chapter 2-5

# Correlation

## Introduction

One of the most frequently asked questions in inferential statistics is whether or not there is a relationship between two variables. When data are normally distributed, the linear relationship between two variables can be described by the Pearson Product Moment Correlation Coefficient (r). Few people have trouble understanding the concept of the correlation coefficient; however, it is probably also the most misused and misinterpreted of all statistics. The purpose of this chapter is to illustrate the meaning of the correlation coefficient and how it can be derived using EXCEL or SPSS. Importantly, this chapter will also point out the possible pitfalls in the use of the correlation coefficient.

## Correlation Coefficient (r)

The Pearson Product Moment correlation coefficient (r) assesses the degree of **linear association** between two variables. The coefficient can vary from -1 through 0 to +1. Figure 2-5.1 depicts various values for the correlation coefficient according to scatterplots of the data. A perfect straight line relationship is present when the correlation coefficient is 1. The + or - sign merely indicates the direction of the slope. If the correlation coefficient is positive, then as one variable gets bigger, so too does the other; if it is negative, then as one gets bigger the other gets smaller, as illustrated by the r = +1 and r = -1 graphs. The linear correlation coefficient is a ratio of the variability in Y relative to X when the best fitting straight line is determined. A freehand ellipse has been drawn around the boundaries of the data points in the other graphs, in order to show the shape of data expected for any given value of r. It should be noted than an ellipse is not the expected shape of the data since the variability in Y is expected to be the same for all values of X, but it does provide a convenient way of displaying differences in shape of the distributions and therefore the ratio of variances typical of various correlation coefficients. It can be seen that the higher the correlation coefficient the slimmer the associated ellipse.

