

An Introduction to statistics

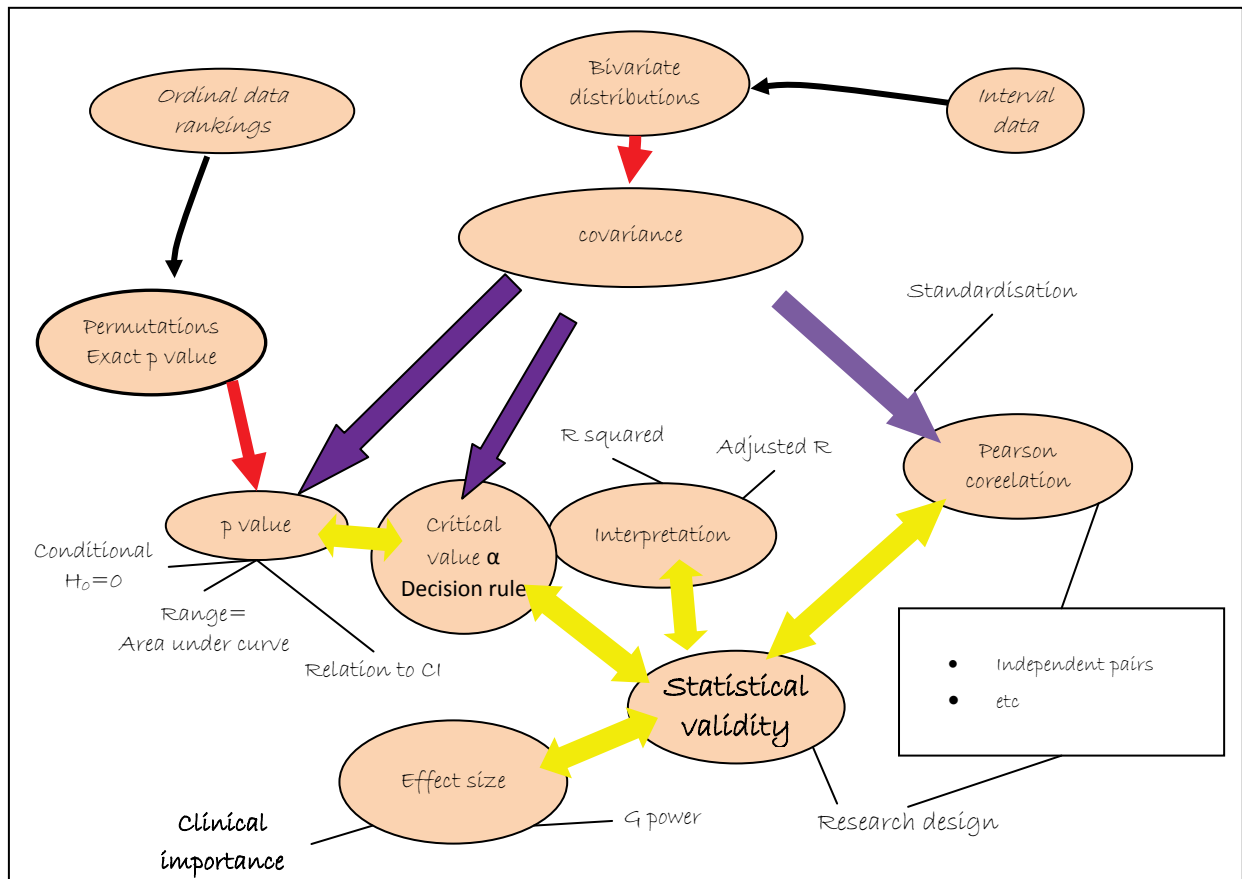
Correlation

Written by: Robin Beaumont e-mail: robin@organplayers.co.uk

<http://www.robin-beaumont.co.uk/virtualclassroom/stats/course1.html>

Date last updated Wednesday, 19 September 2012

Version: 2



How this chapter should be used:

This chapter has been designed to be suitable for both web based and face-to-face teaching. The text has been made to be as interactive as possible with exercises, Multiple Choice Questions (MCQs) and web based exercises.

If you are using this chapter as part of a web-based course you are urged to use the online discussion board to discuss the issues raised in this chapter and share your solutions with other students.

This chapter is part of a series see:

<http://www.robin-beaumont.co.uk/virtualclassroom/contents.html>

Who this chapter is aimed at:

This chapter is aimed at those people who want to learn more about statistics in a practical way. It is the eighth in the series.

I hope you enjoy working through this chapter. Robin Beaumont

Acknowledgment

My sincere thanks go to Claire Nickerson for not only proofreading several drafts but also providing additional material and technical advice.

Contents

1. ASSOCIATIONS, RELATIONSHIPS AND CORRELATION.....	4
1.1 THE BIVARIATE NORMAL PDF	4
1.2 SCATTER PLOTS.....	5
1.3 COVARIANCE	5
1.4 THE STANDARDISED COVARIANCE – CORRELATION COEFFICIENT	8
1.4.1 <i>Examples</i>	8
1.4.2 <i>Interpretation of r</i>	8
2. TAKING INTO ACCOUNT SAMPLING VARIABILITY - ADJUSTED $R = \text{RHO}(P)$	9
3. SAMPLING DISTRIBUTION OF $\text{RHO}(P)$ POPULATION CORRELATION COEFFICIENT=0	9
4. TWO SIDED P VALUE INTERPRETATION	9
4.1 DECISION RULE FOR THE CORRELATION COEFFICIENT	10
4.2 CONFIDENCE INTERVAL	10
4.3 EFFECT SIZE	11
5. COEFFICIENT OF DETERMINATION (R^2)	11
5.1 COMPARISON BETWEEN R^2 AND R	12
5.2 ASSUMPTIONS OF THE PEARSON CORRELATION COEFFICIENT	12
6. RANK CORRELATION.....	12
6.1 MONOTONIC VERSUS LINEARITY	13
6.2 ANGLE BETWEEN LINES OF BEST FIT	14
6.3 STANDARDISED SCORE PLOTS	15
7. COMMON MISCONCEPTIONS ABOUT CORRELATION.....	16
8. CARRYING OUT CORRELATIONS	17
8.1 DO IT IN R COMMANDER	17
8.2 DOING IT DIRECTLY IN R.....	18
8.3 MULTIPLE SCATTERPLOTS AND CORRELATIONS IN R DIRECTLY	19
8.4 TIPS AND TRICKS	21
8.5 ENTERING DATA AND RUNNING CORRELATIONS FROM WITHIN R FOR SMALL DATASETS	21
8.6 SPSS	22
8.7 OBTAINING RANK ORDER CORRELATIONS.....	22
9. RELIABILITY AND R / R^2	22
10. MULTIPLE CHOICE QUESTIONS.....	24
11. SUMMARY	26
12. REFERENCES	26
13. APPENDIX R CODE	27

1. Associations, Relationships and Correlation

So far we have focused on looking at one variable at a time. Although you could argue that when we looked at the differences between different groups of cases for a particular variable such as height divided into groups of males and females we were considering two variables simultaneously. In other words a ratio variable (height) was considered along with the results divided up into two groups using a nominal variable (sex). In contrast to this we will now consider the situation where the two variables are both ordinal, interval or ratio.

Would it be sensible to say that there might be an association or relationship between sex and height? People often do? The problem is with the words 'association' and 'relationship' in common parlance they tend to mean 'being closely connected or moving together' this is too nebulous and open to misinterpretation (The Wikipedia article concerning correlation has some good graphical examples of this http://en.wikipedia.org/wiki/Correlation_and_dependence). Therefore as a consequence of this I will avoid using the terms and restrict myself to the more precise term of correlation (Howells 1992 p.265 provides a reason for the restricted use of the term association). Correlation can be defined as:

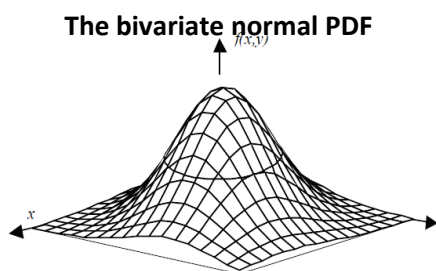
'The degree to which the points cluster about the line of best fit' (Howell 1992 p.223)

I would add 'for two variables that possess, interval or ratio measurement'. The line of best fit is also called the **regression line** for reasons that will be discussed in the chapter on simple regression. The statement above assumes that the correlation is concerned with a 'straight' line in other words it is a **linear** relationship. There are statistics and methods available for quantifying the degree of clustering around lines that are not straight which we will also discuss later.

Sir Francis Galton (1822-1911) first proposed the term "co-relation" between two variables in 1888 when studying the height and length of forearm of 348 men (Snedecor & Cochran 1980 p. 178) and we will in a subsequent section investigate this in more detail. In fact much of statistics in the first decade of the 20th century was concerned with measuring and correlating variables within large data sets, just looking at the titles of the articles in the journal Biometrika from 1900 to 1920 is fascinating and by 1920 the famous Karl Pearson included an article entitled, "Notes on the history of Correlation" (Pearson 1920).

1.1 The bivariate normal PDF

It is important to realise that we are no longer really thinking of two separate variables but a value that has two aspects, think of each value having an x and y part, equivalent to a **point** on a two dimensional plane (i.e. x and y axis). To plot the distribution of these points we need a third dimension, to show how many of these points occur for each value. This means we have moved from a flat (two dimensions) normal distribution to one with three dimensions which looks like an upturned pudding basin.



Exercise:

Visit: <http://demonstrations.wolfram.com/TheBivariateNormalDistribution/> to have a look at a live bivariate distribution and see how varying the parameters changes its shape.

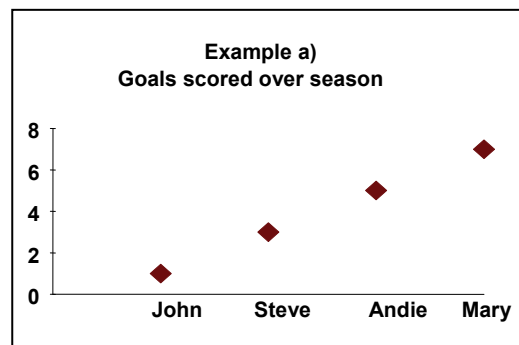
We will start our discussion of correlation by considering the usual way of representing data that we think might be correlated, that is the degree to which two ordinal/interval/ratio variables 'co-vary'.

1.2 Scatter Plots

Two examples of scatter plots are given below.

Example a) Misuse considering nominal data

This scatter plot provides details of goals scored over a session by various players. It shows what, I think, someone might mistakenly call an association or relationship. Clearly it is not possible to investigate the correlation between these two variables using the above definition as the x axis consists of levels of a nominal variable. Notice that the x axis values could have been placed in any order it just so happens that they have been placed in what looks like a 'correct' order. Remember nominal data cannot be ordered.

