm automatically when the workbook is opened. In the Visual Basic Editor locate and double-click the *ThisWorkbook* module in the Project Explorer. This module exists to hold macros specific to the workbook itself. At the top of the code window there are two drop-down lists. The left-hand one will currently read *General*. Open the list and choose **Workbook**. The Visual Basic Editor automatically creates the *Workbook_Open* macro for you. Any code you place in this macro will be executed automatically when the workbook opens. (If you want to see what else you can do here take a look at the other items on the right-hand list.) Complete the macro code as follows (*Listing 7*):

Listing 7

```
Private Sub Workbook_Open()
Worksheets("Sheet1").Activate
frmExpenses.Show
End Sub
```

Complete Code Listing for the UserForm

Here is a complete listing of the code in the UserForm's code module (*Listing 8*)

Listing 8

```
Private Sub cmdCancel_Click()
Unload Me
End Sub
Private Sub cmdClear_Click()
' Clear the form
For Each ctl In Me.Controls
If TypeName(ctl) = "TextBox" Or TypeName(ctl) = "ComboBox" Then
ctl.Value = ""
ElseIf TypeName(ctl) = "CheckBox" Then
ctl.Value = False
End If
Next ctl
End Sub
Private Sub cmdOK_Click()
Microsoft Excel VBA Fact Sheet: Build a UserForm for Excel
© Martin Green www.fontstuff.com 9
Dim RowCount As Long
Dim ctl As Control
' Check user input
```

```
If Me.txtFirstName.Value = "" Then
MsgBox "Please enter a First Name.", vbExclamation, "Staff Expenses"
Me.txtFirstName.SetFocus
Exit Sub
End If
If Me.txtLastName.Value = "" Then
MsgBox "Please enter a Last Name.", vbExclamation, "Staff Expenses"
Me.txtFirstName.SetFocus
Exit Sub
End If
If Me.cboDepartment.Value = "" Then
MsgBox "Please choose a Department.", vbExclamation, "Staff Expenses"
Me.txtFirstName.SetFocus
Exit Sub
End If
If Me.txtDate.Value = "" Then
MsgBox "Please enter a Date.", vbExclamation, "Staff Expenses"
Me.txtFirstName.SetFocus
Exit Sub
End If
If Me.txtAmount.Value = "" Then
MsgBox "Please enter an Amount.", vbExclamation, "Staff Expenses"
Me.txtFirstName.SetFocus
Exit Sub
End If
If Me.txtDescription.Value = "" Then
MsgBox "Please enter a Description.", vbExclamation, "Staff Expenses"
Me.txtFirstName.SetFocus
Exit Sub
End If
If Not IsNumeric(Me.txtAmount.Value) Then
MsgBox "The Amount box must contain a number.", vbExclamation, "Staff Expenses"
Me.txtAmount.SetFocus
Exit Sub
End If
If Not IsDate(Me.txtDate.Value) Then
MsgBox "The Date box must contain a date.", vbExclamation, "Staff Expenses"
Me.txtDate.SetFocus
Exit Sub
End If
' Write data to worksheet
RowCount = Worksheets("Sheet1").Range("A1").CurrentRegion.Rows.Count
With Worksheets("Sheet1").Range("A1")
.Offset(RowCount, 0).Value = Me.txtFirstName.Value
.Offset(RowCount, 1).Value = Me.txtLastName.Value
.Offset(RowCount, 2).Value = Me.cboDepartment.Value
.Offset(RowCount, 3).Value = DateValue(Me.txtDate.Value)
.Offset(RowCount, 4).Value = Me.txtAmount.Value
.Offset(RowCount, 5).Value = Me.txtDescription.Value
.Offset(RowCount, 6).Value = Format(Now, "dd/mm/yyyy hh:nn:ss")
If Me.chkReceipt.Value = True Then
.Offset(RowCount, 7).Value = "Yes"
Else
.Offset(RowCount, 7).Value = "No"
End If
End With
' Clear the form
For Each ctl In Me.Controls
If TypeName(ctl) = "TextBox" Or TypeName(ctl) = "ComboBox" Then
ctl.Value = ""
ElseIf TypeName(ctl) = "CheckBox" Then
ctl.Value = False
End If
Next ctl
End Sub
```

****