
11

CORRELATION

Correlation has an intuitive appeal. Most everyone understands that correlation is concerned with whether changes in one thing are linked to changes in something else. Thus, for example, you may observe that wealthier people seem to have better overall health than those with lower wealth levels. Or, stated differently, as wealth increases, health increases as well.

Correlation is a way of understanding the association between two variables. What does association mean? It refers to “relatedness” or the extent to which two events “vary with one another.” To use this language with the wealth–health example just cited, you might say that the increase in wealth (however it is measured) is accompanied by increases in health (according to its measures). In a general sense, you can say that the values of the one variable (health) change positively as the values of the other variable (wealth) increase. Thus, the values of both variables increase together, or “covary.”

Not every correlation is what it seems, however. There may be additional variables not taken into account in the analysis that give the original two variables the “appearance of covarying.” This possibility of spuriousness (see Chapter 2) hints at the complexity of a seemingly obvious relationship. Is it really wealth that results in health, or is it something (or things) else? In this and the following chapters on regression, I intend to explore this question further.

THE CORRELATION DESIGN

331

TABLE 11.1 Measures of Correlation

Correlation Measures			
Nominal Level Data	Ordinal Level Data	Nominal by Ordinal	Proportional Reduction in Error (PRE)
Contingency Coefficient	Spearman's Rho	<i>Biserial</i>	η^2
			Partial η^2
			r^2
Phi Coefficient	Kendall's Tau	Point Biserial	Gamma
Cramer's V			(Cohen's) Kappa
<i>Tetrachoric</i>			

tetrachoric correlation measures two continuous (interval-level) variables that have been *dichotomized*. A researcher may use this procedure by categorizing income as high, medium, or low and also measuring health as “weak” or “strong.” Both of these are continuous variables, but some researchers dichotomize them for various reasons (despite the fact that it isn't a good idea!). Tetrachoric correlation might measure the relationship between these two variables or other dichotomized variables.

Table 11.1 also shows a column (the fourth column) of measures known as proportional reduction in error (PRE) measures. I have already discussed some of those measures at the interval level (η^2 and partial η^2), but there are several others at different levels of data (indicated in Table 11.1 in different shades), for example, gamma for ordinal and Cohen's kappa for nominal data. These are measures in which correlation is used to “reduce error.” You saw in Chapter 9 (one-way ANOVA) that η^2 is a measure of how much variance in the outcome variable is explained by the categories of the predictor variable. That is, how much variation (error) is reduced by knowing the relationship between the two variables? Eta square and other such measures are expressed in the amount of *variance explained*.

We cannot hope to cover all of these in this book. I wanted you to see the tremendous variety of correlation measures; and there are yet more measures I didn't include in Table 11.1. I will discuss Pearson's correlation coefficient in this chapter which is the most common correlation procedure with interval data.

THE CORRELATION DESIGN

The correlation study is a post facto design since the researcher is *relating* two sets of scores that have been gathered on an individual case. You would not use correlation in an experimental design since in the latter, you attempt to detect *differences* in group scores after an intervention. Correlation measures “sameness” and an experiment measures “difference.” Multiple regression procedures, which are based on correlation, can be used in experimental designs under certain circumstances (see Abbott, 2010).

PLOTting THE CORRELATION: THE SCATTERGRAM

335

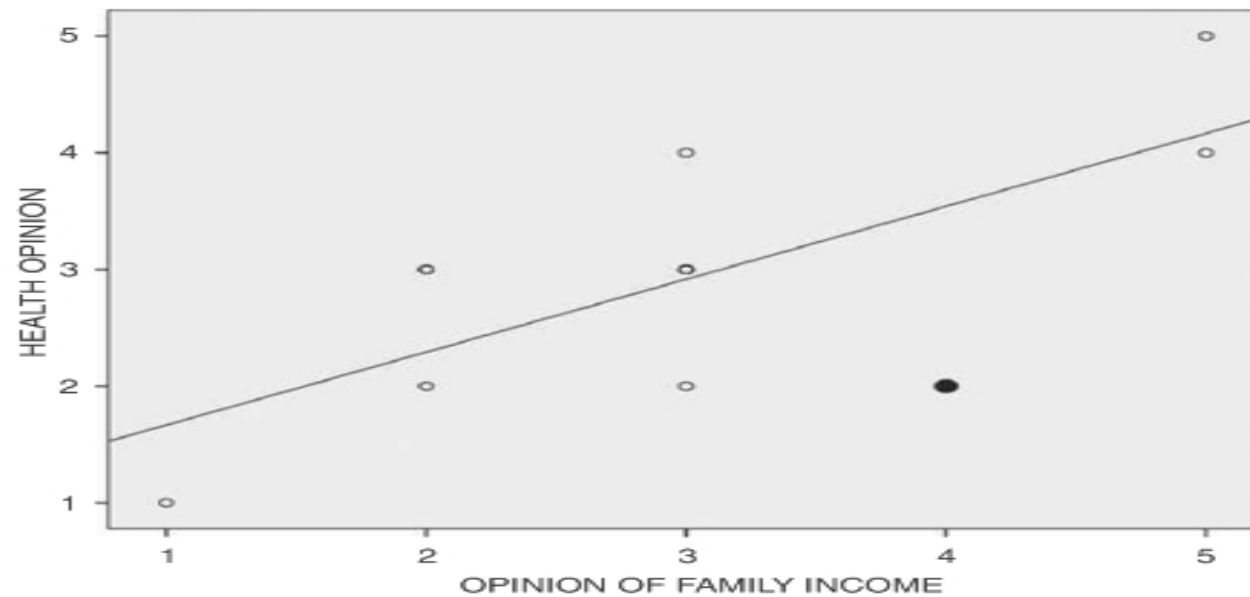


Figure 11.2 The scattergram between health and income.

the top pair of values in Table 11.2 indicates a respondent who indicated opinions that their income was average (value of 3) and that their health was good (value of 3). I filled in the dot in Figure 11.2 to show the pair of scores for the tenth respondent in Table 11.2 (the respondent indicating above average income and fair health).

As you can see, the pairs of scores are entered into the plot simultaneously so that each dot represents the pair of scores of an individual respondent. Typically, the outcome variable is placed on the Y-axis and the predictor variable is placed on the X-axis.

In correlation designs it is not always apparent which variable is the outcome and which is the predictor; you simply have to understand which is which from the research question. The “inherent” research question in this example study is whether income (opinion) influences health (opinion). Therefore, income is the predictor variable (X) in this study. It could just as easily be the other way around! A researcher could posit that health statuses influence a person’s income level.

Patterns of Correlations

Correlations can be positive like the one in Figure 11.2, or they can be negative and even nonlinear. Figure 11.3 shows some of the possibilities for correlation patterns in scattergrams. The upper left panel in Figure 11.3 shows the positive correlation with the actual data in Figure 11.2. However, the results could have been different. The upper right panel shows a negative (or inverse) correlation in which health decreases with higher income. The lower left panel shows a curvilinear relationship in which health is better at middle income levels but lower when income is either lower or

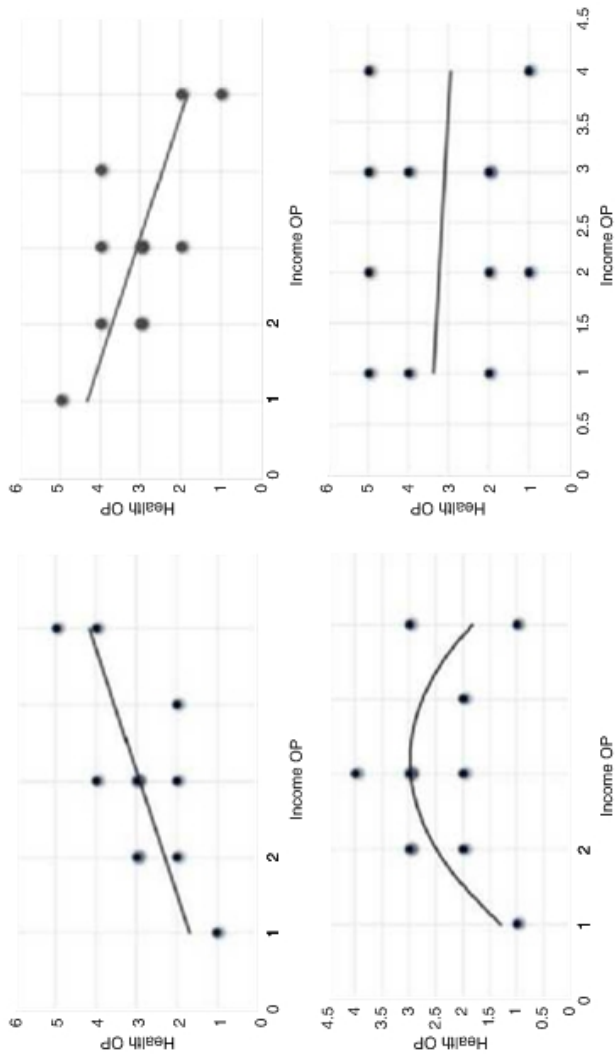


Figure 11.3 Correlation patterns in scattergrams.

a recognizable pattern, as indicated by a “flat” trend line. These panels are only to show the correlation possibilities that might be produced in an actual study.