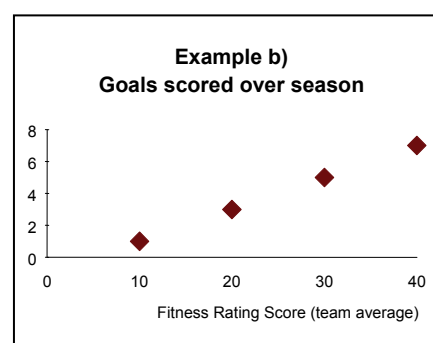


Question: what might be a better way of presenting this data?

Answer: The usual method of presenting nominal data is to use a bar chart.

Example b) appropriate use of scatter plot

This scatter plot provides details for two ratio variables, goals scored over the season and a mythical fitness rating score taken just before the game.



It is all very well looking at scatter plots (in fact it is an essential preliminary step) when considering the possibility that two variables may be correlated. However what we need is an actual measure of the degree to which the two sets of scores 'co-vary'. We will start considering such measures by looking at a measure called the covariance.

1.3 Covariance

Let's start by reconsidering the **sample variance** (introduced in the 'spread' chapter, equation below) this

was the average of the squared deviations from the mean for a single variable. Expanding this expression we end up with the equation opposite. We know this measure increases as the values for the variable become more spread out, also we know that the value in the first bracket is the same as that in the second. Substituting the scores from the second variable in the second bracket gives us what is known as the covariance.

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

Looking at the equation it appears that the covariance is an average measure of the deviations from both means. By considering several scenarios we will investigate what this means.

$$Covariance = cov_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Scenario 1 – all points on line rising.

$$\text{Covariance} = \text{cov}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Take the following pairs of scores (x,y); (1,1), (2,2), (3,3), (4,4) – if you make a quick sketch of the points they all fit on a straight line with an angle of 45% also both have a mean of $10/4=2.5$ so the covariance for these is

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{(1-2.5)(1-2.5) + (2-2.5)(2-2.5) + (3-2.5)(3-2.5) + (4-2.5)(4-2.5)}{4-1}$$

$$= \frac{(-1.5)(-1.5) + (-0.5)(-0.5) + (0.5)(0.5) + (1.5)(1.5)}{3}$$

$$= \frac{2.25 + .25 + .25 + 2.25}{3} = \frac{5}{3} = 1.666$$

What use is this value of 1.666 well it just so happens to be equal to the product of both the standard deviations for each of the groups. Which in this instance is $s_x s_y = 1.29099 \times 1.29099 = 1.666$ So we can say provisionally that if all the points are on a line of 45% then the covariance value is equal to the product of the sample variances.

Scenario 2 – all points on line – falling.

Take the following pairs of scores (x,y); (1,4), (2,3), (3,2), (4,1) –if you make a quick sketch of the points they would all fit on a straight line with an angle of -45% this time. Both variables again have the same mean 2.5 and also the same variance 1.29099 so what is the covariance this time:

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{(1-2.5)(4-2.5) + (2-2.5)(3-2.5) + (3-2.5)(2-2.5) + (4-2.5)(1-2.5)}{4-1}$$

$$= \frac{(-1.5)(1.5) + (-0.5)(0.5) + (0.5)(-0.5) + (1.5)(-1.5)}{3}$$

$$= \frac{-2.25 - .25 - .25 - 2.25}{3} = \frac{-5}{3} = -1.666$$

So this time we get the same value except it is negative. And once again it is equal to the product of the two standard deviations.

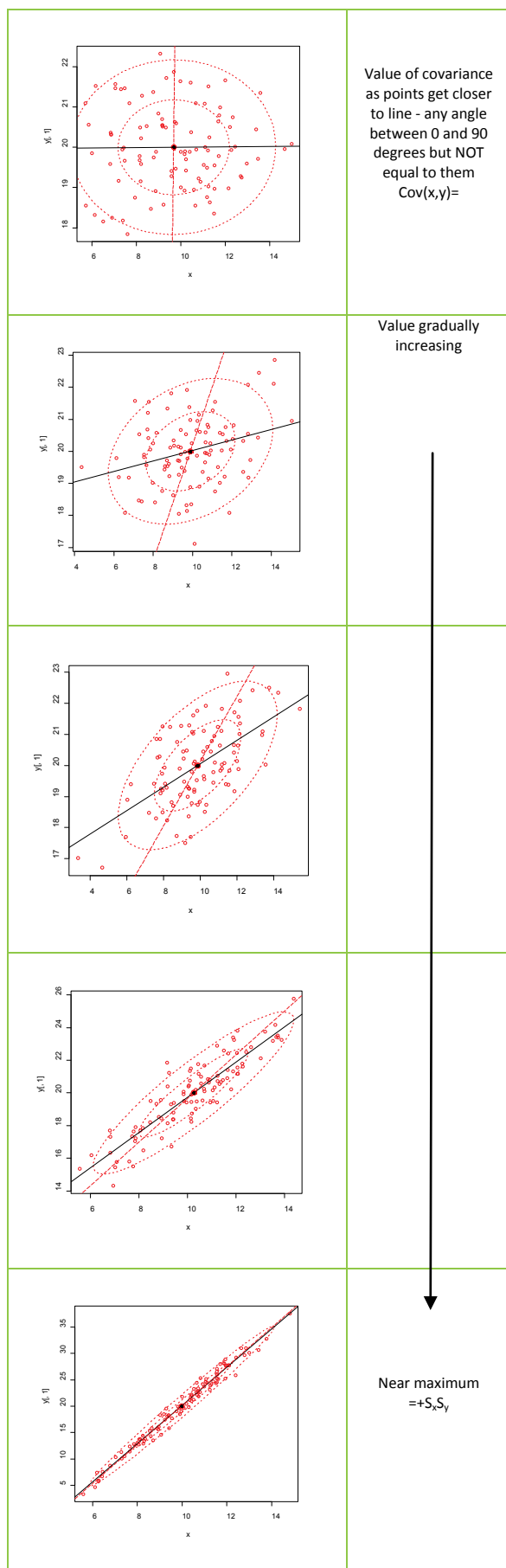
In fact regardless of the angle of the line if all the points fall on it the covariance is equal to the sum of the standard deviations. Just to make sure you have got it:

When all the points are on a straight line the covariance = product of the standard deviations for each variable
= $s_x s_y$

There is one little proviso to the above; if either all the x or y values are the same then the variance of each respectively will be zero (s_x or s_y) so as zero times anything equals zero so will the covariance. Graphically this is the same as saying that if the line is either perfectly horizontal or vertical the covariance will be zero.

Exercise 1.

Taking the following pairs of scores (x,y); (1,1), (2,1), (3,1), (4,1) first sketch them and then work out the covariance. What does it show?



Given that when the line is not either perfectly horizontal or vertical and when all the points are on a straight line, it is all very well knowing that the covariance is equal to the product of the two standard deviations, but it would be nice to know how this descriptive statistic behaves when you do not get a perfect fit? Does it get larger or smaller, in other words are the values we have found so far the entire range of all possible values of covariance's for a particular dataset.

I could create several other datasets to demonstrate what happens when a set of points are mostly off a straight line. However I feel you may have had enough now so will just tell you that the above values are the minimum and maximum values the covariance can achieve for a dataset.

When the points do not cluster to any degree along a straight line the covariance is near zero and gradually either increases or decreases in value until it is equal to the product of the standard deviations depending upon the line rising or falling given our proviso about horizontal and vertical lines. The diagrams opposite illustrates this for the situation with a rising line – I could have easily have produced a diagram for the falling line situation.

The maximum value of the covariance is a bit of an indistinct number which is specific to each datasets standard deviation? It would be nice to be able to have a number that we can remember that tells us, for all data sets, how close they cluster around a straight line.

Once again the z score approach comes to our rescue—remember by subtracting the mean from each of the scores in a dataset and dividing by the standard deviation we end up with a set of scores with a mean of zero and a standard deviation of one which we can then compare with the standard normal pdf. Lets apply the same approach to this situation

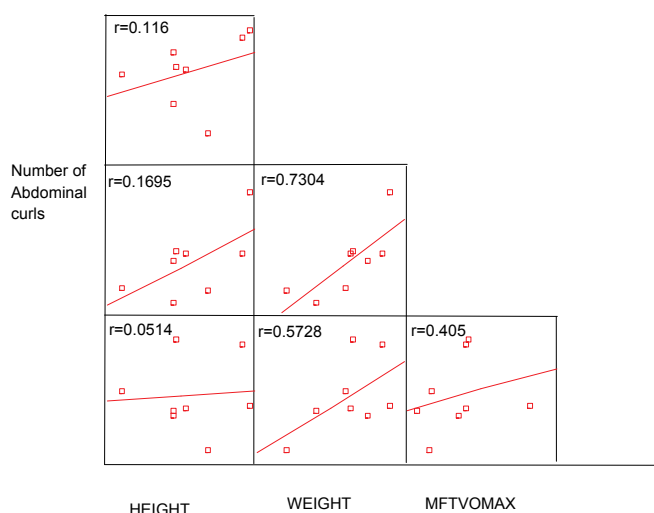
1.4 The standardised covariance – correlation coefficient

As we know what the maximum value our covariance can achieve (the product of the standard deviations) by dividing the observed covariance by this value we know that the maximum value of $-1/+1$ is achievable for any dataset. We will call this new value the Pearson Product-moment correlation coefficient and use the letter 'r' to identify it.

$$r = \frac{\text{cov}(x, y)}{S_x S_y} = \frac{\text{Covariance}}{\text{product of standard deviations}}$$

A coefficient is the mathematical word for a particular type of 'value'. A correlation coefficient, or value, is therefore another descriptive / summary statistic which is in this instance concerned with the degree of 'co-variation' between two variables.

1.4.1 Examples



We will now consider what correlation's from actual data sets look like. Below are a set of scatter plots of data relating to the number of abdominal curls, height, weight and Vo_2Max estimations taken from a group of students.

It is rather a sobering demonstration of what actual data looks like! The highest correlation is between height and Vo_2Max . Weight and abdominal curl score has a r of 0.1695 pretty near zero not what one would have guessed. There are several explanations of these findings, which in an academic

paper would be thoroughly discussed. Possible explanations include biased data collection and the small numbers of subjects.

1.4.2 Interpretation of r

Variables that produce a correlation value close to $+1$ or -1 are said to be highly correlated. A value close to $+1$ yields a high **positive correlation** and a value close to -1 yields a high **negative correlation**. When variables produce a correlation near 0 they are said to be **uncorrelated**. Clearly all this is open to personal interpretation. Just to emphasis it you cannot say that a sample with a correlation of $r=.67$ has a relationship of 67% (Howell, 2007 p238). However the r squared value is a useful measure and we will discuss this latter.

