

## EXERCISE 17

## The Pearson correlation

---

### Before you start

Before starting to work through this practical Exercise, we recommend that you read Chapter 11. The **Pearson correlation  $r$**  is one of the most widely used (and abused) of statistics. Despite its apparent simplicity and versatility, it is only too easy to misinterpret a correlation. The purpose of the present Exercise is not only to show you how to use SPSS to obtain correlations, but also to illustrate how a given value for  $r$  can sometimes be misleading.

### A famous data set

This exercise involves the analysis of four sets of paired data, which were constructed many years ago by Anscombe (1973), in order to make some important points about correlations.

Participant	X1	Y1	Y2	Y3	X2	Y4
1	10.0	8.04	9.14	7.46	8.0	6.58
2	8.0	6.95	8.14	6.77	8.0	5.76
3	13.0	7.58	8.74	12.74	8.0	7.71
4	9.0	8.81	8.77	7.11	8.0	8.84
5	11.0	8.33	9.26	7.81	8.0	8.47
6	14.0	9.96	8.10	8.84	8.0	7.04
7	6.0	7.24	6.13	6.08	8.0	5.25
8	4.0	4.26	3.10	5.39	19.0	12.50
9	12.0	10.84	9.13	8.15	8.0	5.56
10	7.0	4.82	7.26	6.42	8.0	7.91
11	5.0	5.68	4.74	5.73	8.0	6.89

Each set yields exactly the same value for the **Pearson correlation**. The scatterplots, however, will show that in only one case are the data suitable for a Pearson correlation; in the others, the Pearson correlation gives a highly misleading impression of the relationship between the two variables. Ideally a scatterplot should indicate a **linear relationship** between the variables i.e. that all the points on the scatterplot should lie along or near to a diagonal straight line as shown in Chapter 11, Figure 1. Vertical or horizontal lines are not examples of

linear relationships; moreover, the Pearson correlation is not defined when a data set comprises only one value of one variable in combination with various values of another.


The data are presented in Table 1. The four sets we shall examine are variable  $X1$  with each of the variables  $Y1$ ,  $Y2$ , and  $Y3$ , and finally variable  $X2$  with variable  $Y4$ .

## Preparing the SPSS data set

After naming the first variable in **Variable View** as *Case*, name the remaining variables (ensure that the variables are all of type **Scale** in the **Measure** column of **Variable View**) as shown in the data table above. The value in the **Decimals** column should be 2. Switch to **Data View**, enter the data and save the set to a file called *Anscombe*. (This file will be used again in Exercise 20.)

## Exploring the data

Obtain scatterplots of the four data sets, as described in Sections 11.3.1 and 5.7 using the **Scatter/Dot** facility within **Chart Builder**. These plots can be produced either one at a time by choosing **Simple Scatter** from the choice of scatterplot diagrams or, more dramatically, by

selecting **Scatterplot Matrix** , which obtains a grid of scatterplots made up of all pairwise combinations of several variables. In the present Exercise, however, we only want the plots of variables  $Y1$ ,  $Y2$  and  $Y3$  against variable  $X1$ , and of variable  $Y4$  against variable  $X2$ . Thus it is better to use **Scatterplot Matrix** for the plots with  $X1$  and **Simple Scatter** for the plot of  $Y4$  against  $X2$ .

If the matrix scatterplot is selected and variables  $X1$ ,  $Y1$ ,  $Y2$  and  $Y3$  are transferred to the **Scattermatrix?** box, only the first column of plots, (those with  $X1$  on the horizontal axis), will be of interest. Transfer the variable names  $X1$ ,  $Y1$ ,  $Y2$  and  $Y3$  one at a time by highlighting the variable name and then dragging it to the left-hand end of the **Scattermatrix?** box where a smaller box with + inside will appear after the first variable name has been transferred. The order of the variables should be  $X1$ ,  $Y1$ ,  $Y2$  and  $Y3$  from left to right: if not, the order can be changed within the **Elements Properties** panel.

- **What do you notice about the scatterplots in the first column? Which one is (in its present state) suitable for a subsequent calculation of a Pearson correlation? Describe what is wrong with each of the others.**

Return to **Chart Builder**, select **Scatter Scatter**, and prepare a simple scatterplot of variable  $Y4$  against variable  $X2$ .

- **Is the plot suitable for a Pearson correlation?**

The plot of  $Y1$  against  $X1$  shows a substantial linear relationship between the variables. The thinness of the imaginary ellipse of points indicates that the **Pearson correlation** is likely to be high. This is the kind of data set for which the Pearson correlation gives an informative and accurate statement of the strength of the linear association between two variables. The other plots, however, are very different: that of  $Y2$  against  $X1$  shows a perfect, but clearly non-linear, relationship;  $Y3$  against  $X1$  shows a basically linear relationship, which is marred by a glaring outlier;  $Y4$  against  $X2$  shows a column of points with a single outlier up in the top right corner.

## Pearson correlations for the four scatterplots

Using the procedure described in Section 11.3.2, obtain the correlations between  $X$  and  $Y$  for the four sets of paired data. This is most easily done by entering the variables  $X1$ ,  $Y1$ ,  $Y2$ ,  $Y3$  in the first run of the procedure so as to get a correlation matrix, and then  $X2$  and  $Y4$  in the second run.

- **What do you notice about the value of  $r$  for each of the correlations?**

Anscombe's data strikingly illustrate the need to inspect the data carefully to ascertain the suitability of statistics such as the Pearson correlation.

## Removing the outliers

It is instructive to recalculate the **Pearson correlation** for the data set ( $X1$ ,  $Y3$ ) when the values for Participant 3 have been removed. The outlier is the value  $12.74$  on the variable  $Y3$ . Use the **Select Cases...** procedure to select all participants except Participant 3.

Return to the **Scatterplot** and **Bivariate Correlations** dialog boxes for  $X1$  and  $Y3$  (ignore the other variables) to re-run these procedures using the selected cases. Check that in the listing, only 10 rather than 11 cases have been used. You should find that the Pearson correlation for  $X1$  and  $Y3$  is now  $+1$ , which is what we would expect from the appearance of the scatterplot.

## Conclusion

This Exercise has demonstrated the value of exploring the data first before calculating statistics such as the **Pearson correlation**. While it is true that Anscombe's data were specially constructed to give his message greater force, there have been many misuses of the Pearson correlation with real data sets, where the problems created by the presence of outliers and by basically non-linear relationships are quite common.

## Finishing the session

Close down SPSS and any other windows before logging out.