Strength of Correlations in Scattergrams

The correlation panels in Figure 11.3 showed correlations with different directions or patterns. The dots could “extend upward to the right” (positive correlation), “downward to the right” (negative correlation), or in other patterns. Figure 11.4 shows scattergrams that indicate the strength of the correlation.

The top two panels represent positive correlations. The dots are arrayed from bottom left to upper right. But notice that the *extent of the scatter around the line* is different. I use a line to represent the scatter and pattern of the dots (in a later chapter, I will discuss how to calculate the equation for this “line of best fit”). When the dots have a wide scatter, like the scattergram in the top right panel, the correlation is weaker. This would indicate that as values of one variable (books read) increase, the values of the other also increase, but not consistently. The upper left panel shows a positive correlation with a tighter pattern of dots that indicate more of a consistent increase in the values of both variables. I use the darker color to show stronger correlations.

The lower panels of Figure 11.4 show negative correlations. In the bottom right panel, the correlation is inverse in that as one variable increases in value, the other decreases, but the values do not change consistently. The negative correlation in the bottom right panel shows a much higher correlation since the dots are very close to the line indicating a more consistent change in values.

The panels underscore the fact that “negative” correlations are not necessarily bad since negative only refers to direction. As you can see in the bottom left panel of Figure 11.4, this negative correlation is very strong.

You can easily freehand draw a scattergram with pairs of data. However, both SPSS and Excel provide simple procedures to create the graphs.

USING SPSS® TO CREATE SCATTERGRAMS

SPSS has a straightforward way to create scattergrams using the main Graphs menu. Figure 11.5 shows how to create the scattergram through the “Graph – Legacy Dialogs – Bar” path of menu choices. Near the bottom of the list is “Scatter/Dot” which will produce a specification box that allows you to design your scattergram. In this system, you do not need to move the data columns so that one of the variables is listed first, as in Excel.

When you choose “Scatter/Dot,” the menu box in Figure 11.6 appears. This box allows you to choose the scattergram that matches the complexity of the research data. For my fictitious example, the “Simple Scatter” choice is appropriate.

Choosing “Simple Scatter” produces a specification window like the one shown in Figure 11.7. As you can see, I specified that income opinion should be placed on the X-axis since I might hypothesize that income differences give way to different

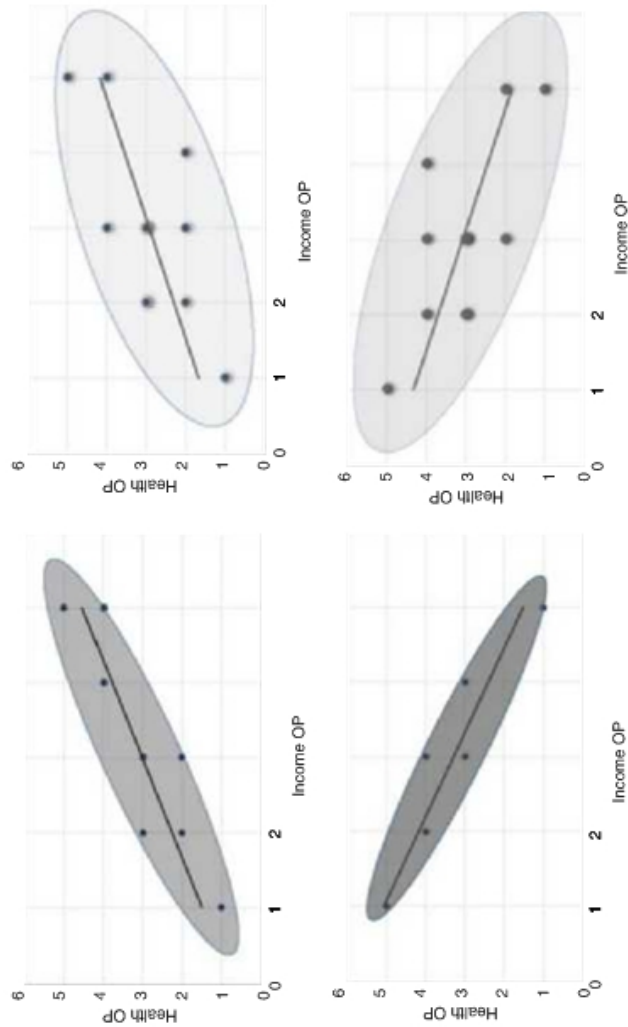


Figure 11.4 Strength of correlations in the scattergram.

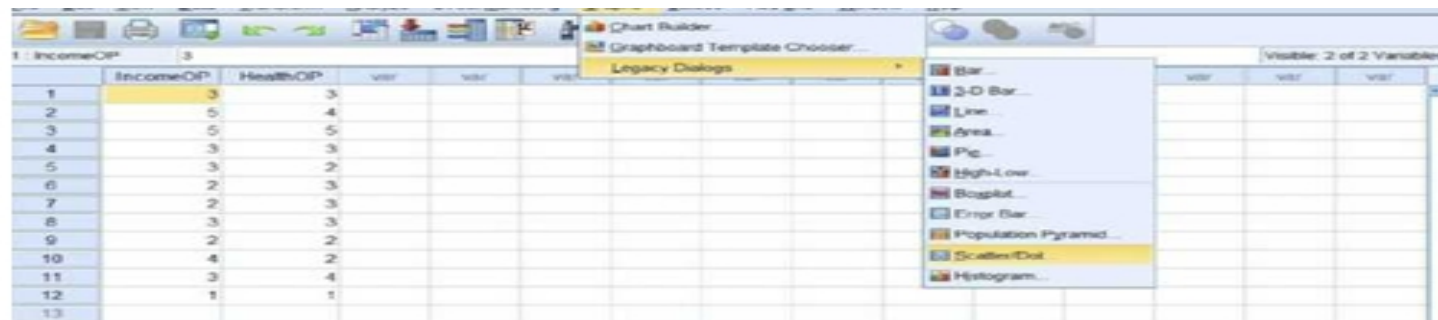


Figure 11.5 The SPSS® graph menu for creating scattergrams.

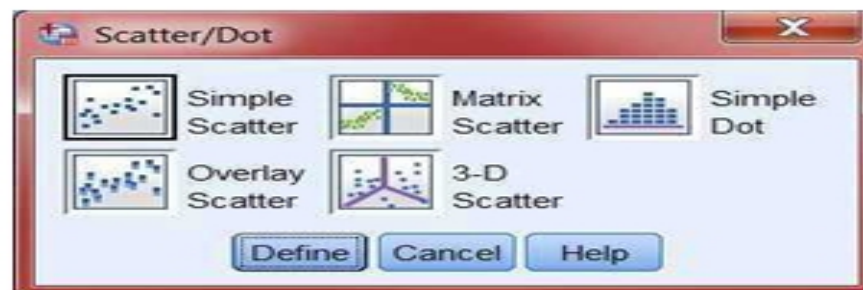


Figure 11.6 The Scatter/Dot menus.

health statuses. Health opinion is placed in the Y-axis. I can select the “Titles” and “Options” buttons in the upper right corner of the window to further specify the graph, but I prefer to edit the scattergram once it is produced.