It must be stressed that the correlation 'value' only describes the degree of clustering of points near the line of best fit. Clearly this may happen, or not, by accident. And obviously we would prefer a way of knowing if this is the case or not. In other words we need to be able to infer from our sample to a population which either has a perfect correlation or does not have one and compare it with our sample. For technical reasons we tend to consider a population which as a zero correlation coefficient, similar to the situation with the paired t statistic we considered which had a mean of zero. To consider this we need to go through a similar process to the one we did concerning the various t –statistics, develop a sample estimate of the population parameter, develop a sampling distribution (PDF-the null hypothesis) and from which gain a p value given our particular PDF. Let's start by thinking about how to estimate the population correlation coefficient from our sample, in other words taking into account the sampling variability.

2. Taking into account sampling variability - Adjusted $r = \rho$ (p).

As with the other statistics we have discussed there is a version for the population (which we very rarely know) along with a sample estimate. In this instance the above value is a biased estimate of the population correlation coefficient. **Even though it is this one that is normally reported!** The same issues come into play which we discussed when investigating samples earlier on (remember – if not turn back and revise the chapter), most importantly sample size. So it is no surprise that a sample estimate of the population correlation coefficient is calculated with some ‘n-something’ added, you can see the formulae below. The important thing to realise is that usually it is lower than the unadjusted correlation coefficient as we are not usually taking into account the random sampling issues with the discrepancy decreasing as the sample size increases.

$r_{adj} = \sqrt{1 - \frac{(1 - r^2)(N - 1)}{N - 2}}$ As with the population and sample values for the mean and variance we also represent the estimate for the population correlation coefficient by a greek letter, this time its ρ (p).

3. Sampling distribution of ρ (p) population correlation coefficient=0

To be able to provide a P-value for our correlation (r) we need to be able to define first the sampling distribution, of r (Norman & Streiner 2008 p.140). but to start with we need to know if the statistic (r) has a standard error, that is the standard deviation of (r) across random samples from a population with a correlation coefficient of ρ . This value can be defined and is:

$$SE_r = \sqrt{\frac{1 - r^2}{N - 2}}$$

So now we can create an expression analogous to the t statistic; the observed correlation divided by the variability one would expect across random samples:

$$t_{df=N-2} = \frac{\text{Observed correlation}}{SE_r} = \frac{r}{\sqrt{\frac{1 - r^2}{N - 2}}} = \frac{\text{signal}}{\text{noise}} = \frac{r\sqrt{N - 2}}{\sqrt{1 - r^2}}$$

Luckily for us this statistic follows a t pdf (Fisher 1915) so we can obtain a p-value.

You may remember that when the degrees of freedom (df) for the t distribution (in this instance the number of pairs -2) was over 30 there was little difference between the t pdf and the normal which obviously still holds and also once again we can simply use the computer to produce a p value.

4. Two sided P value interpretation

In this section I will be repeating within the context of correlation much that I said in the chapter on ‘assessing a single mean’. The p value is a conditional probability which includes a range of r values. Specifically we now have:

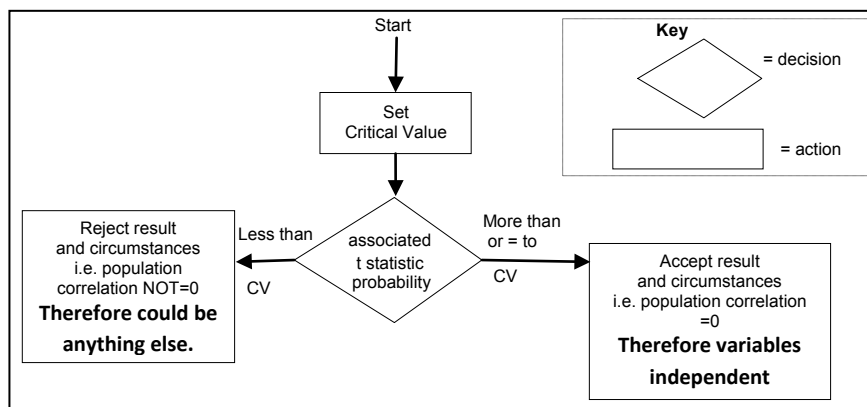
p value associated with our r value =

$P(t \text{ statistic value} | \text{assumptions}) = P(t \text{ statistic}_{(df=n-2)} | p=0) = P(t \text{ statistic}_{(df=n-2)} | \text{null hypothesis is true})$

$p \text{ value} = 2 \cdot P(t_{(n-2)} < t | H_0 \text{ is true}) = 2 \cdot [\text{area to the left of } t=t \text{ under a } t \text{ distribution with } df=n-2]$
(as symmetric)

4.1 Decision rule for the correlation coefficient

We can do exactly the same thing with the other statistics, as shown below:



4.2 Confidence interval

In previous chapters we have calculated the confidence interval by using a measure of the standard

deviation of the parameter estimate (i.e. the standard error). This time it is a bit trickier because the sampling distribution of the correlation is only normal when the estimated population correlation is assumed to be zero otherwise the distribution tends to bunch up to either the minimum or maximum value.

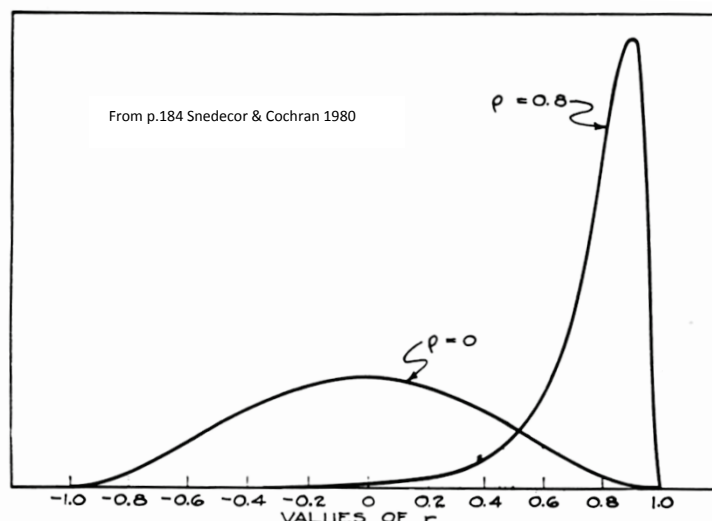


Fig. 10.5.1—Distribution of sample correlation coefficients in samples of 8 pairs drawn from two normally distributed bivariate populations having the indicated values of ρ .

The diagram opposite from George Snedecors classic book on statistics (Snedecor & Cochran p.184) illustrates this, this is what one would expect, as we now only have a range of -1 to +1 so nearer to the extremes our correlation is the more squashed up the distribution becomes.

To overcome this problem with the non normal shape of the distribution three approaches have been taken:

- Transform the values so that they follow a standard normal distribution – This was the traditional method using what is called Fishers transformation, then we can use our good old z score again to create the necessary confidence interval. This is the method both R uses in the above output and also in Andy Fields SPSS syntax given latter.
- Use the t statistic value above (this is the method SPSS uses to calculate a p value).
- Computer intensive approaches – Either create a permutation test or use the boot strapping procedure we described in earlier chapters. For a good introduction see the Wikipedia article: http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient SPSS has an add-on package called bootstrapping which allows you to do this with a click of the mouse.

The R literature does not describe the method it uses in the `cor.test()` function

Exercise 2.

Download the small simulation program at: http://www.vias.org/simulations/simsoft_rdistri.html and see how the distribution shape changes for different values of both ρ and n .

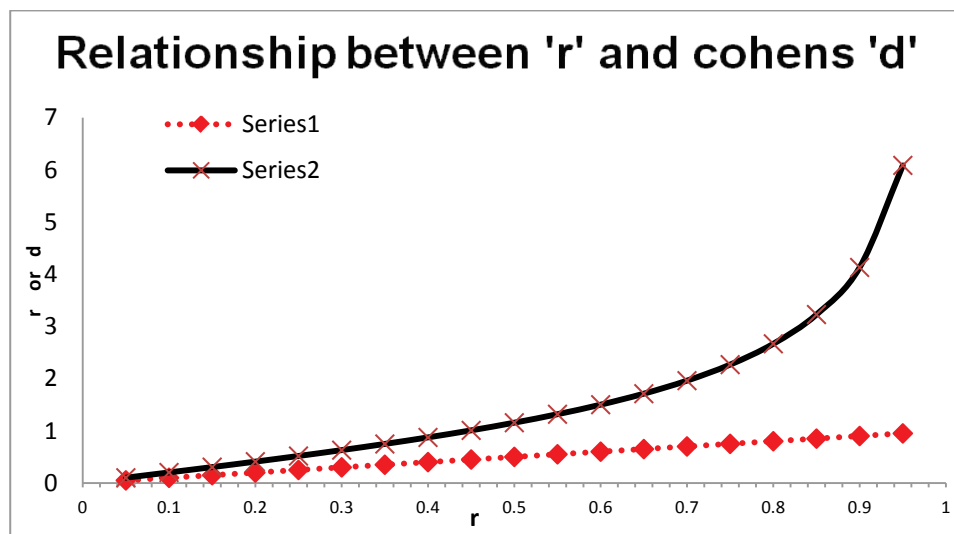
4.3 Effect size

Norman & Streiner 2008 is an excellent resource for those difficult to find aspects such as effect size measures and does not let you down when it comes to correlation. The effect size measure d is calculated thus:

$$d = \frac{2r}{\sqrt{1-r^2}} \text{ and } r = \frac{d}{\sqrt{d^2 + 4}}$$

The table opposite provides details of the relationship between d and r and you can see that as r gets larger. Also importantly from the graph below notice that d more rapidly rises as we get closer to the maximum correlation value.

r	d
0.05	0.1001
0.1	0.2010
0.15	0.3034
0.2	0.4082
0.25	0.5164
0.3	0.6290
0.35	0.7473
0.4	0.8729
0.45	1.0078
0.5	1.1547
0.55	1.3171
0.6	1.5000
0.65	1.7107
0.7	1.9604
0.75	2.2678
0.8	2.6667
0.85	3.2271
0.9	4.1295
0.95	6.0849
1	Division by zero

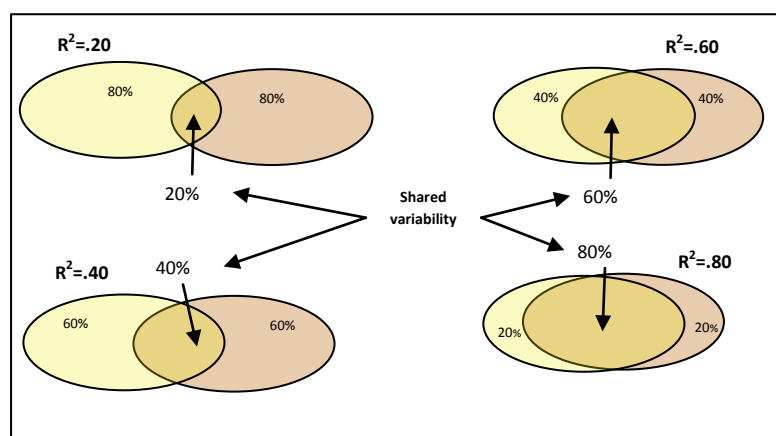


A more widely used measure which can also be classed as an effect size measure is r^2 called the **coefficient of determination** or more commonly called **r squared**.

