coefficient, and Pearson's correlation coefficient. I will typically refer to it as Pearson's *r* for the sake of brevity.

Pearson's *r* is used to illustrate the relationship between two continuous variables, such as years of education completed and income. The correlation between any two variables using Pearson's *r* will always be between –1 and +1. A correlation coefficient of 0 means that there is no relationship, either positive or negative, between these two variables. A correlation coefficient of +1 means that there is a perfect positive correlation, or relationship, between these two variables. In the case of +1, as one variable increases, the second variable increases in exactly the same level or proportion. Likewise, as one variable decreases, the second variable would decrease in exactly the same level or proportion. A correlation coefficient of –1 means that there is a perfect negative correlation, or relationship, between two variables. In this case, as one variable increases, the second variable decreases in exactly the same level or proportion. Also, as one variable decreases, the other would increase in exactly the same level or proportion.

You most likely will never see a correlation between two variables of –1 or +1 in the social sciences as while two variables may be very highly related, the chance of error or random variation is too great to have a perfect correlation. A positive correlation means that generally, as one variable increases, the other will increase, and as one variable decreases, the other will decrease. Also, a negative correlation means that in general, if one variable increases, the other will decrease, and as one variable decreases, the other will increase. Very important here is the notion of significance, which I introduced you to in Chapter 1. When determining Pearson's *r*, or other correlation coefficients, it is important to be aware of whether your correlation is in fact significant or not at the .05 level.

Let's now compute the Pearson's *r* for some data. The table below consists of data made up for this example.

Years of Education (x)	Income (in Thousands of $) (y)
8	12
12	15
8	8
14	20
12	18
16	45
20	65
24	85
24	100
24	90

As you may have noticed, I tried to create a positive relationship between years of education and income—I am hoping that this will result in a strong positive correlation coefficient that will be significant.

The equation for Pearson's r is as follows:

$$r = \frac{\sum xy - N\bar{x}\bar{y}}{\sqrt{\left(\sum x^2 - N\bar{x}^2\right)\left(\sum y^2 - N\bar{y}^2\right)}}$$

This equation requires us to first calculate the sum of the product of all our data pairs, the means of both variables, and the sum of the squared values of both variables.

So first,

$$\begin{aligned}\sum xy &= (8 \times 12) + (12 \times 15) + (8 \times 8) + (14 \times 20) + (12 \times 18) + (16 \times 45) \\ &\quad + (20 \times 65) + (24 \times 85) + (24 \times 100) + (24 \times 90) \\ &= 96 + 180 + 64 + 280 + 216 + 720 + 1{,}300 + 2{,}040 + 2{,}400 + 2{,}160 \\ &= 9{,}456\end{aligned}$$

$$\bar{x} = \frac{8 + 12 + 8 + 14 + 12 + 16 + 20 + 24 + 24 + 24}{10} = \frac{162}{10} = 16.2$$

$$\bar{y} = \frac{12 + 15 + 8 + 20 + 18 + 45 + 65 + 85 + 100 + 90}{10} = \frac{458}{10} = 45.8$$

$$\sum x^2 = 8^2 + 12^2 + \mathrm{L} + 24^2 = 2{,}996$$

$$\sum y^2 = 12^2 + 15^2 + \mathrm{L} + 90^2 = 32{,}732$$

N = Number of cases or data pairs = 10.

Now, plugging these values into our equation, we get the following:

$$\begin{aligned}r &= \frac{\sum xy - N\bar{x}\bar{y}}{\sqrt{\left(\sum x^2 - N\bar{x}^2\right)\left(\sum y^2 - N\bar{y}^2\right)}} \\ &= \frac{9456 - 10(16.2)(45.8)}{\sqrt{(2996 - 10(16.2^2))(32732 - 10(45.8^2))}} \\ &= \frac{9456 - 7419.6}{\sqrt{(2996 - 2624.4)(32732 - 20976.4)}} = \frac{2036.4}{\sqrt{4368380.96}} = 0.9743\end{aligned}$$

I will use this same example in the sections on IBM SPSS and Stata—in those sections, you will be able to see that the result for Pearson's r using either of these programs is identical to the value we have calculated by hand.