When you choose “OK” from the window in Figure 11.7, SPSS produces a graph in the output file like the one shown in Figure 11.2. If you double-click on the graph, you can make a series of edits using the available menu screens. I added the line to the “basic” scattergram to produce the graph in Figure 11.2.

USING EXCEL TO CREATE SCATTERGRAMS

With the table of data entered in an Excel spreadsheet, simply highlight the entire table (including labels) and choose the “Insert” menu from the main menu list. Figure 11.8 shows this selection with the example data. As you can see, I highlighted the table of values and chose Insert. I then clicked on the Scatter icon listed among the other graph icons. At that point, I could choose several scattergram types in a drop-down menu. For the simple linear correlation, the upper left choice is appropriate. As the data increase in complexity, you can experiment with the other scattergram choices.

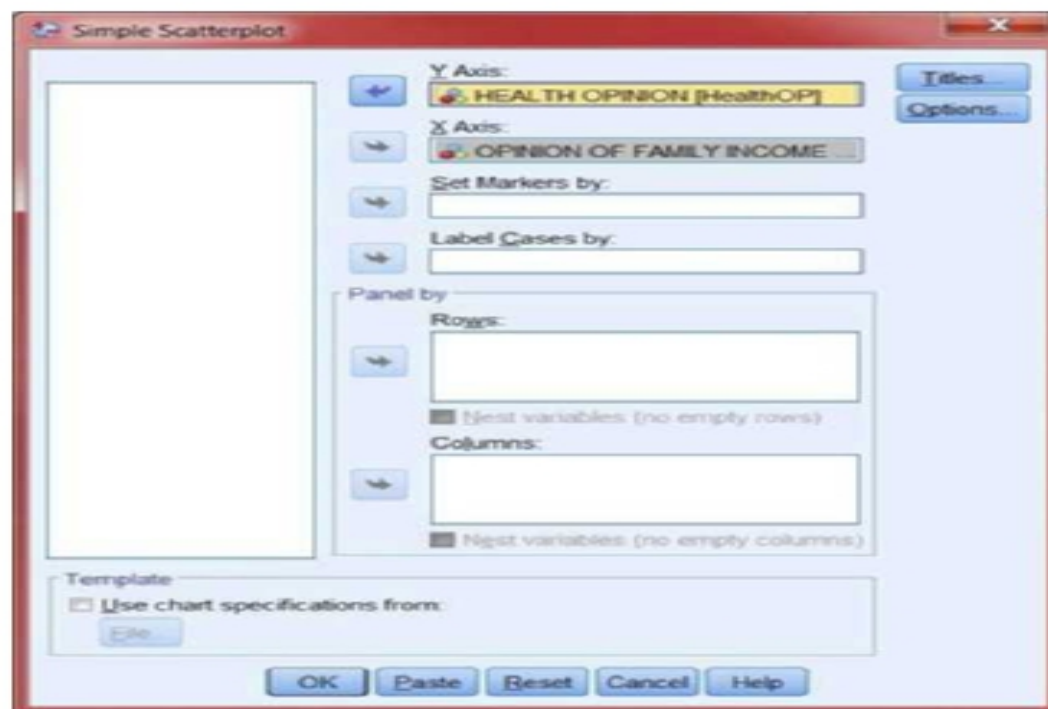
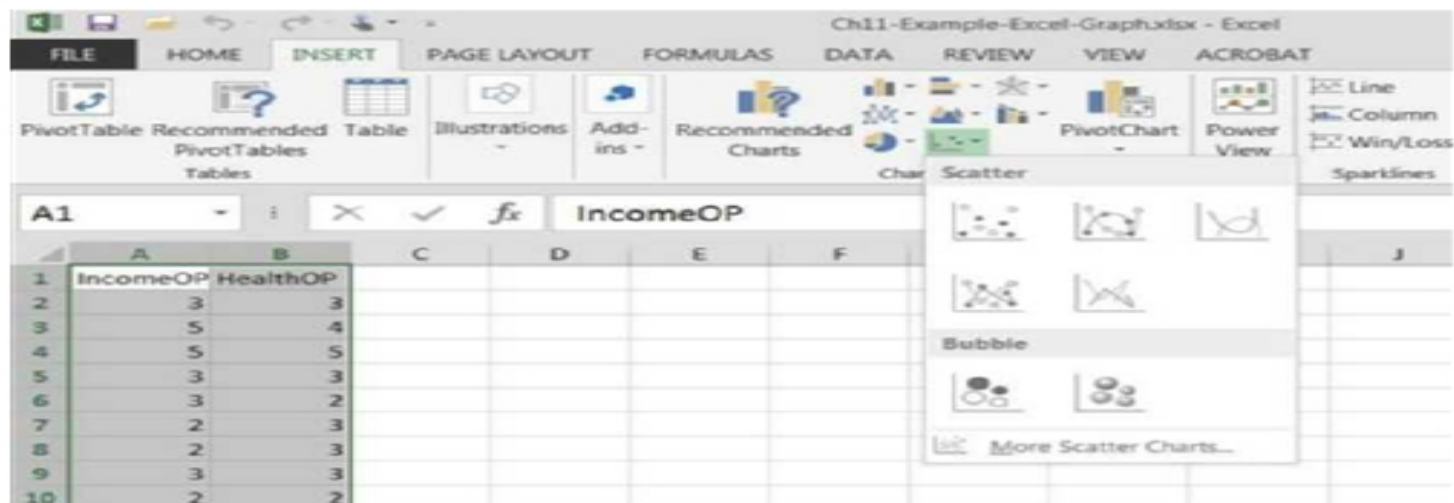


Figure 11.7 The scattergram specification window in SPSS®.



11	4	2
12	3	4
13	1	1
14		

Figure 11.8 The Excel scattergram specification.

CALCULATING PEARSON'S r

341

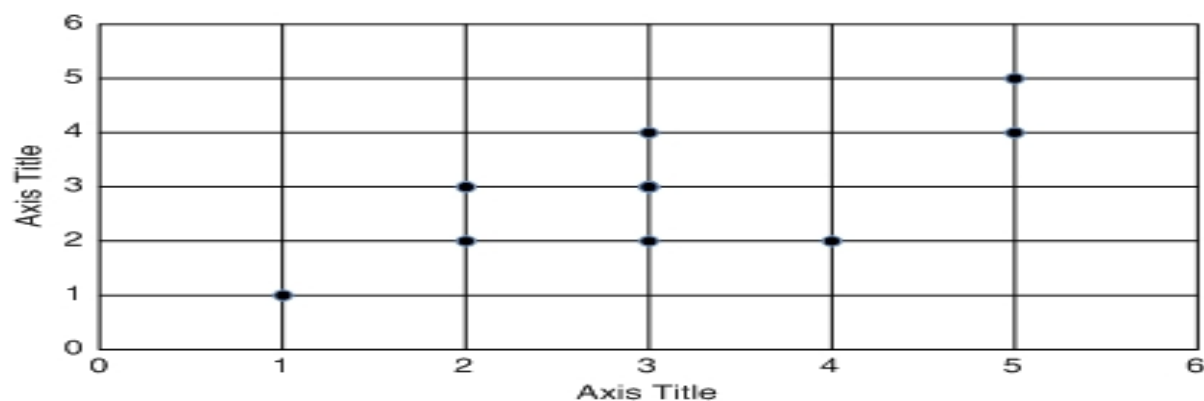


Figure 11.9 The Excel scattergram.

Figure 11.9 shows the scattergram created by the menu choices earlier. As you can see, the dots are displayed according to the pairs in the data table. One note here is that you must add the axis titles to the scattergram by editing it using the small icons that appear to the right of the graph when you select it. This is important to do since you want to make sure the variables are specified on the appropriate axes. As you can see in Figure 11.9, the axes are not titled, so you need to add the titles. In this case, the first variable (IncomeOP) in the Excel data table (see Figure 11.8) is placed on the X-axis as is appropriate.