5. Coefficient of Determination (r^2)

The correlation coefficient is often assumed to equal some type of proportion for example $r = 0.18$ is said to indicate that 18% of the points are on or near the line of best fit, unfortunately this is not the case, although it must be admitted that such a straightforward interpretation of the measure would be welcomed by most people. However it is possible to provide a readily understandable interpretation by using the square of the correlation.

r^2 can be considered as a proportion, unlike r , as it is in effect measuring the proportion of explained variation compared to the total variation. Put another way it is the proportion of predicted to total variability. Therefore when all the observed variation is accounted for by the predicted portion (the line of best fit) it will equal one.



5.1 Comparison between r^2 and r

r	r^2
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25
0.6	0.36
0.7	0.49
0.8	0.64
0.9	0.81
0.95	0.90
0.98	0.96
0.99	0.98

Authors hold a wide range of views concerning which is the best measure to report either r or r^2 . See Howell 1992 p. 242. I prefer to use r^2 . Most statistical applications, tends to produce correlation's however, but it is a simple matter of squaring the values to obtain r^2 and it is instructive to examine a table of the r and associated r^2 values to see how large correlation's can appear to mean a remarkably small amount if we re-interpret them as coefficients of determination, for example a correlation of 0.5 means that only 25% of the variability is accounted for by the correlation model.

5.2 Assumptions of the Pearson correlation coefficient

Because we have assumed that we are sampling from a bivariate normal distribution it implies that both the x and y variables are themselves normally distributed. Peat & Barton 2005 (p.156) list the following assumptions:

- both variables must be normally distributed
- the sample must have been selected randomly from the general population
- the observations (i.e. each pair) are independent of one another
- the relation between the two variables is linear
- the variance is constant over the length of the data

While in the above a linear relationship is stated this is only true for the Pearson correlation and as we will see below this is slightly relaxed when we consider the various correlations based upon ranks. The last assumption is important and often missed when you have data over a wide range of values.

6. Rank correlation

The nonparametric equivalent to the Pearson correlation coefficient, the **Spearman rho correlation**, often indicated by the Greek letter ρ (rho), does what it says – you just convert the dataset into a set of ranks, if they are not ordinal data, and then apply the Pearson formula to the ranks. Most text books provide a separate formula for the Spearman correlation however this equation that looks so very different from the Pearson correlation is in fact identical, if you don't believe me look at Marques de Sá JP 2007 (p.69).

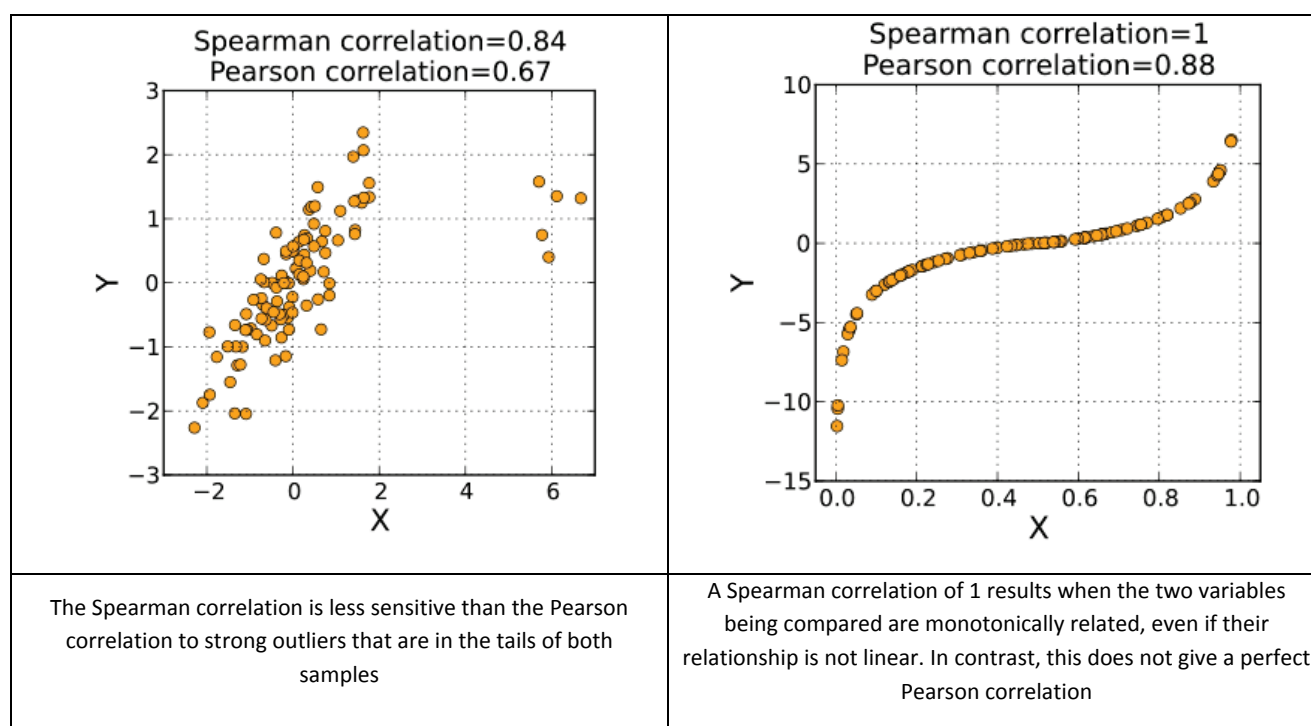
The problem with the Spearman correlation is the same as that of the other non parametric statistics we have already discussed – Ties, if there are a large number of them the result is unreliable but an alternative statistic called **Kendall's tau (τ)** coefficient can be used here. Kendalls tau comes in three varieties called Kendall's tau-a, Kendall's tau-b and I think you can guess the third variety. Wikipedia has a good article about both the spearman and Kendall correlations.

6.1 Monotonic versus linearity

A monotonic function is one that either always increases or decreases. That is we have the possible situation of a snake like line as shown in the right hand side plot below. Following my own advice about Wikipedia I have used two of the diagrams from the spearman correlation article (http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient) below.

The left one illustrates how the Spearman correlation is more robust concerning outliers, you should have a feeling of something similar here to what we discussed in one of the very early chapters concerning how the median behaves compared to the mean with outliers.

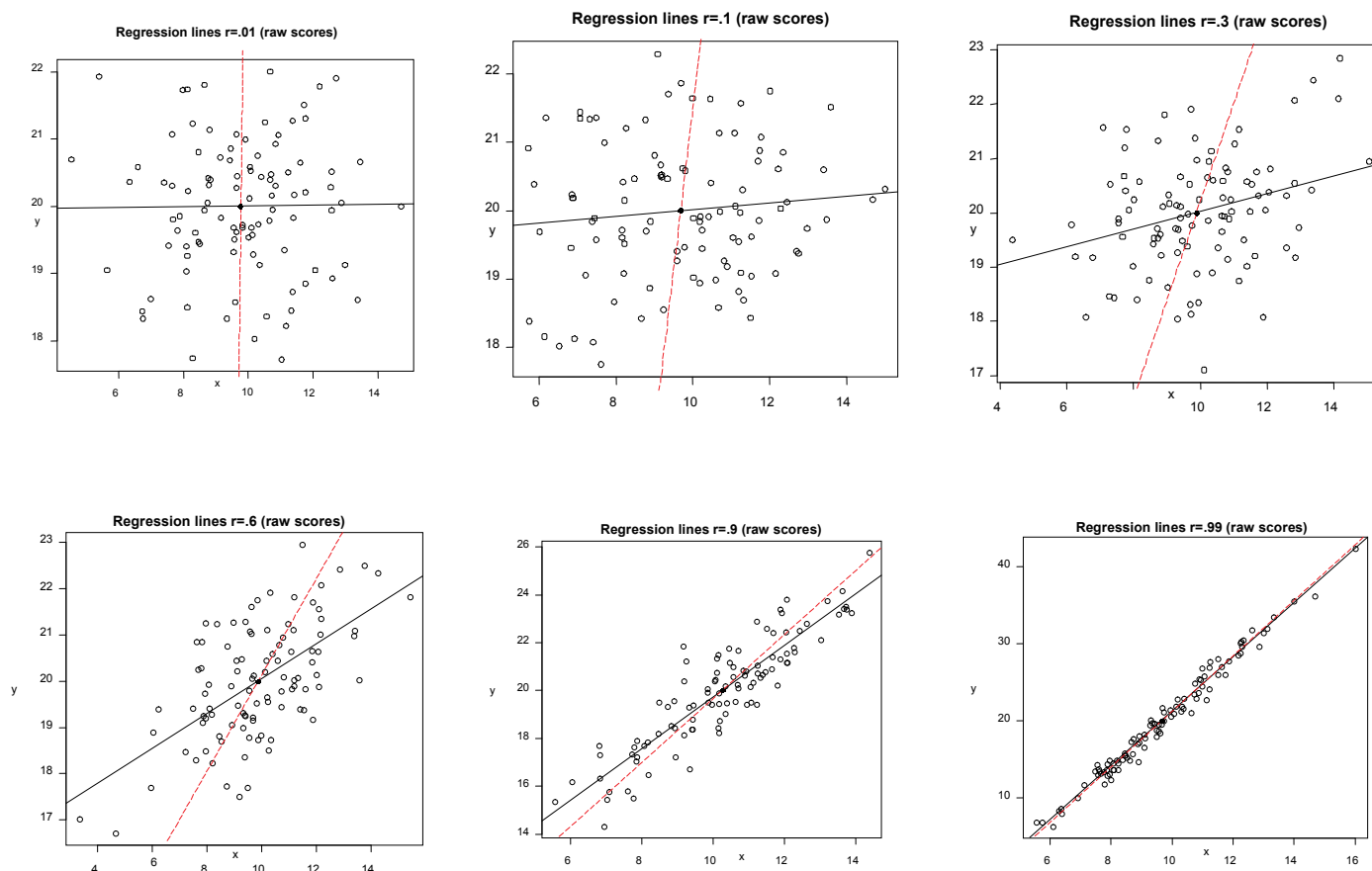
The second aspect is also the result of ranking the data, the actual gaps between each value are irrelevant just their order so as long as the x and y value go in the same direction it adds towards the correlation. This is illustrated nicely in the second diagram.



By the rank order correlation considering monotone functions rather than linear functions, which we had with the Pearson correlation, means that we can consider associations that we feel are not linear but are either increasing or decreasing.