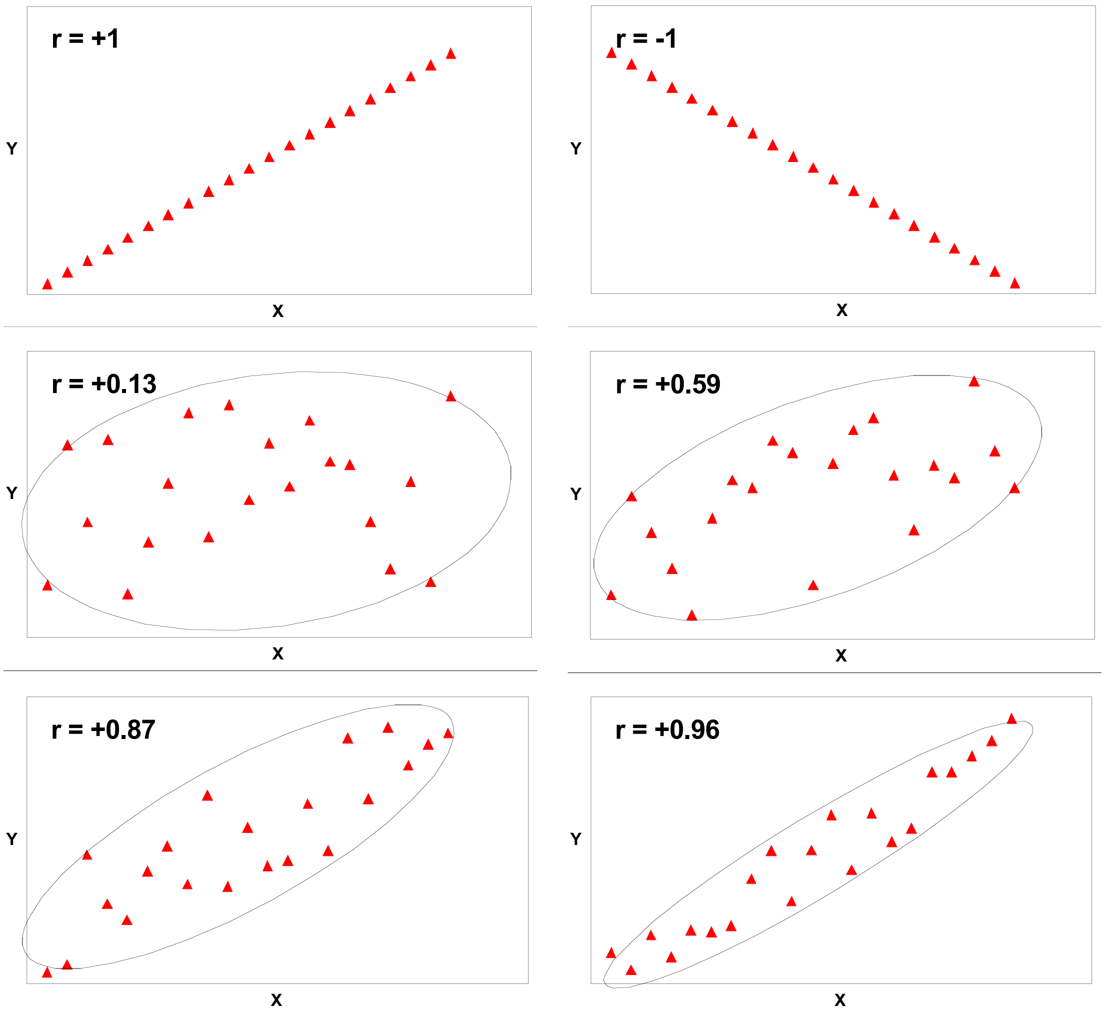


**Figure 2-5.1: Illustration of various values of the correlation coefficient**

**Limited to a Linear Fit:** Figure 2-5.2 shows a plot of data where there was a statistically significant correlation coefficient of 0.906 ($p < 0.05$) between two variables. An r =.906 seems high, but a quick look at the plot clearly shows that a straight line is not the best fit to the data. A curvilinear relationship would fit better. You should always plot out your data and look for nonlinearity, or test for a better fit with other nonlinear equations. We will be discussing nonlinear curve fitting in chapter 2-7 on modeling.

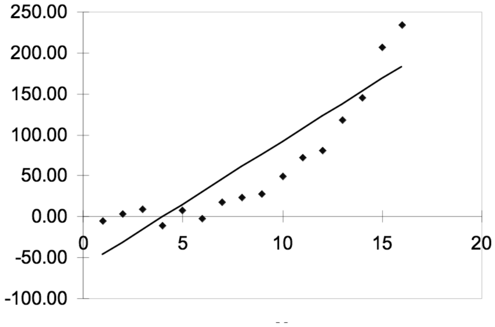


**Figure 2-5.2: Scatterplot of data where there was a statistically significant correlation coefficient of 0.906 ($p < 0.05$)**

## Calculation of the Correlation Coefficient (r)

Below is the equation to calculate the correlation coefficient, which is essentially a ratio of variances. This is typically a function on most calculators and is certainly included in all types of statistical analysis software, so you would rarely have to calculate it by hand.

$$r = \frac{\sum x_1 x_2}{\sqrt{(\sum x_1^{\ 2})(\sum x_2^{\ 2})}} \quad \text{where } x_1 = X_1 - \overline{X_1} \quad \text{and } x_2 = X_2 - \overline{X_2}$$

| Right Hand Length (cm) $X_1$ | Left Hand Length (cm) $X_2$ | $x_1 = X_1 - \overline{X_1}$ | $x_2 = X_2 - \overline{X_2}$ | $x_1^{\ 2}$ | $x_2^{\ 2}$ | $x_1 x_2$ |
|---|---|---|---|---|---|---|
| **18.6** | **17.9** | -0.22 | -0.86 | 0.05 | 0.73 | 0.19 |
| **17.9** | **17.3** | -0.92 | -1.46 | 0.84 | 2.12 | 1.34 |
| **19.4** | **19.4** | 0.58 | 0.64 | 0.34 | 0.41 | 0.37 |
| **18.1** | **18.6** | -0.72 | -0.16 | 0.52 | 0.02 | 0.11 |
| **17.3** | **17.1** | -1.52 | -1.66 | 2.31 | 2.75 | 2.52 |
| **17.6** | **17.8** | -1.22 | -0.96 | 1.49 | 0.92 | 1.17 |
| **18.4** | **18.2** | -0.42 | -0.56 | 0.18 | 0.31 | 0.23 |
| **17.7** | **17.7** | -1.12 | -1.06 | 1.25 | 1.12 | 1.18 |
| **20.1** | **20.2** | 1.28 | 1.44 | 1.64 | 2.08 | 1.85 |
| **20** | **19.5** | 1.18 | 0.74 | 1.39 | 0.55 | 0.88 |
| **18.1** | **18.3** | -0.72 | -0.46 | 0.52 | 0.21 | 0.33 |
| **20.7** | **20** | 1.88 | 1.24 | 3.54 | 1.54 | 2.34 |
| **19.4** | **19.8** | 0.58 | 1.04 | 0.34 | 1.09 | 0.61 |
| **19.6** | **19.1** | 0.78 | 0.34 | 0.61 | 0.12 | 0.27 |
| **20.8** | **20.7** | 1.98 | 1.94 | 3.92 | 3.77 | 3.85 |
| **18.5** | **18.8** | -0.32 | 0.04 | 0.10 | 0.00 | -0.01 |
| **17.3** | **17.5** | -1.52 | -1.26 | 2.31 | 1.58 | 1.91 |
| **18.2** | **18.4** | -0.62 | -0.36 | 0.38 | 0.13 | 0.22 |
| **18.7** | **19** | -0.12 | 0.24 | 0.01 | 0.06 | -0.03 |
| **19.8** | **19.7** | 0.98 | 0.94 | 0.96 | 0.89 | 0.92 |
| **19** | **18.9** | 0.18 | 0.14 | 0.03 | 0.02 | 0.03 |
| $\overline{X_1}$ | $\overline{X_2}$ | | | $\sum x_1^{\ 2}$ | $\sum x_2^{\ 2}$ | $\sum x_1 x_2$ |
| **18.82** | **18.76** | | | **22.73** | **20.43** | **20.27** |