The scattergram in Figure 11.9 is the “plain” graph that Excel produced using the specifications earlier. *You can edit the graph by double-clicking on parts of the graph and then right-clicking for edit choices.* A simple way to make edits is to use the main menu bar when you have the graph selected. The “Chart Tools” menu provides several different ways that you can make the graph appear. For example, one choice includes axis titles, another includes the raw data table, and so on. You can include the Trendline through the data using these methods.

CALCULATING PEARSON'S r

There are several ways to calculate Pearson's r manually. Formulas exist using the "deviation" method or the "calculation method" or even the "Z score method." I will present two formulas in this book since both point out different facets of Pearson's r .

The Z score method appears to be the most simple until you realize that, in order to use it, you have to transform every score to a Z score. The calculation method is my preferred method since it uses symbols and formulas I introduced in past chapters and includes a simple data table of values.

342

CORRELATION

THE Z SCORE METHOD

The many formulas for calculating Pearson's r manually result from measuring various elements that express different components of the correlation relationship.

The Z score formula is as follows:

$$r_{XY} = \frac{\sum Z_X Z_Y}{N}$$

In this formula, the correlation between two variables X and Y (symbolized by r_{XY}) is calculated by *summing the products* of the X_Z scores and the Y_Z scores and then dividing by the number of pairs (*in correlation, the N indicates pairs, not the total set of scores*). In order to carry out this formula, z scores for each of the X and Y values must be created using the formula I discussed in Chapter 4:

$$Z = \frac{X - M}{SD}$$

Creating z scores manually is tedious, especially with many pairs of scores. On the other hand, creating the z scores in an Excel spreadsheet is quite simple, as I discussed in Chapter 4. Table 11.3 shows the values as I created them in an Excel spreadsheet using the embedded z formula. I first calculated the mean and SD values for both variables as shown.

Once these calculations are made, the following Z score formula will calculate Pearson's r (using the relevant calculations in the final column of Table 11.3):

$$\begin{aligned} r_{XY} &= \frac{\sum Z_X Z_Y}{N} \\ &= \frac{8.39}{12} \end{aligned}$$

$$\frac{12}{12}$$

$$= 0.699$$

Note: Be careful with the calculation of the numerator of the formula. It is not the sum of the Z scores for the IncomeOP column (0) times the Z score sum of the HealthOP column (0). This mistake would look like the following:

$$r_{XY} = \frac{(\sum Z_X)(\sum Z_Y)}{N}$$

The reason that this is inappropriate is that the sums of these particular columns would be equal to 0! Like the deviation method for calculating SD that we discussed earlier, there will be an equal number of negative and positive z scores in a normally distributed set of scores. Therefore, you must *multiply each pair of Z scores* to arrive at the values in the last column of Table 11.3. Otherwise, this is the result:

$$r_{XY} = \frac{(0)(0)}{12}$$

$$= 0$$

THE Z SCORE METHOD

343

TABLE 11.3 The Data Table Showing Z Scores

	IncomeOP	Z(Income)	HealthOP	Z(Health)	$Z_X * Z_Y$
	3	0.00	3	0.08	0.00
	5	1.74	4	1.04	1.82
	5	1.74	5	2.01	3.49
	3	0.00	3	0.08	0.00
	3	0.00	2	-0.88	0.00
	2	-0.87	3	0.08	-0.07
	2	-0.87	3	0.08	-0.07
	3	0.00	3	0.08	0.00
	2	-0.87	2	-0.88	0.77
	4	0.87	2	-0.88	-0.77
	3	0.00	4	1.04	0.00
	1	-1.74	1	-1.85	3.21
Σ		0.00		0.00	
Mean	3.00		2.917	$\Sigma ZX * ZY$	8.39
SD	1.15		1.037	r	0.699

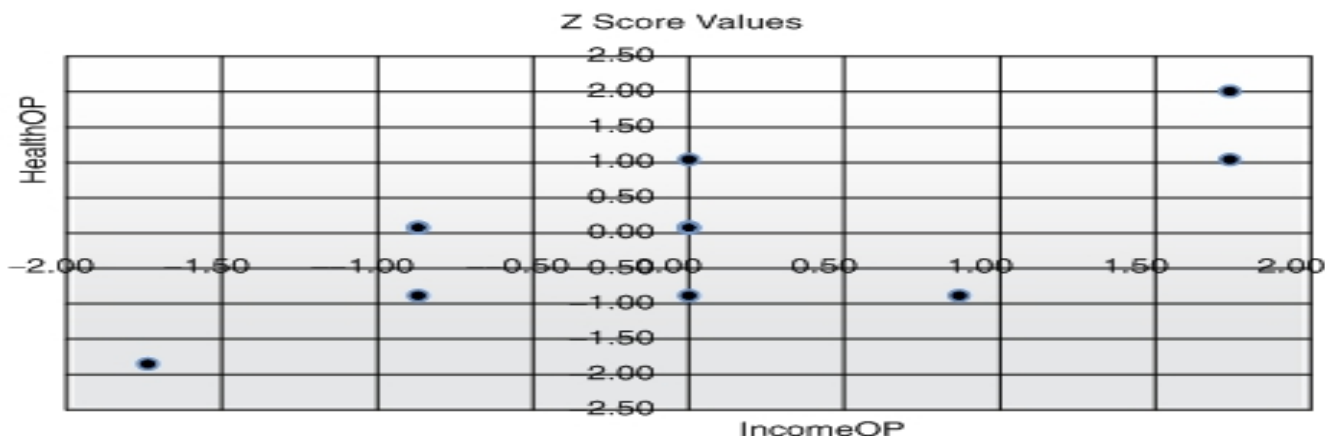


Figure 11.10 The Z score scattergram in Excel.

The reason to use the Z score formula is to use the transformed *X* and *Y* values as Z scores that present the *X* and *Y* values in the same scales. This will become more apparent when I discuss regression procedures, but for now, look at Figure 11.10 that presents the scattergram with Z score values for both the *X* and *Y* raw score values. The *X*-axis is the Z score values of IncomeOP and the *Y*-axis represents the Z values of the HealthOP.

As you can see, all the values of the two scores are now expressed in the same scale, as Z score values the majority of which (over 99%) range between -3.00 and $+3.00$.

Whereas in Figure 11.9 the values were expressed in *raw score units* (i.e., the units of the original scores), the transformed scores are expressed in *z score units*. To take an example, the lower left dot expressed in Figure 11.9 is (IncomeOP = 1, HealthOP = 1). In Figure 11.10, the Z scores are (IncomeOP = -1.74 , HealthOP = -1.85).

Using Z scores creates the same scale for raw score variable units of any magnitude. This does not change the correlation or the placement of the dots in the scattergram.

THE COMPUTATION METHOD

There are several computation formulas for Pearson's r . I prefer the following because it uses the sum of squares (SS) measures that I have used throughout former chapters. The overall formula looks straightforward. Conceptually, the formula expresses cor-

relation (“ r ”) as the proportion of the “cross product” SS (or the SS of the $X*Y$ pairs) to the square root of the overall product of the X and Y sums of squares³:

$$r_{XY} = \frac{SS_{XY}}{\sqrt{(SS_X)(SS_Y)}}$$

Recall that the SS is a general (unstandardized) variance measure of the population of scores. The computation formula thus includes variance (SS) measures of the XY products as well as the individual X and Y SS values. The formula for SS we discussed in Chapter 3 is

$$\sum X^2 - \frac{(\sum X)^2}{N}$$

Each of the SS values (for X and Y) in the Pearson’s r formula can be computed with these formulas. The product SS (the numerator in the Pearson’s r formula) can be calculated with a variation of the same SS formula by simply using the product column (XY) of values:

$$SS_{XY} = \sum XY - \frac{(\sum X)(\sum Y)}{N}$$

Here is how the income–health example data would be computed using the computation formula:

The SS_{XY}

$$SS_{XY} = \sum XY - \frac{(\sum X)(\sum Y)}{N}$$

$$SS_{XY} = 115 - \frac{(36)(35)}{12}$$

$$SS_{\text{Income} \times \text{Health}} = 10.00$$

³This makes conceptual sense if you recognize the denominator as a combined variance measure of X and Y together, since each SS is multiplied and transformed to a combined variance measure. Thus, the formula expresses the “variance” of the XY product measures as a proportion of the overall variance of both sets of scores. This is conceptually what ANOVA does – compares variance components.