6.2 Angle between lines of best fit

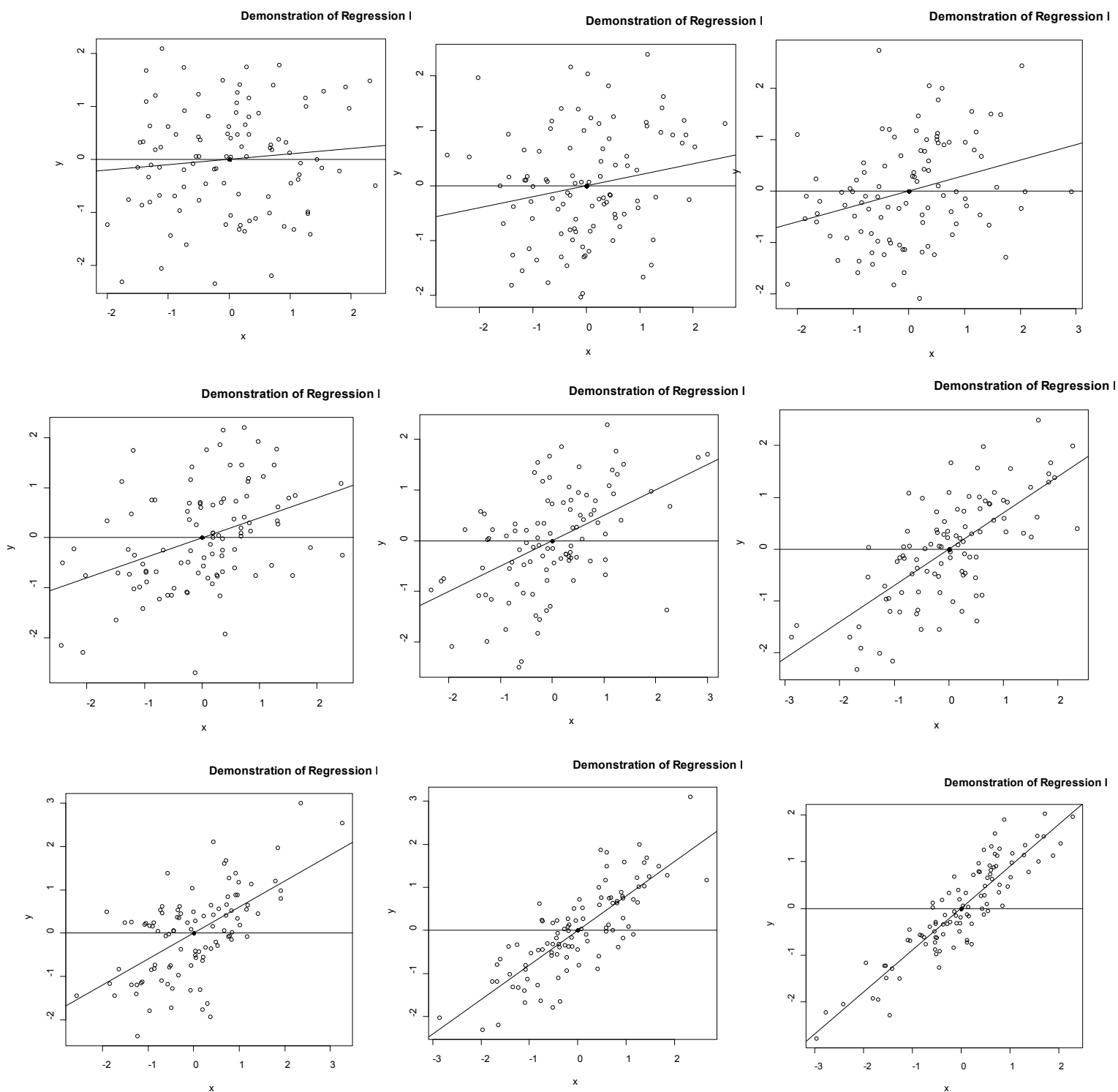
The line of best fit we have talked about will be the centre of our attention in a number of chapters. But I thought that a gentle introduction here to a particular aspect of it is pertinent. The scatterplot below shows two lines one that minimises the horizontal lines from the points to the line (dotted line) and the other which minimises the vertical lines from the points to the line (the usual line of best fit). Now if you consider the cosine of the angle between them when the lines are at right angles the cosine(90 degrees)= 1 and when all the points are on both straight lines the cosine(0) = 1 in fact **if you convert the x v values to z scores the gradient (i.e slope) of the standard line of best fit (minimising the vertical errors) is also equal to r.** (see Rodgers & Nicewander, 1988, Pagano 1990, p123).



You can see from the above that even a correlation of .9 looks like it has a far amount of dispersal around both lines. I always find that it is best to standardise the x and v values before plotting the data as then you can compare them to both possibly past experiences and also a standard image you might have.

To help you I have provided some examples over the page.

6.3 Standardised score plots



Ignore the Horizontal line it is purely for reference purposes.

Correlations of .1 to .9 of standardised scores (i.e. z scores) notice how the angle of the line gradually increases to 45%. If it were a negative correlation it would be -45% (reflection in the x axis). So a perfect correlation (1) has an angle of 45%

7. Common misconceptions about correlation

- Correlation does not imply causation
- Correlation is only to do with monotonic / linear associations
- Non-homogeneous Groups
- A significant p value does not prove the correlation

Just because two variables co-vary does not mean that one causes the other, however it can be said that correlation is a pre-condition for measuring causation. Read

http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation for further details

We have discussed the particular type of association that Pearson and rank correlations measure. Non linear relationships require special techniques to allow you to access them not discussed in this chapter.

Some samples represent a number of distinct groups, for example cholesterol levels may have been taken from both males and females and also of patients of different ages, it is possible that each subgroup will provide a different correlation, there is also a danger that the sample becomes so heterogeneous that the various associations cancel one another out resulting in what appears to be a very low correlation. Two techniques are available for removing the influence of possible intervening variables, partial correlation (see: http://www.psychwiki.com/wiki/What_is_a_partial_correlation%3F) and semi-partial (also called part) correlation which is part of multiple regression.

There are other important aspects but these are more closely related to the line of best fit and will be discussed in the chapter concerning simple regression.

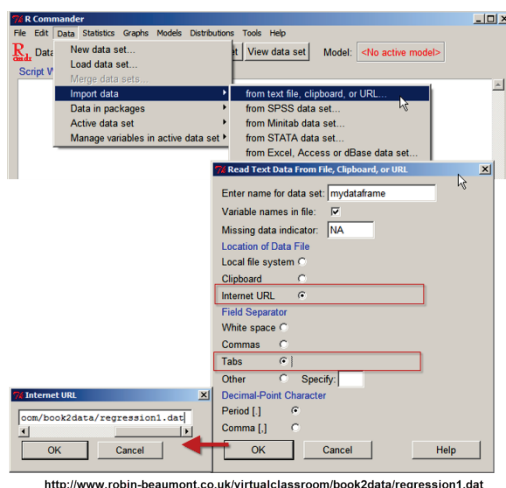
It is important to remember that the associated p value of a correlation is providing information concerning the likelihood of obtaining a correlation of the given value or one more extreme given that the correlation is assumed to be zero in the population. That is when we obtain a significant correlation (i.e. p value less than the critical value) all we are saying is that the population correlation is NOT zero we are **definitely not saying that the population value is equal to the observed correlation in our sample.**

8. Carrying out correlations

To demonstrate how you obtain correlations I have used two examples, the first of which is from Harris & Taylor 2008 page 55 which describes 12 HbA1c and fasting blood glucose measurements taken from 12 diabetic subjects. We will investigate the correlation between HbA1c and fasting blood glucose.

8.1 Do it in R Commander

Preliminaries



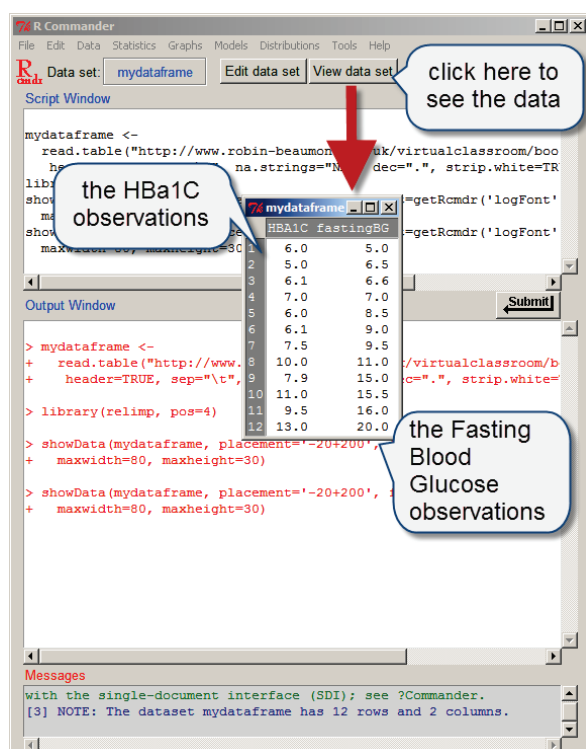
From within R you need to load R commander by typing in the following command:

```
library(Rcmdr)
```

You can obtain the data directly from my website which you do by selecting the R commander menu option:

Data-> from text, the clipboard or URL

I have given the resultant dataframe the name mydataframe, also indicating that it is from a URL (i.e. the web) and the columns are separated by tab characters.



Clicking on the OK button brings up the internet URL box, you need to type in it the following to obtain my sample data:

```
http://www.robin-beaumont.co.uk/virtualclassroom/book2data/regression1.dat
```

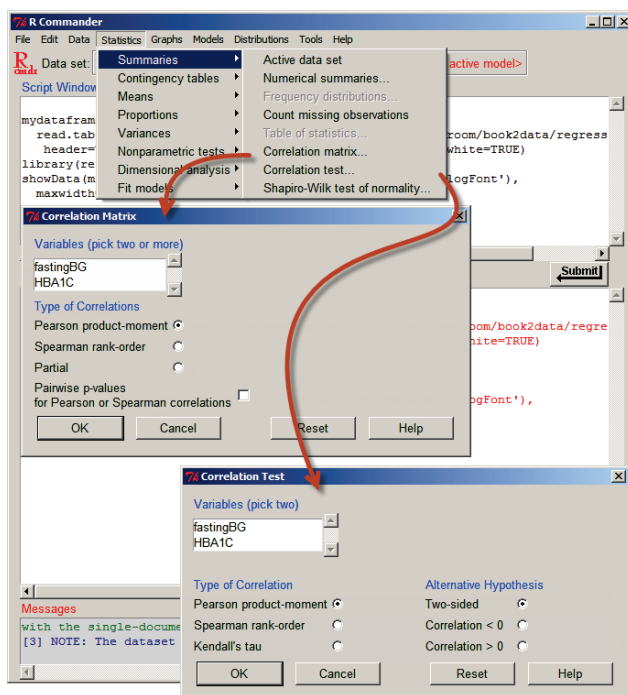
Click OK

It is always a good idea to check the data and you can achieve this easily in R commander by clicking on the View data set button.

We now have a dataframe with the necessary data in it we can request the correlation values along with their associated p-values.

Exercise 3.