$$r = \frac{\sum x_1 x_2}{\sqrt{(\sum x_1^{\ 2})(\sum x_2^{\ 2})}} = \frac{20.27}{\sqrt{22.73 \times 20.43}} = 0.94$$

**Table 2-5.1: Calculation of the Correlation Coefficient of Right Hand Length versus Left Hand Length in University Men (n = 21)**

## Significance of the Correlation Coefficient (r)

While it is understood that the correlation coefficient is a measure of the degree of association of two variables, the question is, how confident are we that the degree of association in the sample reflects that in the population from which the sample was drawn? Just as with all inferential statistics, the correlation coefficient has an associated probability distribution. The statistical significance of the correlation coefficient is determined by the sample size (**n**) and the preset level of acceptance (***p***). Table 2-5.2 shows the critical values of the correlation coefficient for two levels of acceptance ***p*** = 0.05 (95%) and ***p*** = 0.01 (99%). The degrees of freedom are calculated as **n** – 1. Obviously a higher correlation coefficient is needed for statistical significance at **p** = 0.01 than ***p*** = 0.05. If we refer back to Figure 2-5.1, it can be seen that there are 21 data points in each graph. This gives us n – 1 = 20 degrees of freedom and hence critical values of r of 0.423 (p = 0.05), and 0.537 (p = 0.01). If we had preset our acceptance level at ***p*** = 0.01, then all of the relationships shown in figure 2-5.1 would have been statistically significant except for r = +0.13. However, if the sample size had been 401 (therefore degrees of freedom = 400), then the +0.13 would have been a statistically significant value for r since the critical value is 0.128 for 400 degrees of freedom at ***p*** = 0.01.

| **Degrees of Freedom** | **Probability** | | **Degrees of Freedom** | **Probability** | |
|---|---|---|---|---|---|
| | **0.05** | **0.01** | | **0.05** | **0.01** |
| **1** | .997 | 1.000 | **24** | .388 | .496 |
| **2** | .950 | .990 | **25** | .381 | .487 |
| **3** | .878 | .959 | **26** | .374 | .478 |
| **4** | .811 | .917 | **27** | .367 | .470 |
| **5** | .754 | .874 | **28** | .361 | .463 |
| **6** | .707 | .834 | **29** | .355 | .456 |
| **7** | .666 | .798 | **30** | .349 | .449 |
| **8** | .632 | .765 | **35** | .325 | .418 |
| **9** | .602 | .735 | **40** | .304 | .393 |
| **10** | .576 | .708 | **45** | .288 | .372 |
| **11** | .553 | .684 | **50** | .273 | .354 |
| **12** | .532 | .661 | **60** | .250 | .325 |
| **13** | .514 | .641 | **70** | .232 | .302 |
| **14** | .497 | .623 | **80** | .217 | .283 |
| **15** | .482 | .606 | **90** | .205 | .267 |
| **16** | .468 | .590 | **100** | .195 | .254 |
| **17** | .456 | .575 | **125** | .174 | .228 |
| **18** | .444 | .561 | **150** | .159 | .208 |
| **19** | .433 | .549 | **200** | .138 | .181 |
| **20** | .423 | .537 | **300** | .113 | .148 |
| **21** | .413 | .526 | **400** | .098 | .128 |
| **22** | .404 | .515 | **500** | .088 | .115 |
| **23** | .396 | .505 | **1,000** | .062 | .081 |

**Table 2-5.2: Critical Values of the Correlation Coefficient**

This highlights the difference between statistically significance and practical significance discussed in chapter 2-4. Although +0.13 is a weak relationship, it is statistically significant, meaning that if we kept sampling from the population 99 times out of 100 we would expect to see this degree of association in the data. Note that a statistically significant correlation coefficient does not mean a strong relationship between variables, merely that we are confident enough that this degree of relationship exists in the population.