THE HYPOTHESIS TEST FOR PEARSON’S r

345

The SS_Y

$$SS_Y = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

$$SS_{\text{Health}} = 115 - \frac{1225}{12}$$

$$SS_{\text{Health}} = 12.92$$

The SS_X

$$SS_X = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$SS_{\text{Income}} = 124 - \frac{1296}{12}$$

$$SS_{\text{Income}} = 16.00$$

Pearson's r

$$r_{XY} = \frac{10.00}{\sqrt{(16.00)(12.92)}}$$

$$= \frac{10.00}{14.38}$$

$$= 0.695$$

As you can see, the resulting Pearson's r of 0.695 is equivalent to the r value computed with the Z score formula (with minor rounding differences). Both formulas yield the same value.

THE HYPOTHESIS TEST FOR PEARSON'S r

What does the Pearson's r value indicate? Recall that the study asked if income was correlated to health. The scattergrams presented strong visual evidence that there was a relationship, but how can the resulting r of 0.695 help to judge the strength of the relationship?

The r of 0.695 is certainly closer to 1.00 than it is to 0, so you can use this as one gauge of its meaningfulness. However, like the other statistical tests I have discussed, you assess the statistical significance of a test ratio by comparing it to a sampling distribution.

Therefore, the first way to determine whether the r value is significant is to perform the hypothesis test as I have with other statistical procedures. If the calculated value exceeds the tabled value (i.e., if the calculated value is further out in the tail of the exclusion area), you can conclude that the test value is statistically significant.

Use the same steps as you did with other statistical tests:

1. The null hypothesis (H_0): $\rho = 0$

The null hypothesis states that the correlation between the variables in the population from which the sample values came (symbolized by the parameter value of rho or ρ) is 0.

2. The alternative hypothesis (H_A): $\rho \neq 0$

The alternative hypothesis states that the population correlation is not 0.

3. The critical value $r_{df(0.05)}$

The comparison value for the value of r that defines the exclusion area at a given probability level (I used 0.05 in the above value) is determined by a table of values as are other statistical procedures. The tables for correlation are extensive and detailed. The critical values for correlation table, which lists exclusion values for Pearson's r , is found in the Appendix.

Like other statistical procedures, testing the null hypothesis for correlation uses degrees of freedom since you estimate a population parameter from a sample statistic. The degrees of freedom for correlation is $N - 2$, where N is the number of pairs. You "lose" two degrees of freedom since you are estimating a population value with two sample sets of values.

The table for correlation specifies that the two-tailed 0.05 critical (exclusion) value for r at 10 degrees of freedom (12 pairs minus 2 = 10) is $r_{df(0.05)} = 0.576$. Values that are greater than 0.576 would therefore be considered statistically significant.

4. The calculated value of $r = 0.695$

5. Statistical decision

You reject the null hypothesis since the calculated value (0.695) exceeded the critical value of exclusion (0.576). *A significant correlation indicates that a calculated value with this magnitude is unlikely to occur by chance if there is no correlation in the population.*

6. Interpretation

You can consider the calculated r of 0.695 to be statistically significant. This indicates that it is likely not a chance finding. In the words of the example question, respondents' opinions of their incomes and their health statuses are positively correlated.

The Comparison Table of Values

Because the correlation tables are so extensive and because I will discuss SPSS and Excel results, I do not include comprehensive correlation tables in this book. However, there is an alternate way to identify the critical values by using the T table of values which is included. Cohen (1988) provides the following formula that you can use in the T tables as a comparison value. In effect, you "transform" the r value into a t ratio that you can then compare against the critical values of T in the T table

EFFECT SIZE: THE COEFFICIENT OF DETERMINATION**347**

(using $N - 2$ degrees of freedom in the table):

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

Therefore, in the present example

$$\begin{aligned} t &= \frac{0.695\sqrt{12-2}}{\sqrt{1-0.483}} \\ &= 3.057 \end{aligned}$$

The r value of 0.695 is “transformed” to a t ratio of 3.057. When you consult the T table of values using $N - 2$ degrees of freedom, you note that the critical value of $t_{0.05,10} = 2.228$. Therefore, the calculated r is significant since it resulted in a t score that exceeded the critical T value.

EFFECT SIZE: THE COEFFICIENT OF DETERMINATION

A second way to judge the meaningfulness and strength of a correlation beyond the hypothesis is the effect size. I have shown with every statistical procedure ways to measure the “impact” of a relationship. With correlation, you can judge the strength of the relationship by the *coefficient of determination* or r^2 . *This value is simply the square of r and refers to the amount of variance in one variable explained by the other.* What this means is this: you can consider the fact that a distribution of scores (e.g., health opinion values) vary a certain amount or are spread out around a mean score.

The question is why do health opinion values vary? In a world where every person was the same, there would be no variance – everyone would offer the same health opinion. But we don’t live in that kind of world, so something or things are responsible for people getting different health opinion values. Figure 11.11 shows how the example study variables relate to one another in terms of “explaining variance.”

If you establish that a correlation is significant (i.e., between income and health opinions), the variables’ measures overlap (like the Venn diagram in Figure 11.11), and you can understand this overlap to be the amount of variance in the outcome measure (health opinions) that is accounted for by the predictor variable (income opinion). Thus, knowing a respondent’s income opinion is a partial explanation of their health opinions. The income opinion does not explain all the variation in health opinion, since there is still a lot of “unexplained variance” (the amount of the health opinion circle in Figure 11.11 not overlapping with income opinion circle), but it “chips

away from the overall spread of the scores. The r^2 value is this explained variance.

Figure 11.12 shows another way to visualize effect size in correlation. The overall distribution represents the entire variability of health opinion values. With a correlation of 0.695, the r^2 of 0.483 reduces the amount of “unexplained variance” in reading

348

CORRELATION



Figure 11.11 The effect size of correlation—explaining variance.

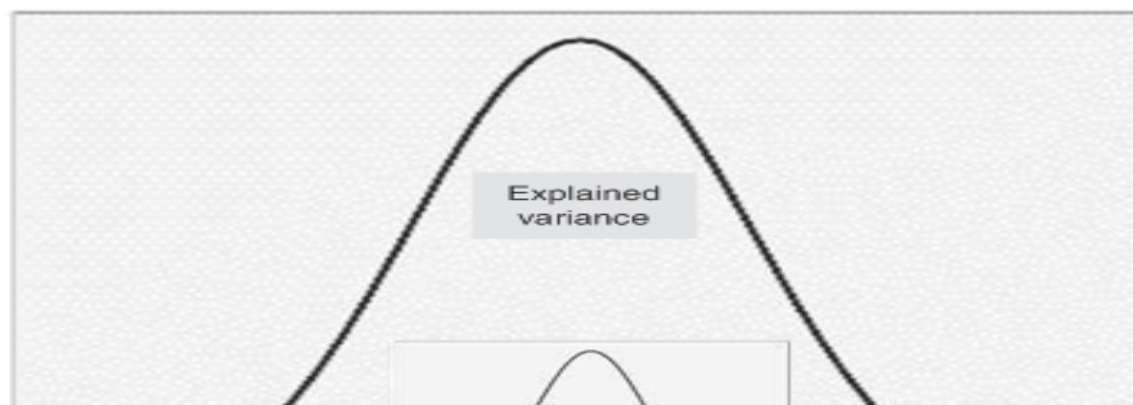




Figure 11.12 The effect size components produced by correlation.

to $(1.00 - 0.483 = 0.517)$. Thus, other variables not in the analysis might help further to “explain away” the remaining unexplained variation.