Create a scatterplot of the above data using the appropriate variables and menu option in R Commander



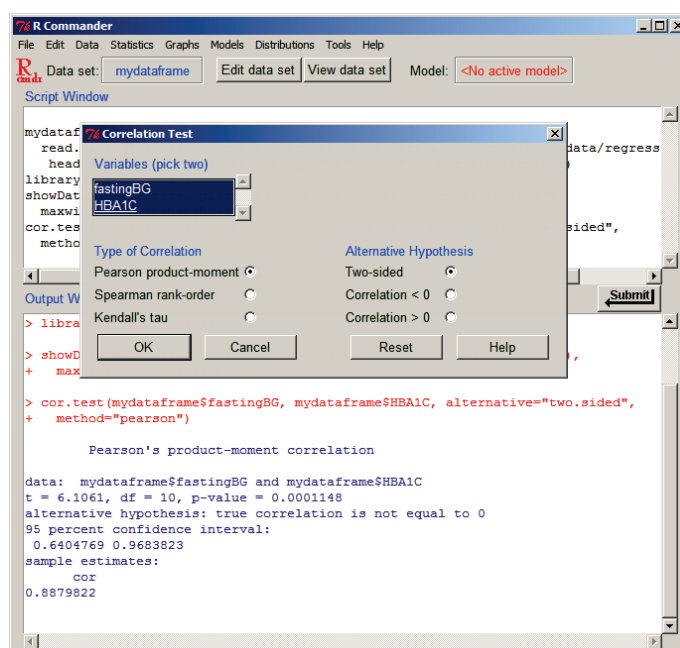
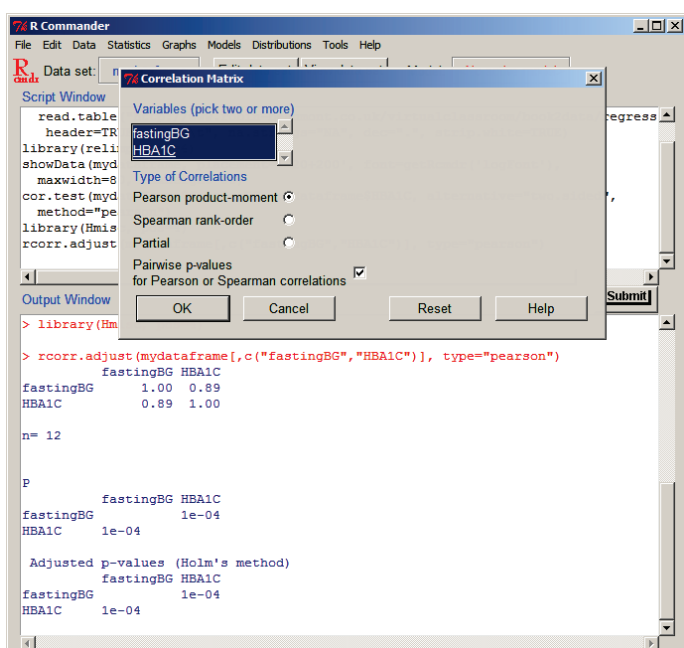
The correlation

R Commander has two menu options for obtaining correlations.

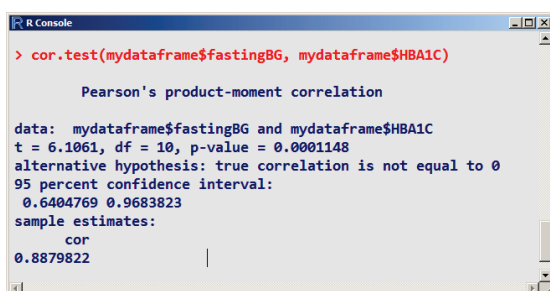
The correlation matrix option provides the correlation values (r) in the form of a table, the values reported to two decimal places (i.e. 0.89), along with a table of p-values underneath it. The 1e-04 is what is known as E notation (see glossary).

The 'adjusted p-value' is a p-value that takes into account the fact that the more correlations you have in the table to more likely you are to find significant p-values so adjusts them accordingly.

The Correlation test option provides more detail including a confidence interval, but has the downside that you can only investigate one pair at a time.



8.2 Doing it directly in R



If correlations within p-values we simply enter the following command in the R console window:

```
cor(mydataframe)
```

or we can specify a certain number of decimal places by

```
round(cor(mydataframe), 4) # will give r values to 4 decimal places
```

For a single correlation with a 95% CI and p-value enter:

```
cor.test(mydataframe$fastingBG, mydataframe$HBA1C)
```

8.3 Multiple scatterplots and correlations in R directly

Often journal articles have a series of small scatterplots and correlation values appearing together in a table like structure. You can achieve this in R.

To demonstrate this I will use a dataset from the DAAG package called `ais`, which I have extracted.

The data frame consists of 202 observations on Australian athletes for the following 13 variables. The data were collected to study how various characteristics of the blood varied with sport body size and sex of the athlete.

- `rcc` - red blood cell count, in $10^{12} l^{-1}$
- `wcc` - while blood cell count, in 10^{12} per liter
- `hc` - hematocrit, percent
- `hg` - hemoglobin concentration, in g per decaliter
- `ferr` - plasma ferritins, $ng dl^{-1}$
- `bmi` - Body mass index, $kg cm^{-2} 10^2$
- `ssf` - sum of skin folds
- `pcBfat` - percent Body fat
- `lbm` - lean body mass, kg
- `ht` - height, cm
- `wt` - weight, kg
- `sex` - a factor with levels f m
- `sport` - a factor with levels B_Ball Field Gym Netball Row Swim T_400m T_Sprnt Tennis W_Polo

Notice that the last two variables I have grayed out as they would make little sense in using for a correlation as they are factors.

To obtain the required plot I will first load the `psych` package:

```
library(psych)
```

Now load the required data from my website:

```
mydataframe <- read.delim("http://www.robin-beaumont.co.uk/virtualclassroom/book2data/ais_daag.dat", header=TRUE)
```

Get a printout of the names of the columns and attach the dataframe:

```
names(mydataframe)
attach(mydataframe)
```

Now use the `pairs.panels` command in the `psych` package to get the required plot

```
pairs.panels(mydataframe)
```

I have used the simplest approach here some options:

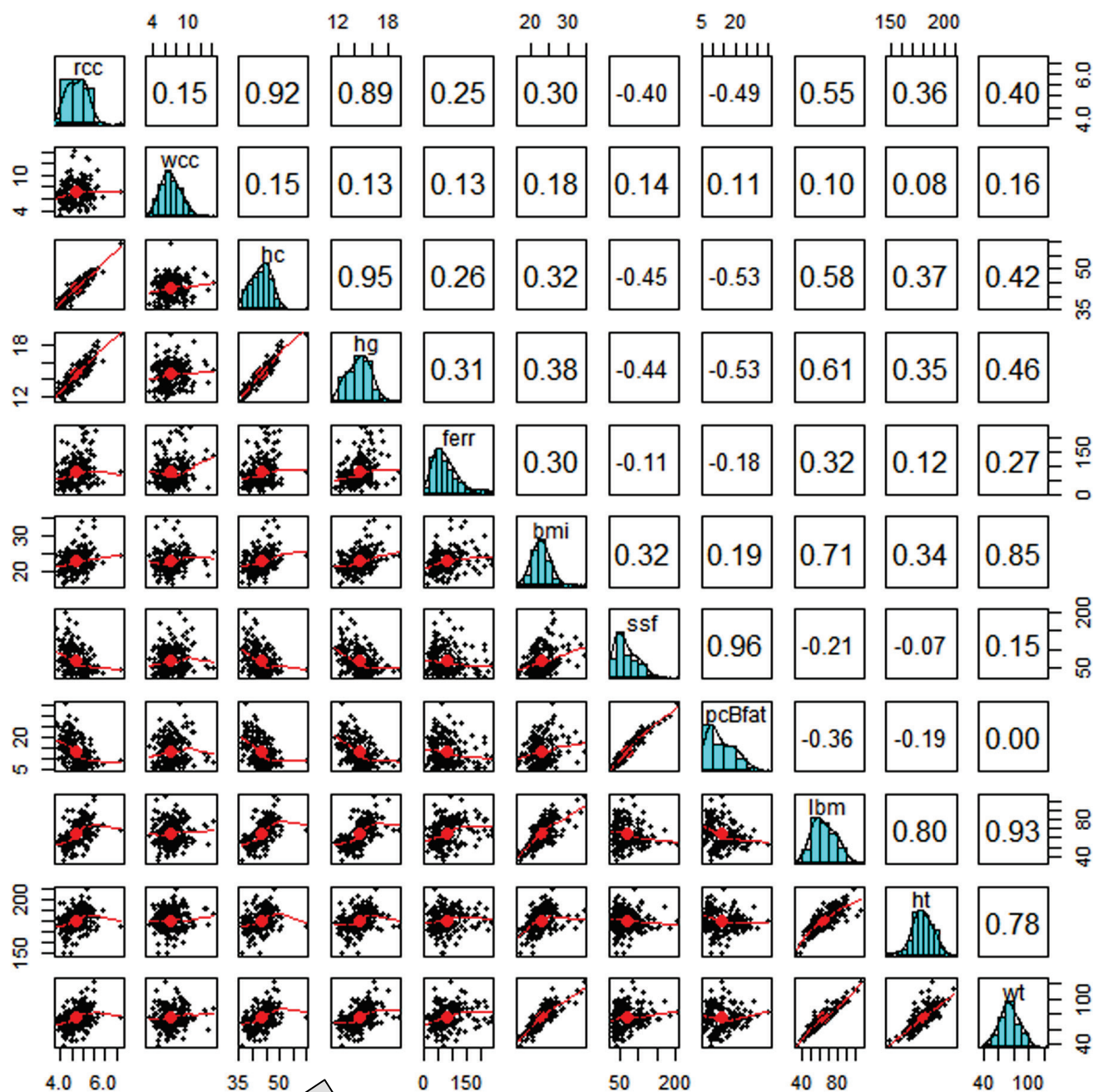
- `lm=True` ; produces straight linear regression lines.
- `ellipses=FALSE`; removes the red ellipses.
- `method="spearman"` or `"kendall"` produces other types of correlations

The last command produced a diagram which included the last two variables (columns 12 and 13) to omit those I can use the following command:

```
pairs.panels(mydataframe[1:11])
```

The result is shown on the following page:

Pearson correlations - larger equals more statistically significant



Scatterplots with a correlation ellipse and a LOESS line fit

name and histogram for each variable

My youtube video at the end of this section describes how you can obtain a result similar to the above in R commander.