## Range of the Data affects the Correlation Coefficient

Figure 2-5.3 shows data for university men (n =20) and women (n = 23) for maximum grip strength versus skinfold-adjusted forearm girth. The correlation coefficient r for men was found to be +0.78 and for women was +0.75. When the two groups were combined to calculate a new correlation coefficient a value of +0.91 was found.

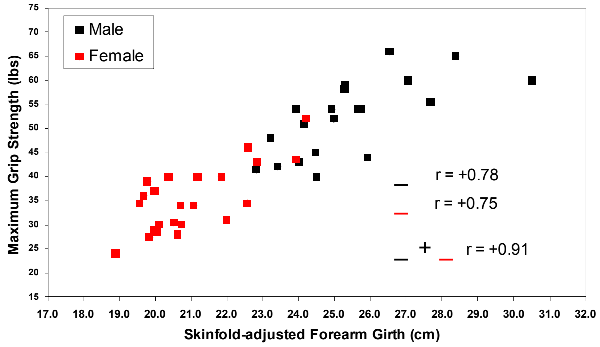


**Figure 2-5.3: Maximum Grip Strength vs Skinfold-adjusted Arm Girth. University men (n=22) and women (n=21)**

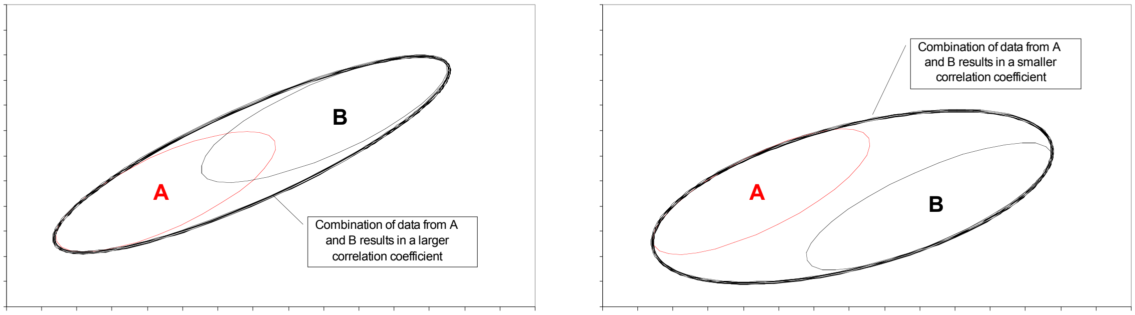
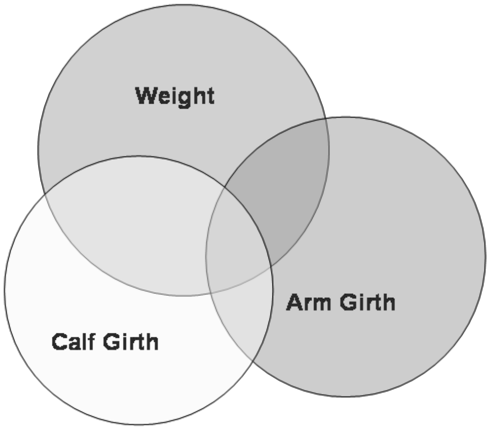

A
B
Combination of data from A and B results in a larger correlation coefficient
Combination of data from A and B results in a smaller correlation coefficient
A
B

**Figure 2-5.4: Diagramatic representation of correlation coefficients resultant from a combination of data from two groups A and B.**

Figure 2-5.3 shows the data for this analysis. Since men are bigger and stronger their data points tend to be higher on both the X and Y axes. If you eyeball the best line fits through the men and women's data respectively you will see that they are very similar in slope and in fact a single line could fit through them. This results in the r value increasing in the combined data. The Y axis variability is similar in the new group but the X axis variability is increased due to the smallness of the females and the greater size of the males. This is not always the case when two groups are combined. Figure 2-5.4 illustrates this situation where the r increases after combination (left chart) and another scenario where the r would decline in comparison to the r values for the two groups prior to combination. This is because the groups are not nicely aligned. There are no quick and easy rules for combination of r values. This stresses the need to always visualize your data by plotting such charts.