I will discuss further insights into r^2 in the chapters on regression. *For now you can understand it as the percentage of the variance in one variable (outcome) contributed by another (predictor).* Knowing how much variance is explained is helpful to the

DIAGNOSTICS: CORRELATION PROBLEMS

349

researcher, but what guidelines exist that might help judge the extent of the explained variance? Cohen (1988) provides the following conventions for the r^2 :

- Small effect size: $r^2 = 0.01$ ($r = 0.10$)
- Medium effect size: $r^2 = 0.09$ ($r = 0.30$)
- Large effect size: $r^2 = 0.25$ ($r = 0.50$)

These r^2 size conventions may appear to be small, and for some studies they may not be appropriate. The researcher must ultimately decide the meaningfulness of the correlation effect size since it is tied to the nature of the research question. Thus, I might not be excited about an r^2 of 0.09 in a small exploratory study of the relationship between income and health among electronics workers, but I might be very excited about the same effect size in a study of this relationship in a large national study relating total wealth measures to specific health behaviors.

The r^2 is one of the PRE methods I mentioned at the opening of this chapter. If you look at Figure 11.1, r^2 is listed in the last column with the other PRE methods that I discussed in the ANOVA chapters. All these are related measures in that they focus on explaining variance. I will develop this notion further in discussing regression methods.

DIAGNOSTICS: CORRELATION PROBLEMS

Several factors affect the size of a correlation and therefore its power (i.e., the ability to reject the null hypothesis when it should be rejected). The following are several such factors.

Correlations and Sample Size

As I stressed in previous chapters, hypothesis tests establish whether a finding is likely a chance finding. (We just discussed the effect size, which is distinguished from the statistical rejection of the null hypothesis.) Be aware that *the sample size dramatically affects the size of Pearson's r* .

For example, with 15 cases, the researcher can reject the null hypothesis (two-tailed test at 0.05) with $r = 0.514$. With a sample of 30, the researcher can reject the null hypothesis when $r = 0.361$. With a sample of 102, the researcher can reject the null hypothesis at 0.195. It only takes an $r = 0.098$ to reject the null hypothesis with $N = 402$! *Therefore, with the same study variables, but simply larger samples, the researcher will reject the null hypothesis more easily.*

Correlation Is Not Causation

I noted this injunction previously. The problem with discovering correlations is the temptation to assume that the variables are *causally* related. Thus, you might be tempted to assume that in the example study I have been using, income opinions cause health opinions to increase. A correlation cannot make this claim. There are

350

CORRELATION

many things that are related to opinions about one's health other than opinions about one's income – and some of them may be better explanations of increases in scores. What are some? If you were to list other *potential* causes, you could probably produce a long list (e.g., health behavior practices, diet, exercise, genetic predispositions, etc.).

Restricted Range

The strength and size of a correlation is dramatically affected by “restricted range,” or the selection of scores that do not display full variability. Figure 11.13 shows this problem with a scattergram. If a study sample is relatively homogenous on the variables (e.g., all the subjects appear to be similar in IncomeOP and HealthOP), the resulting correlation can appear to be very low. In the top panel of Figure 11.13, respondents of all different income and health opinions were plotted, indicating a strong correlation. If the study were to be restricted to respondents who were very similar to one another, like those respondents enrolled in health clubs, the resulting correlation would be less pronounced and perhaps not significant. The power of a correlation is related to the variability of the study variables.

Extreme Scores

“Outlier” scores can have a dramatic effect on the calculated r . Often, deleting only a single extreme score has a surprisingly marked effect on the results. With correlation, extreme scores, especially in studies with small sample sizes, can be problematic.

I offer the same suggestion with correlation studies as with studies using other statistical procedures. The researcher must use their understanding of the data and

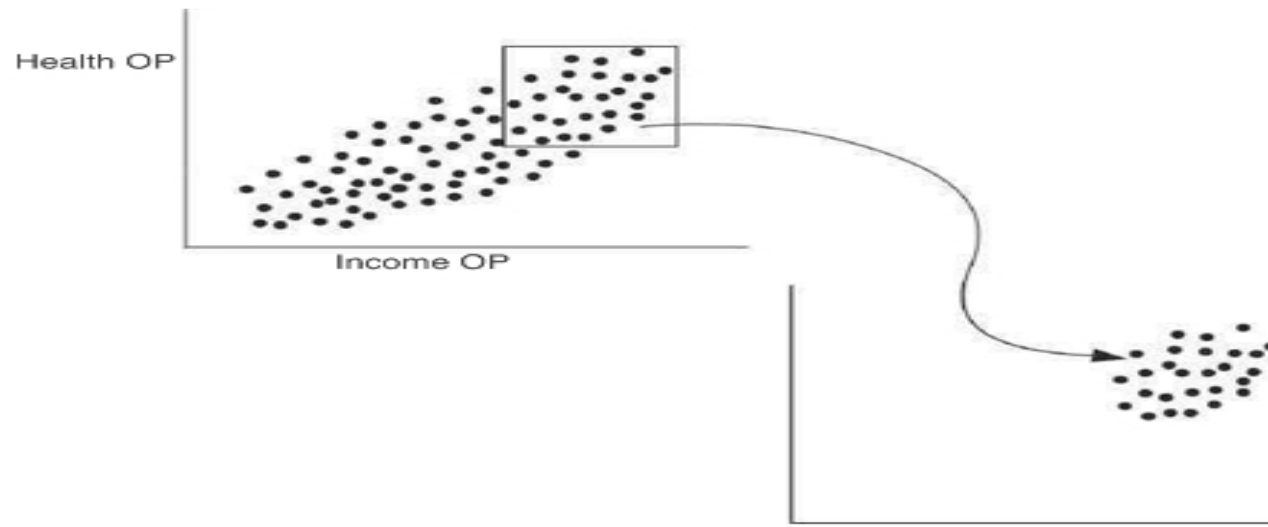


Figure 11.13 The correlation problem of restricted range.

DIAGNOSTICS: CORRELATION PROBLEMS

351

the research situation in order to manage extreme scores. They can transform such scores, delete them, or retain them as legitimate depending on their understanding.

The first step in limiting the effects of extreme scores is to detect them! In other works (e.g., Abbott, 2010), I devote significant attention to diagnostic procedures for correlation and regression studies. The serious researcher should seek these works out before engaging in a correlation study.

Heteroscedasticity

This essentially refers to violating the equal variance assumption among study variables. Homoscedasticity is the condition in which an outcome variable has equal variation across the levels of the predictor variable. If this condition is not present as

variation across the levels of the predictor variable. If this condition is not present (as, for example, if the scores of one study variable were markedly skewed), the correlation would show a distortion (heteroscedasticity). Figure 11.14 shows the two conditions of homoscedasticity (top panel) and heteroscedasticity (bottom panel). The top panel of Figure 11.14 shows that the values of one variable are distributed evenly across the values of the other variable (homoscedasticity). The bottom panel shows unevenness of the values of one variable across the other variable (heteroscedasticity) indicated by the different sized circles encompassing the scatter of dots on the graph. In some cases, heteroscedasticity may result in a departure from linearity.

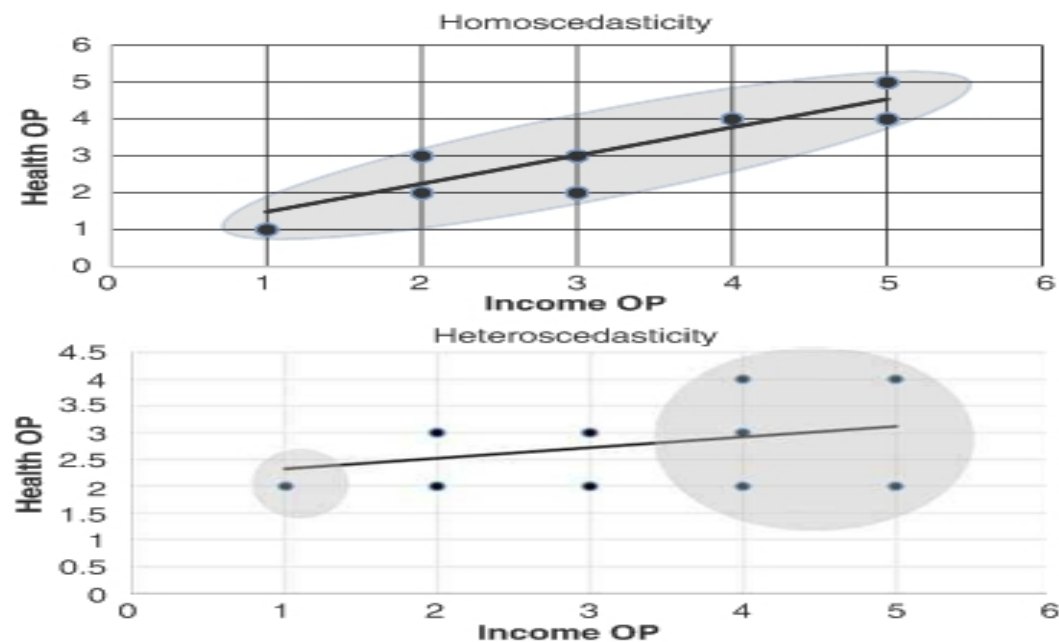


Figure 11.14 The correlation conditions of homoscedasticity and heteroscedasticity.