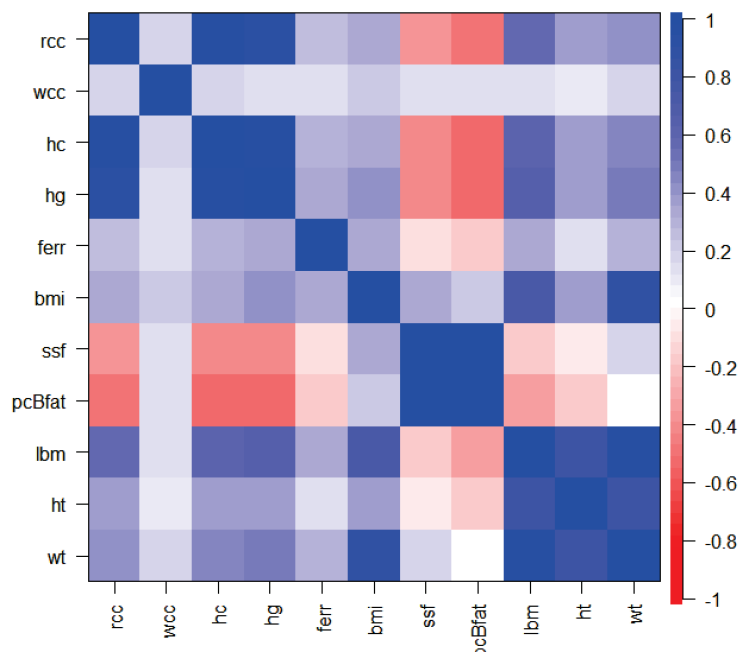
Another method of seeing a large number of correlations is to have a table where each of the cells within the table has a colour which reflects the strength of the correlation. This can easily be done using the psych package `cor.plot()` command, typing the following into the R console window, assuming you have the psych package installed and loaded, produces the output below.

```
cor.plot(cor(mydataframe[1:11]))
```

some options:

- `colors=FALSE`; produces gray scale.
- `n=51`; number of levels of shading, 51 is the default.
- `main="a title"`

Correlation plot



The key on the right hand side indicates that dark red or blue is a large negative or positive correlation. This graphical approach immediately highlights and clustering of correlations.

It also highlights the danger of correlations with variables which are just combinations of other variables for example someone who did not know that BMI is calculated from height and weight would be excited by the dark blue around these three variables in the bottom right hand corner.

8.4 Tips and Tricks

While most correlations are for continuous or ordinal data, you can produce special types of correlations for binary or nominal data.

Remember that the can produce multiple scatterplots like the one shown on the previous page using the R Commander menu options see my YouTube video below for information

YouTube help

[tube.com/watch?v=a9ndjEYAB10&feature=share&list=PL9F0EBD42C0AB37D0](https://www.youtube.com/watch?v=a9ndjEYAB10&feature=share&list=PL9F0EBD42C0AB37D0)

8.5 Entering data and running correlations from within R for small datasets

For a small dataset you can do everything in a few lines in R.

```
R Console
File Edit Misc Packages Windows Help

> x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)
> y <- c( 2.6, 3.1, 2.5, 5.0, 3.6, 4.0, 5.2, 2.8, 3.8)
> ## example of correlation
> ##
> cor.test(x, y, method = "pearson", alternative = "two.sided", conf.level = 0.95)

Pearson's product-moment correlation

data:  x and y
t = 1.8411, df = 7, p-value = 0.1082
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.1497426  0.8955795
sample estimates:
cor
0.5711816
```

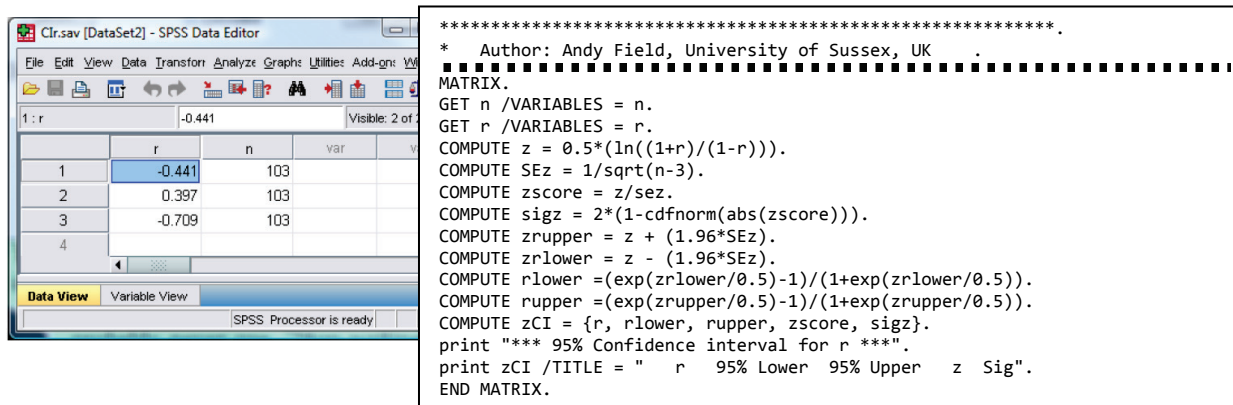
R code for pearson correlation coefficient and p value and CI

```
x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)
y <- c( 2.6, 3.1, 2.5, 5.0, 3.6, 4.0, 5.2, 2.8, 3.8)
## example of correlation
cor.test(x, y, method = "pearson", alternative = "two.sided", conf.level = 0.95)
```

8.6 SPSS

The Youtube video on the previous pages demonstrates how to obtain scatterplots and correlations in SPSS.

SPSS can be encouraged to produce confidence intervals for correlations by using a few simple commands. Andy Fields excellent introductory statistics book (Field 2009, p.173) provides some details with step by step details at http://www.uk.sagepub.com/field3e/Downloads/addmaterial/DSUSaddweb_ch06.pdf



8.7 Obtaining rank order correlations

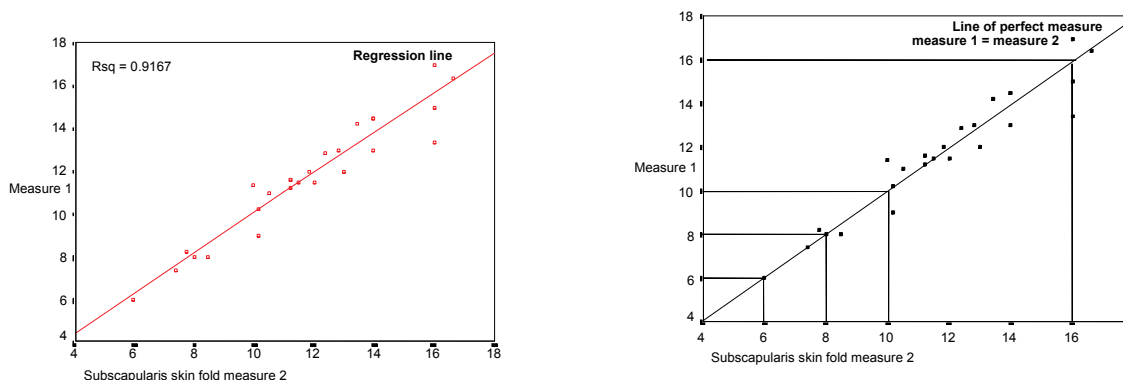
I mentioned earlier that non parametric correlations were useful for monotonic functions. In r you just substitute the 'pearson' term for either spearman or kendall:

```
cor.test(x, y, method = "kendall", exact = TRUE)
cor.test(x, y, method = "spearman")
```

SPSS provides all three statistics but no confidence intervals.

9. Reliability and r / r^2

Instead of using two different variables it is possible to consider the same variable measured at two separate times. For example suppose we are interested in investigating the reliability of taking skin fat measurements. We devise an experiment where 40 students each have their measurements taken twice by two separate skin calliper operators. The results are given below:



We can say therefore that 91% of the variation in measuring subscapularis skin fold measurements is accounted for by the regression line. However we cannot consider this regression line as the line of 'perfect

measure' as this would be through the point where $x_i=y_i$ for every value i.e. $x=4$ when $y=4$ and $x=8$ when $y=8$ etc. It appears from this scatter plot that people tend to become more inaccurate as the skin fold measurements increase. For details of how to analysis this data correctly see Bland & Altman 1996, their excellent article (except for the error with the formula and the table data).

Taking into account the fact that in this instance it is really irrelevant which value of a particular pair of x y values is the x or y value a special type of correlation has been developed, the **Intraclass correlation**, unfortunately there are several varieties and I would advice those who are interested to consult (Shrout & Fleiss, 1979; Nichols 1998 or McGraw & Wong 1996) A good introductory article is the Bland and Altman mentioned above and I have used the data (the correct annotated version of it) from their article to demonstrate the intraclass correlation in SPSS using the menu option Analyse->Scale->Reliability

I will not discuss the various options (you can also request the ANOVA table mentioned in the Bland and Altman article)– there are many and this is just a taster, but notice that SPSS produces a slightly higher value (.76) than that quoted in the article (.75)

Intraclass Correlation Coefficient

	Intraclass Correlation ^a	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.764 ^b	.494	.899	7.457	19	19	.000
Average Measures	.866 ^c	.661	.947	7.457	19	19	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Data from Bland and Altman 1996
Pairs of measurements of FEV1 (litres)
a few weeks apart
from 20 Scottish schoolchildren, taken
from a larger study (D Strachan,
personal communication)

Subject number	First reading	Second reading
1	1.20	1.24
2	1.21	1.19
3	1.42	1.83
4	1.43	1.38
5	1.49	1.60
6	1.58	1.36
7	1.58	1.65
8	1.59	1.60
9	1.60	1.58
10	1.61	1.61
11	1.62	1.68
12	1.64	1.61
13	1.65	2.05
14	1.74	1.80
15	1.78	1.76
16	1.80	1.76
17	1.80	1.82
18	1.85	1.73
19	1.88	1.82
20	1.92	2.00

Trick

Because the ordering of the x y values is immaterial you can also obtain an intraclass correlation simply by appending the x values to the y column and appending the y values to the x column and then perform a standard Pearson correlation on the new dataset. Doing this produces a value of .752 basically equivalent to the value produced by Bland and Altman.

Bland and Altman developed a special plot to show the level of agreement for a interval/ratio level variable. called a Bland Altman plot see:

http://en.wikipedia.org/wiki/Bland%20%80%93Altman_plot

You can draw these in R by using theBA.plot method in the MethComp package.