## Coefficient of Determination ($r^2$)

A useful statistic is the Coefficient of Determination ($r^2$), which is merely the correlation coefficient squared. The $r^2$ value quantifies the **proportion of the variance in one variable explained by the other**. Figure 2-5.5 is a Venn diagram illustrating this concept. The circles can be thought of as depicting the total variance (variability) in each of the three variables, weight, arm girth and calf girth within the sample.



| | | |
|---|---|---|
| Weight<br>vs<br>Arm Girth | r = 0.5 | $r^2$ = 0.25 |
| Weight<br>vs<br>Calf Girth | r = 0.6 | $r^2$ = 0 .36 |
| Arm Girth<br>vs<br>Calf Girth | r = 0.4 | $r^2$ = 0 .16 |

**Figure 2-5.5: Venn diagram illustrating proportions of variance explained by each variable**

The overlap of the circles represents the proportion of the variance of one explained by the other variable, this can be turned into a percentage explained variance by multiplying by 100:

**% Explained Variance = 100 x $r^2$**

The other portion of the variance is referred to as the error or residual variance:

**% Error (Residual) Variance = (100 – (100 x $r^2$))**

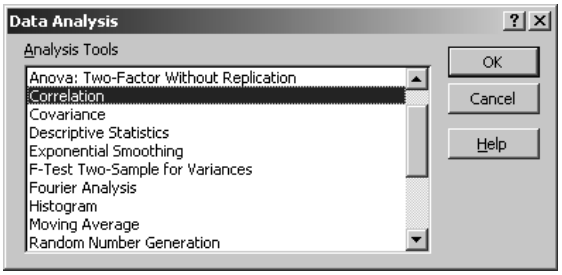
By example, the correlation coefficient between weight and arm girth is 0.5, $r^2$ is therefore 0.25. If you multiply this by 100 you get 25%. We therefore interpret $r^2$ = 0.25 as 25% of the variance in weight is explained by arm girth (or similarly, 25% of the variance in arm girth is explained by weight). Hence, a 25% overlap in the circles for weight and arm girth. The unexplained or residual variance is 100% – 25% or 75%. Calf girth has a 0.6 correlation with weight; therefore, 36% of the variance in weight is explained by this variable and the residual variance is 64%. It should be noted that the 25% of weight variance explained by arm girth is not totally contained within the variance explained by arm girth. The reason for this is that arm girth and calf girth are not perfectly correlated themselves, but have only a 0.4 correlation therefore only 16% of common variance explained and 84% residual variance. We will discuss this concept again in chapter 2-5 when discussing multiple regression equations.

## Correlation Using MS EXCEL

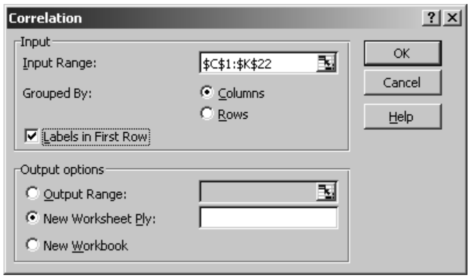
The calculation of the correlation coefficient performed earlier in this chapter was to allow the reader to have a better understanding of the meaning of r. The active researcher typically uses some form of computer calculation. In MS EXCEL the correlation coefficient can be calculated within a cell using the function =CORREL(), where the cell addresses of the two columns of data to be correlated are typed within the parentheses.

Correlation coefficients can also be calculated using the DATA ANALYSIS package listed in the TOOLS menu. Figure 2-5.6 shows the set up and results of using this on the grip strength data. When selected, the DATA ANALYSIS dialog box will be displayed and you can scroll down to find the Correlation item. Click OK and the Correlation dialog box will be displayed. The input range is the cell addresses of the two columns of data. If you have labels as the first entry in the column this selection needs to be checked off. The output can go to either the same sheet or a new sheet based upon your selection. In the example shown, nine columns of data with 21 men in each were selected. The output shown in Figure 2-5.6 illustrates what is known as a correlation matrix, with each variable correlated with every other variable. Only the half below the diagonal has values since the half above the diagonal would merely be a mirror image of the lower half. The diagonal is always composed of 1s, since this represents the correlation of a variable with itself, which by definition would be perfect, or r=1. Note the number of decimal places given for r. In reporting these values, be sure to round them down to an appropriate number of significant decimal places (1 or 2 is usually appropriate).

A correlation matrix is often a very valuable first look at a multivariable scenario. It can reveal a very low r between two variables that you might ordinarily expect to be correlated with each other. These low values can often be due to one or more extreme values in your data, thus highlighting you to possible errors in data entry. A correlation matrix is also a valuable tool for selecting variables for multivariable analysis and to avoid co-linearity. A multivariable analysis is adversely affected by independent variables that are highly correlated (see Chap 2-6).