Now, we can see that our correlation, .9743, is very high as it is very close to +1, the maximum possible value for Pearson's r. But we still need to calculate the p value in order to determine whether this correlation is statistically significant or not.

To determine this, we will first calculate a *t* ratio using the following equation:

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

Now, plugging our values into the equation, we get the following:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{.9743\sqrt{10-2}}{\sqrt{1-.9743^2}} = \frac{.9743\sqrt{8}}{\sqrt{.0507}} = \frac{2.7557}{.2251} = 12.2386$$

Also, we will need to know our degrees of freedom (*df*). This is equal to the number of pairs of data minus 2:

$$df = N - 2 = 10 - 2 = 8$$

Next, we will need to consult a *t* table to compare our calculated *t* value with the critical *t* value in order to determine statistical significance. Looking at a *t* table, we can see that for 8 degrees of freedom, the critical *t* value for a *p* level of .05 (two-tailed) is 2.306. As our calculated *t* value is greater than the critical *t* value at the .05 level, we can say that the correlation between education and income is significant at the .05 level. Again referring to our table, we can see that our correlation is even significant at the .001 level, as the critical *t* value in this case is 5.041, which is still lower than our calculated *t* value. This means that the probability that the correlation between education and income is simply due to error or chance is less than 0.1%. In this example, I have used the two-tailed critical *t* value, which is more conservative than a one-tailed test and is generally preferred. If you are not making a **directional hypothesis** (examples of a directional hypothesis: those with greater levels of education will have higher incomes or males have higher incomes than females), then you would use a **two-tailed test**, as it does not make any specification regarding direction. For example, a two-tailed test would be used if you're simply hypothesizing that there will be a correlation between level of education and income, but not specifying the direction of the correlation. However, if you were making a directional hypothesis, for example that those with more education are more likely to have higher incomes, the **one-tailed test** could be used. However, when the direction between your two variables corresponds to the direction stated in your hypothesis, the one-tailed test is less conservative than the two-tailed test and so tends to be used less often.

In the next section, the concept of ***R*-squared** will be discussed. The *R*-squared value represents the proportion of variance in the dependent variable (the variable you are trying to predict or explain) that is explained by the independent variable(s) (the variables that you are using to explain or predict the dependent variable). In this example, it would make sense that we would use years of education to predict the respondent's income and not vice versa. What's interesting is that we simply need to square the value we arrived at after calculating Pearson's r to attain the R-squared. Thus,

$$R^2 = r^2 = .9743^2 = .9493$$

Later on, in the Stata section, I will replicate this result. We can interpret this by stating that level of education explains 94.93% of the variance in income. Here, I simply moved the decimal point two places to the right to arrive at this value.

Finally, it is important to state again that Pearson's r is only used for continuous variables. To determine the correlation between variables that are ordered and categorical or dichotomous, there are a number of special options, including **Kendall's tau**, **Spearman's rank correlation coefficient** or **Spearman's rho**, the **polyserial correlation**, the **polychoric correlation**, **phi**, the **tetrachoric correlation**, and others. Many of these tests require specialized software programs or certain specific add-ons to IBM SPSS or Stata. These additional measures of correlation are described in more detail in Appendix C, Section 4, Part F.

Chi-Square: Theory

The chi-square statistic is used to show whether or not there is a relationship between two categorical variables. It can also be used to test whether or not a number of outcomes are occurring in equal frequencies or not, or conform to a known distribution. For example, when rolling a die, there are six possible outcomes. After rolling a die hundreds of times, you could tabulate the number of times each outcome occurred and use the chi-square statistic to test whether these outcomes were occurring in basically equal frequencies or not (e.g., to test whether the die is weighted). The chi-square statistic was also developed by Karl Pearson.

This is the chi-square equation:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Here,

χ^2 = the chi-square statistic

O_i = the observed frequency

E_i = the expected frequency

i = the number of the cell (cell 1, cell 2, etc.)