CORRELATION USING SPSS® AND EXCEL

353

Descriptives				Statistic	Std. Error
HEALTH OPINION	Mean			2.92	.313
	95% Confidence Interval for Mean	Lower Bound		2.23	
		Upper Bound		3.61	

OPINION OF FAMILY INCOME	5% Trimmed Mean		2.91	
	Median		3.00	
	Variance		1.174	
	Std. Deviation		1.084	
	Minimum		1	
	Maximum		5	
	Range		4	
	Interquartile Range		2	
	Skewness		.192	.637
	Kurtosis		.219	1.232
	Mean		3.00	.348
	95% Confidence Interval for Mean		Lower Bound	2.23
			Upper Bound	3.77
	5% Trimmed Mean		3.00	
	Median		3.00	
	Variance		1.455	
	Std. Deviation		1.206	
	Minimum		1	
	Maximum		5	
	Range		4	
	Interquartile Range		2	
	Skewness		.373	.637
	Kurtosis		-.160	1.232

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
HEALTH OPINION	.219	12	.115	.939	12	.487
OPINION OF FAMILY INCOME	.250	12	.037	.910	12	.212

a. Lilliefors Significance Correction

Figure 11.15 The SPSS® descriptive output for the study variables.

Figures 11.17 and 11.18 show the histograms for the outcome variable HealthOP. In both cases, the histograms indicate distributions that appear normal, despite small sample sizes.

Variances are equal: You can assume equal variance since neither variable is *markedly* skewed.

Linear relationship: Figure 11.2 shows the SPSS scattergram between the example study variables. Figure 11.19 shows the Excel scattergram between Math and FR. As you can see, the pattern of the dots is fairly evenly distributed around the line, and the overall shape of the dots is in a straight line.

	<i>IncomeOP</i>	<i>HealthOP</i>
Mean	3.00	2.92
Standard Error	0.35	0.31
Median	3.00	3.00
Mode	3.00	3.00
Standard Deviation	1.21	1.08
Sample Variance	1.45	1.17
Kurtosis	-0.16	0.22
Skewness	0.37	0.19
Range	4.00	4.00
Minimum	1.00	1.00
Maximum	5.00	5.00
Sum	36.00	35.00
Count	12.00	12.00

Figure 11.16 The Excel descriptive statistics.

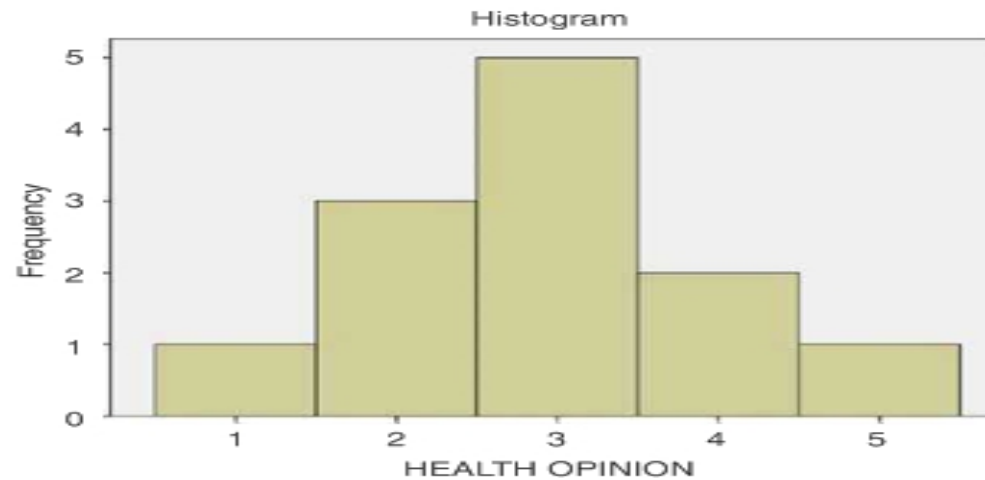


Figure 11.17 The SPSS® histogram for HealthOP.

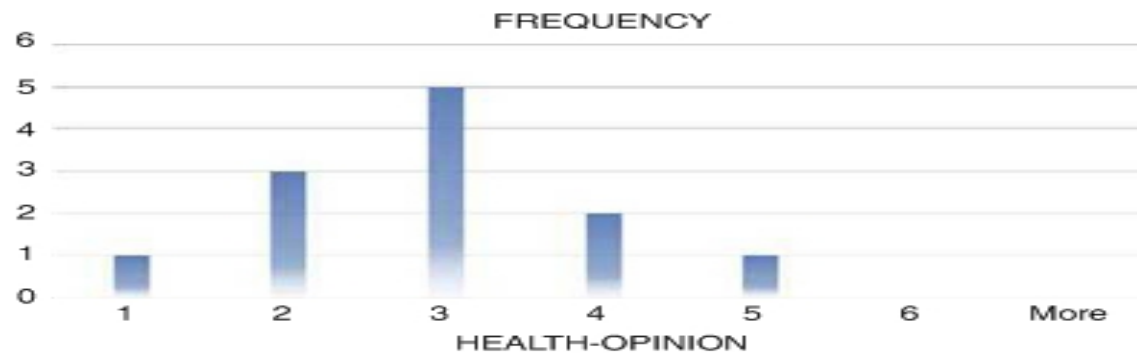
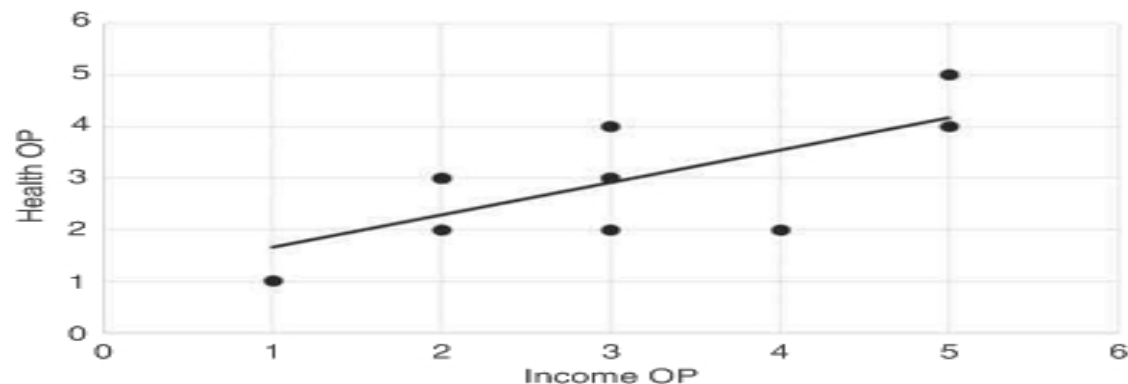


Figure 11.18 The Excel histogram for HealthOP.

CORRELATION USING SPSS® AND EXCEL

355

**Figure 11.19** The Excel scattergram between IncomeOP and HealthOP.**Computation of Pearson's r for the Example Data**

See the manual calculations for the example data in the section "THE COMPUTATION METHOD."

Correlation Using SPSS®

Figure 11.20 shows the SPSS correlation menu. Note that it is called "Bivariate" because you use the program to specify a two-variable correlation ("bi" – two). When you make this choice, you are presented with the specification menu shown in Figure 11.21.