**Data Analysis – Test selection dialog box**



Input
Input Range: $C$1:$K$22
Grouped By: ● Columns ○ Rows
☑ Labels in First Row
Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

**Correlation - Set up dialog box**

| | gripR | gripL | HLR | HLL | FASFR | FASFL | FAGR | FAGL | SAFAGR | SAFAGL |
|---|---|---|---|---|---|---|---|---|---|---|
| gripR | 1 | | | | | | | | | |
| gripL | 0.854597 | 1 | | | | | | | | |
| HLR | 0.499165 | 0.507521 | 1 | | | | | | | |
| HLL | 0.593716 | 0.571354 | 0.93761 | 1 | | | | | | |
| FASFR | -0.55379 | -0.50556 | -0.33931 | -0.41879 | 1 | | | | | |
| FASFL | -0.50959 | -0.43246 | -0.39214 | -0.46046 | 0.950324 | 1 | | | | |
| FAGR | 0.616308 | 0.697226 | 0.574923 | 0.584657 | -0.04922 | -0.03157 | 1 | | | |
| FAGL | 0.553192 | 0.691243 | 0.496687 | 0.498774 | -0.07575 | -0.06611 | 0.95942 | 1 | | |
| SAFAGR | 0.778533 | 0.830347 | 0.652368 | 0.694138 | -0.46049 | -0.42411 | 0.909252 | 0.884246 | 1 | |
| SAFAGL | 0.691667 | 0.786612 | 0.596157 | 0.624007 | -0.42945 | -0.43965 | 0.875612 | 0.925267 | 0.957166 | 1 |

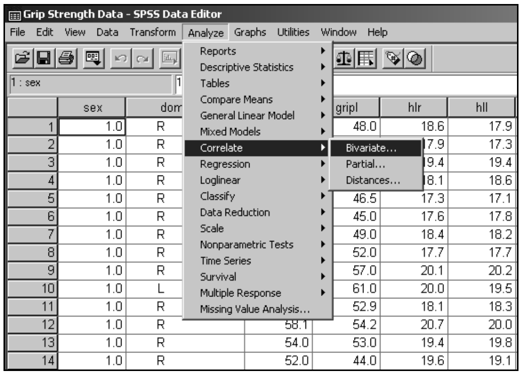
**Figure 2-5.6: Correlation using MS EXCEL,**
**Correlation matrix (bottom) produced from dialog boxes shown (top)**

gripR = right grip strength, gripL = left grip strength, HLR = right hand length, HLL = left hand length,
FASFR = right forearm skinfold, FASFL = left forearm skinfold, FAGR = right forearm girth, FAGL = left forearm girth,
SAFAGR = right skinfold-adjusted forearm girth, SAFAGL = left skinfold-adjusted forearm girth

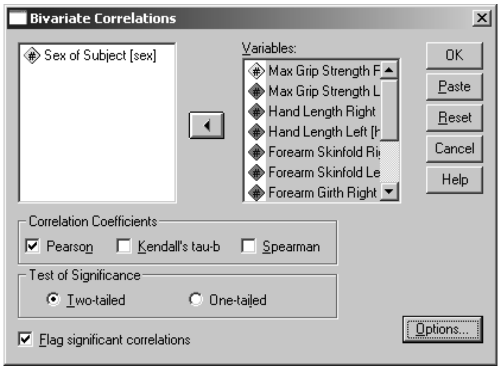
## Correlation Using SPSS

Using SPSS, a correlation matrix can be produced by selecting the CORRELATE option from the ANALYZE menu (Figure 2-5.7). The BIVARIATE choice is then made from the options provided. This will bring up the BIVARIATE CORRELATIONS dialog box. In this dialog box you can select any of your defined variables and move them over to the variables list. Any variable included in this list will be included in the correlation matrix. There are three options for correlation coefficients provided. For continuous variables you should select PEARSON. The other two options are nonparametric coefficients and will be dealt with in Chapter 2-9. The resultant correlation matrix shown in Figure 2-5.7 has more information than the MS EXCEL output. In addition to the correlation coefficients, you are provided with the probability level of a significant relationship and the sample size on which the correlation coefficient is based.

**Pairwise & Listwise Exclusion of Cases:** In the data set that was used for this example there were no missing values for any of the variables. Thus, the sample size (N) for every cell in the correlation matrix is 23. If however, one or more variables had some missing values then you would have to make a decision on how to deal with the missing values in calculating the correlation coefficients.