10. Multiple Choice Questions

1. Correlation is a measure that makes use of a particular distribution, what is it? (one correct choice)

- a. Normal
- b. Exponential
- c. Chi square (df=1)
- d. Bivariate normal
- e. Uniform

2. Correlation is often assessed by eye, which type of plot is usually used for this purpose? (one correct choice)

- a. Histogram
- b. Bar chart
- c. Boxplot
- d. Scatter plot
- e. Funnel plot

3. Which of the following statements is true concerning correlation? (one correct choice)

- a. A correlation is always between -2 and 2, a zero value indicates no clustering towards line
- b. A correlation is always between -1 and 1, a zero value indicates all points on line
- c. A correlation is always between -2 and 2, a zero value indicates all points on line
- d. A correlation is always between -1 and 1, a zero value indicates no clustering towards line
- e. A correlation is always between -1 and 1, a zero value indicates all points on a horizontal line

4. The correlation coefficient is based upon another measure, what is it? (one correct choice)

- a. Variance
- b. Co-relation
- c. Contingency coefficient
- d. Covariance
- e. Cooks distance

5. The calculation of the confidence interval for the correlation coefficient is . . . ? (one correct choice)

- a. No different from other statistics
- b. More complex than usual because of the restricted range
- c. Needs to be interpreted with extreme caution
- d. Un-defined
- e. Equivalent to the coefficient of determination

6. There are a number of effect size measures for the correlation coefficient. Which of the following is **not** considered to be one? (one correct choice)

- a. Coefficient of determination (r^2)
- b. Cohens d
- c. Correlation coefficient
- d. Cooks distance
- e. Correlation coefficient squared

7. The coefficient of determination can be interpreted a number of ways. Which of the following is one of them? (one correct choice)

- a. Proportion of explained variation
- b. Proportion of unexplained variation (i.e. residual)
- c. Proportion of mean variation
- d. Proportion of variance variation
- e. Proportion of points on the line

8. There is a special variety of the correlation coefficient used in the situation where the x and y values are interchangeable such as when comparing two measures, this intraclass correlation can be calculated easily by? (one correct choice)

- a. Appending the y scores to the x scores and then performing a standard correlation.
- b. Appending the y scores to the x scores and then performing a rank correlation
- c. Appending the y scores to the x scores and appending the x scores to the y ones then performing a standard correlation.
- d. Appending the y scores to the x scores and appending the x scores to the y ones then performing a rank correlation.
- e. Appending the y scores to the x scores and appending the x scores to the y ones then performing a paired *t* statistic .

9. Which is the most important assumption that is relaxed when considering Rank correlation compared to those for the Pearson correlation coefficient? (one correct choice)

- a. Linear relationship
- b. Normal distribution
- c. Observation pairs are independent
- d. Sample is randomly selected
- e. Data cannot be nominal

10. Which of the following statements concerning the correlation coefficient is **not** correct? (one correct choice)

- a. Correlation does not imply causation
- b. Usual correlation techniques only consider monotonic/linear associations
- c. Non-homogenous groups can affect the correlation
- d. A significant *p* value provides evidence that the population correlation is equal to that observed
- e. Correlation was originally developed by Sir Francis Galton

11. Which of the following provides the most accurate interpretation of a Pearson correlation coefficient of .733 ($p=.0001$)? (one correct choice)

- a. We are likely to observe a correlation of .733 given that the population correlation is equal to .773 around once in ten thousand times on average in the long run.
- b. We are likely to observe a correlation of .733 or one more extreme given that the population correlation is **not** equal to zero around once in ten thousand times on average in the long run.
- c. We are likely to observe a correlation of .733 or one more extreme given that the population correlation is equal to zero around once in a hundred times on average in the long run.
- d. We are likely to observe a correlation of .0001 or one more extreme given that the population correlation is equal to .733 in the long run.
- e. We are likely to observe a correlation of .733 or one more extreme given that the population correlation is equal to zero around once in ten thousand times on average in the long run.

11. Summary

This chapter has taken a detailed look at correlation, seeing how it is based on the covariance. We then considered its measurement and what the associated p value meant. Also aspects such as the associated effect size and confidence intervals were discussed. A range of scatter diagrams were given for both raw and standardized datasets demonstrating the complete range of correlations. The final section discussed some of the common misconceptions concerning correlation and suggested how it is the basis for a wide range of other statistical methods.

12. References

Bland JM, Altman DG. 1996 Measurement error and correlation coefficients. *BMJ* ;313:41-2. Available from: <http://ccg.cochrane.org/basic-statistics>

Field A 2009 *Discovering Statistics Using SPSS*. Sage

Fisher R A 1915 Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, Vol. 10, No. 4 (May, 1915), pp. 507-521

Harris M, Taylor G 2008 (2nd ed) *Medical Statistics Made Easy*. Scion.

Howell D C 2006 *Statistical Methods for Psychology (Ise) (Paperback)*

Marques de Sá JP 2007 *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*. Springer

McGraw KO, Wong SP. 1996 Forming inferences about some intraclass correlation coefficients [published erratum appears in *Psychol Methods* 1:390]. *Psychol Methods* 1996;1:30-46.

Nichols DP 1998 Choosing An Intraclass Correlation Coefficient From SPSS Keywords, Number 67, Available online at: <http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/articles/whichicc.htm> [accessed 21/04/2010 08:30]

Pagano RR 1980 (3rd ed.) *Understanding statistics* West publishing

Rodgers J L, Nicewander W A 1988 Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66

Rovine M J. von Eye A 1997 A 14th Way to Look at a Correlation Coefficient: Correlation as the Proportion of Matches. *The American Statistician*, Vol. 51, No. 1 (Feb., 1997), 42-46 Stable URL: <http://www.jstor.org/stable/2684692>

Shrout PE Fleiss JL. 1979 Intraclass correlations: Uses in assessing reliability. *Psychol Bull* 86: 420-428

Snedecor G Cochran W 1989 (8th ed.) *Statistical Methods*. Blackwell. [Most of this book is available in Google books]

13. Appendix r code

```

library(ecodist) #library to produce sample of a given correlation
#code to produce regression lines of both x on y and y on x
# Also: allows you to select, sample size, mean,sd, required correlation,
# be removing '#' comment symbol in lines 14-17 allows plotting of standardised values
# gives parameters for both y on x and x on y and the latter in terms of y as well
#highlights the point mean(x), mean(y)
#
# corgen(len, x, r, population = FALSE, epsilon = 0)
# return a vector of length n with required correlation r plus/minus epsilon

  x <- rnorm(100,mean=10,sd=2)
y <- 20 + corgen(x=x, r=.01, epsilon=.001)$y #the dollar sign means you only get the y back
# 4 lines below standardize if min < zero then update y data to yi=yi+|min| + delta
#remove comment from next 4 lines if you want to standardised scores
# x<-(x-mean(x))/sd(x)
# y<-(y-mean(y))/sd(y)
# if (min(y)<0) { y<-y+abs(min(y)) + 1}
# if (min(x)<0) { x<-x+abs(min(x)) + 1}

## First create the standard regression line - y on x = y=mx+c
modela<-lm(y~x) #intercept slope
modpar<-coef(modela) #put the model parameters into a vector
#can manipulate the standard equation several ways:
#also y=mean(y)= gradient(x-mean(x)) #above shows gradient and intercept
# y=(gradient(x - mean(x)) + mean(y)
# y=(mean(y) - gradient*(mean(x)) + gradient*x # before + is all constant= intercept
# after + is b =gradient
#above model is for y on x now need x on y then need it arranged in terms of the y on x form
#x=gradient*y+intercept # 'x on y'
# x=(mean(x) - gradient*(mean(y)) + gradient*y #swapping around x and y's in above
xonygrad<-cov(x,y)/var(y) #var(y) has replaced var(x)
xonyinter<--mean(y)*(xonygrad)+mean(x) #not used see below
#but now need to rotate axis
gradxonyrot<-var(y)/cov(x,y)
xonyinterrot<-mean(y)-(gradxonyrot*mean(x))
#equation below produces slightly inaccurate line?
#xonyinterrot<-xonyinter/gradxonyrot
plot(x,y,type="p",ylab="y",xlab="x",main="Regression lines r=.01 (raw scores)")
# abline takes intercept ,slop parameters
abline(modpar[1],modpar[2]) #intercept, slope
abline(xonyinterrot[1],gradxonyrot[1],col="red",lty=5) #type 5 is dotted
#if using standardised scores use horizontal line below instead of abline above
# abline(h=0) # horizontal line
points(mean(x),mean(y),pch=16) #pch=16 gives solid circle; mean of x, y
# if you have the car package installed you can also use ellipse(x,y,levels =.95,col="black")
mean(x); mean(y) # now some stats
modpar[1] # intercept term for y on x = standard regression
modpar[2] #gradient term for y on x = standard regression
xonygrad # gradiate for x=gradient*y + c
xonyinter # intercept (c) for x=gradient*y+c
gradxonyrot # gradient for x= in terms of y=
xonyinterrot # intercept for x= in terms of y=
cor(x,y) # correlation #to check we got what we asked for!
# completed# completed

```

```

#can substitute the above artificial distribution for real data
#data from snedecor & Cochran 1980, page 176 brother and sisters height
# from pearson & Lee 1902 Biometrika

x <- c(71,68,66,67,70,71,70,73,72,65,66)
y <- c(69,64,65,63,65,62,65,64,66,59,62)
## etc

# Important note saving graphs from word to pdf - you need to 'paste special' into word selecting:
# Picture (windows metafile) this allows you to edit out the additional line that appears some times
# with the abline command that does not appear within word but only when you pdf it!!

```

```
#code to produce regression lines of both x on y and y on x and ellipse
#
library(ecodist) #library to produce sample of a given correlation
library(car) #library for the ellipse function
# corgen(len, x, r, population = FALSE, epsilon = 0)
# return a vector of length n with required correlation r plus/minus epsilon
x <- rnorm(100, mean=10, sd=2)
y <- 20 + corgen(x=x, r=.6, epsilon=.001)$y #the dollar sign means you only get the y back
## First create the standard regression line - y on x = y=mx+c
modela<-lm(y~x) #intercept slope
modpar<-coef(modela) #put the model parameters into a vector
xonygrad<-cov(x,y)/var(y) #var(y) has replaced var(x)
xonyinter<--mean(y)*(xonygrad)+mean(x) #not used see below
#but now need to rotate axis
gradxonyrot<-var(y)/cov(x,y)
xonyinterrot<-mean(y)-(gradxonyrot*mean(x))
data.ellipse(x,y[,1],lty=3,col="red")
# abline takes intercept ,slop parameters
abline(modpar[1],modpar[2],col="black") #intercept, slope
abline(xonyinterrot[1],gradxonyrot[1],col="green",lty=5) #type 5 is dotted
#if using standardised scores use horizontal line below instead of abline above
# abline(h=0) # horizontal line
points(mean(x),mean(y),pch=16) #pch=16 gives solid circle; mean of x, y
# completed
```

```
#####
#####
##### r code to use for next revision
source code of the bothlines function in the MethComp package
Plots both x on y and y on X
```

```
bothlines
function (x, y, Dem = FALSE, sdr = 1, col = "black", ...)
{
  clr <- rep(col, 3)
  if (class(x) == "lm") {
    y <- x$model[[1]]
    x <- x$model[[2]]
  }
  abline(lm(y ~ x), col = clr[1], ...)
  ic <- coef(lm(x ~ y))
  abline(-ic[1]/ic[2], 1/ic[2], col = clr[2], ...)
  if (Dem) {
    Dm <- Deming(x, y, sdr = sdr)
    abline(Dm[1], Dm[2], col = clr[3], ...)
  }
}
```