**Analyze – Correlate - Bivariate**

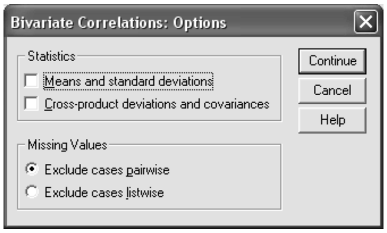


Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

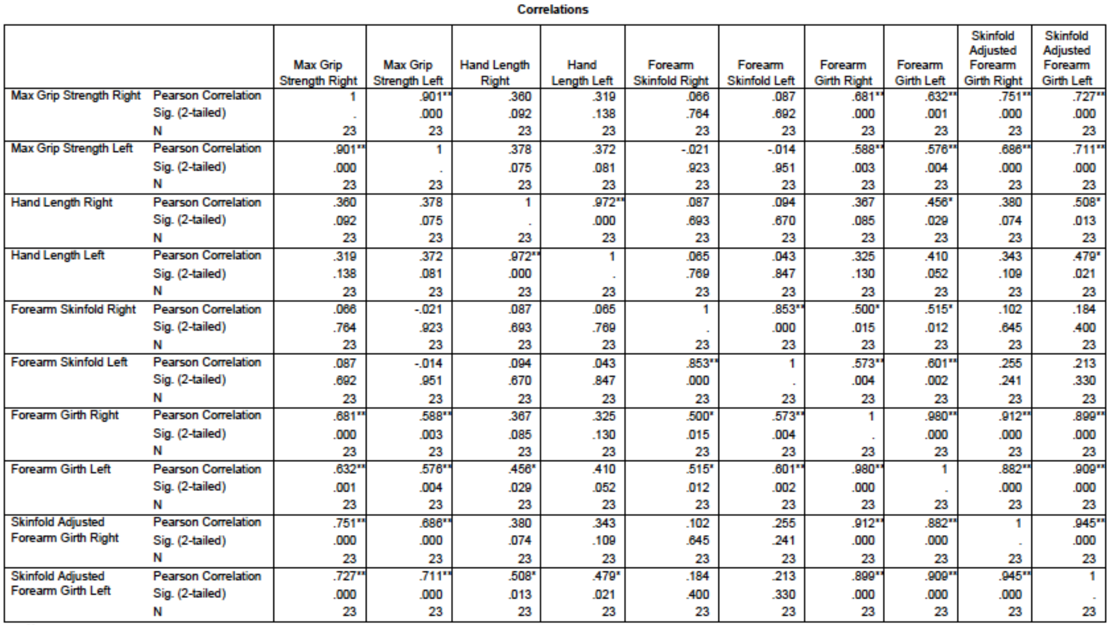
**Bivariate Correlations - Set up dialog box**



Statistics
☐ Means and standard deviations
☐ Cross-product deviations and covariances

Missing Values
● Exclude cases pairwise
○ Exclude cases listwise

**Bivariate Correlations - Options dialog box**



Correlations

| | | Max Grip Strength Right | Max Grip Strength Left | Hand Length Right | Hand Length Left | Forearm Skinfold Right | Forearm Skinfold Left | Forearm Girth Right | Forearm Girth Left | Skinfold Adjusted Forearm Girth Right | Skinfold Adjusted Forearm Girth Left |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Max Grip Strength Right | Pearson Correlation | 1 | .901** | .360 | .319 | .066 | .087 | .681** | .632** | .751** | .727** |
| | Sig. (2-tailed) | . | .000 | .092 | .138 | .764 | .692 | .000 | .001 | .000 | .000 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Max Grip Strength Left | Pearson Correlation | .901** | 1 | .378 | .372 | -.021 | -.014 | .588** | .576** | .686** | .711** |
| | Sig. (2-tailed) | .000 | . | .075 | .081 | .923 | .951 | .003 | .004 | .000 | .000 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Hand Length Right | Pearson Correlation | .360 | .378 | 1 | .972** | .087 | .094 | .367 | .456* | .380 | .508* |
| | Sig. (2-tailed) | .092 | .075 | . | .000 | .693 | .670 | .085 | .029 | .074 | .013 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Hand Length Left | Pearson Correlation | .319 | .372 | .972** | 1 | .065 | .043 | .325 | .410 | .343 | .479* |
| | Sig. (2-tailed) | .138 | .081 | .000 | . | .769 | .847 | .130 | .052 | .109 | .021 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Forearm Skinfold Right | Pearson Correlation | .066 | -.021 | .087 | .065 | 1 | .853** | .500* | .515* | .102 | .184 |
| | Sig. (2-tailed) | .764 | .923 | .693 | .769 | . | .000 | .015 | .012 | .645 | .400 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Forearm Skinfold Left | Pearson Correlation | .087 | -.014 | .094 | .043 | .853** | 1 | .573** | .601** | .255 | .213 |
| | Sig. (2-tailed) | .692 | .951 | .670 | .847 | .000 | . | .004 | .002 | .241 | .330 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Forearm Girth Right | Pearson Correlation | .681** | .588** | .367 | .325 | .500* | .573** | 1 | .980** | .912** | .899** |
| | Sig. (2-tailed) | .000 | .003 | .085 | .130 | .015 | .004 | . | .000 | .000 | .000 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Forearm Girth Left | Pearson Correlation | .632** | .576** | .456* | .410 | .515* | .601** | .980** | 1 | .882** | .909** |
| | Sig. (2-tailed) | .001 | .004 | .029 | .052 | .012 | .002 | .000 | . | .000 | .000 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Skinfold Adjusted Forearm Girth Right | Pearson Correlation | .751** | .686** | .380 | .343 | .102 | .255 | .912** | .882** | 1 | .945** |
| | Sig. (2-tailed) | .000 | .000 | .074 | .109 | .645 | .241 | .000 | .000 | . | .000 |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Skinfold Adjusted Forearm Girth Left | Pearson Correlation | .727** | .711** | .508* | .479* | .184 | .213 | .899** | .909** | .945** | 1 |
| | Sig. (2-tailed) | .000 | .000 | .013 | .021 | .400 | .330 | .000 | .000 | .000 | . |
| | N | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

**Figure 2-5.7: SPSS Correlation. Analyze – Correlate - Bivariate dialog box (top left) Bivariate Correlations dialog box (top right), Bivariate Correlations - Options dialog box (middle), Correlation Matrix for all variables in Grip strength data (Women)**

Figure 2-5.7 shows the BIVARIATE CORRELATIONS dialog box. In it you can choose to select **missing values being excluded pairwise, or listwise**. The distinction between these two is important in terms of the resulting correlation matrix. If listwise exclusion is chosen then if missing values are encountered for any subject (case) then all of the data for that case will be excluded from calculation. The matrix will therefore have the same number of cases (N) in each cell, but it will be less than the total sample size, depending upon the number of missing values found. If pairwise exclusion is chosen, then as each pair of variables is used to calculate the correlation coefficient, any cases with either variable having missing values will be excluded. What this means is that if different variables have different cases with missing values then the number of cases (N) used in each cell of the matrix may be different. It all depends upon which cases have missing values.

This will affect the comparability of coefficients between variables. One coefficient in the matrix may have a smaller N than another, which may or may not dramatically affect the coefficient. If the sample size was in the thousands, 10 cases different between cells would have minimal effect on the correlation coefficient. However, if as in the example here, the sample size is only 23, and it was 18, 19, or 20 for different cells in the matrix, because of the missing values, this could have a profound effect on comparability. The safest method is to use listwise exclusion, thus ensuring similar numbers of cases. Unfortunately, you may have to use pairwise exclusion, if there are many missing values spread around different variables and cases. This is just one example of how multiple missing values can hinder your analysis. The best tactic therefore is to strive as hard as possible to ensure that there are minimal missing values in your data set.

**Statistical Significance of r:** The exact probability of a significant r is provided in each cell in the matrix. These can be compared to predetermined acceptance levels such as $p < 0.05$, in order to assign statistical significance. A default option is to have SPSS flag significant correlations ($p < 0.05$) with an asterisk in the cell. This can be deselected if desired in the BIVARIATE CORRELATIONS, OPTIONS dialog box (Figure 2-5.7 middle).

## Summary of Pitfalls in the Use of the Correlation Coefficient (r)

The Pearson Product-Moment Correlation Coefficient (r), is a very useful statistic; however there are several pitfalls to its use, which one must be wary of:

**It provides a linear fit to the data:** Because you get a statistically significant, even high value of r, does not mean that there is a linear relationship. A curvilinear relationship may fit better. Always plot your data and look for nonlinearity, or preferably, test for a better fit with other equations. Another possibility is transforming one or both variables with a transformation such as $\log_{10}$, in order to make the relationship linear.

**Correlation does not mean causation:** Two variables may be related because of a relationship to a third variable. A correlation coefficient is insufficient evidence to ascribe causality.

**Statistical significance of r does not infer practical significance:** A statistically significant r means that you are confident at a certain level (often 95%) that the degree of association you see in the sample, actually exists in the population sampled from. It does not mean that there is a strong association (high r) or that you can predict one variable from the other.

**The range of the data will affect the correlation coefficient:** If you inadvertently sample individuals with smaller values for the tested variables you would find that correlation coefficients would tend to lower than if the whole range of size of scores had been sampled. Conversely a very large range can give very high values of r.