As you can see in Figure 11.21, I included both variables in the "Variables:" window. The default values for this specification window are shown. The analysis will produce Pearson's r (this is the only value produced in the default analysis), and SPSS will create a ("Two-tailed") hypothesis test and "Flag significant correlations" according to the default checks.

If you select the "Options" button in the upper right part of the window, you can request means, standard deviations, and cross-product values. The latter are helpful if you wish to compare the values with those you calculated manually. For this example, I did not select the options.

Choosing "OK" will produce the correlation matrix shown in Figure 11.22. The outputs of the SPSS matrix are helpful because they provide the values you need for the hypothesis test. As you can see, the correlation significance level (" $p = 0.012$ ")

is listed, which indicates that the correlation is in the exclusion area and therefore considered a significant correlation. The calculated r fell so far into the exclusion region that it has an extremely small probability of representing a chance finding.

As with other output I presented, SPSS lists the *exact probability* of the finding. When you create hypotheses tests using correlation tables to specify exclusion values, you would report significance as " $p < 0.05$," indicating that the calculated value fell into the exclusion area beyond the boundary of the 0.05 level of probability.

356

CORRELATION

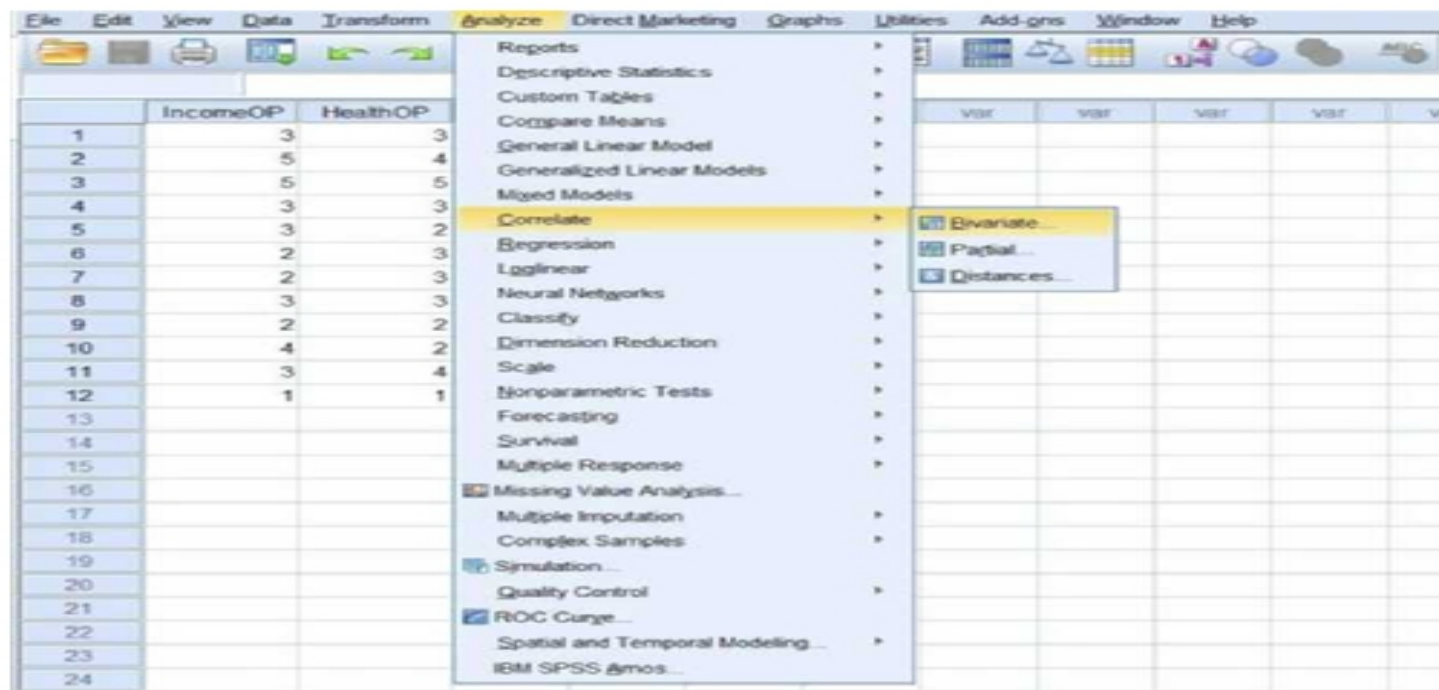
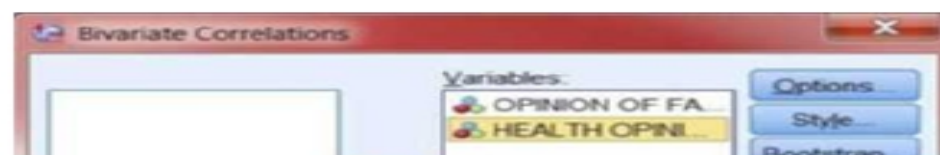


Figure 11.20 The SPSS[®] Correlation menu.



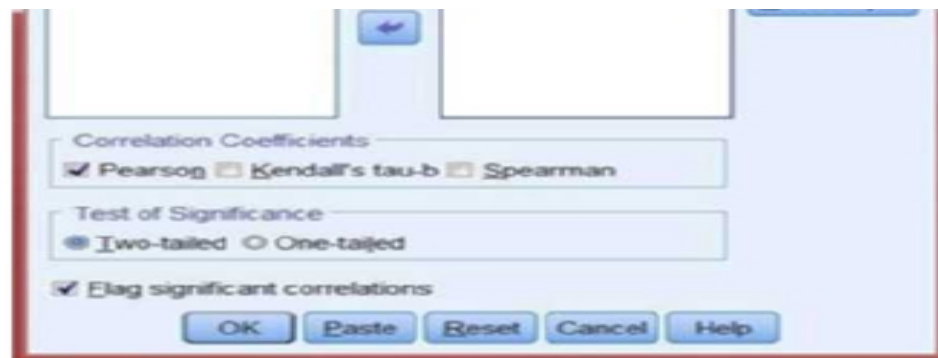


Figure 11.21 The Correlation specification window.

CORRELATION USING SPSS® AND EXCEL

357

Correlations		OPINION OF FAMILY INCOME	HEALTH OPINION
OPINION OF FAMILY INCOME	Pearson Correlation	1	.696*
	Sig. (2-tailed)		.012
	N	12	12
HEALTH OPINION	Pearson Correlation	.696*	1
	Sig. (2-tailed)	.012	
	N	12	12

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

Figure 11.22 The SPSS® correlation matrix.





Figure 11.23 The “Correlation” window in the Excel Data – Data Analysis menu.

Correlation Using Excel

Figure 11.23 shows the menu window when you select “Data – Data Analysis” from the main menu in the spreadsheet where the data are located. This should be a familiar menu now that you have used it in past procedures. When you select “Correlation” you are presented with another specification window as shown in Figure 11.24. As you can see, I selected both columns of data at the same time in the “Input Range:” window. Since I included the labels, I made sure to check the “Labels in First Row” box. The default is to create the data in a separate spreadsheet.

When you select “OK” from the button shown in Figure 11.24, you will see the correlation results in a matrix form, even though there is only one correlation pair. Figure 11.25 shows the IncomeOP–HealthOP correlation as 0.696, which matches the manually calculated value.

Note that Excel does not specify the p value of the result. You must consult the correlation table of values to obtain the comparison value for the hypothesis test.

358

CORRELATION



Figure 11.24 The “Correlation” specification window in Excel.

IncomeOP	HealthOP
----------	----------

IncomeOP	1
HealthOP	0.695608

Figure 11.25 The Excel correlation matrix.

NONPARAMETRIC STATISTICS: SPEARMAN'S RANK ORDER CORRELATION (r_s)

I have not discussed nonparametric correlation procedures thus far, but I listed some examples in Table 11.1. Those listed in the first three columns are nonparametric procedures designed to measure correlations with other than interval data. I will briefly describe one of these in this chapter, Spearman's Rho.

Up to now, I have discussed using two interval-level variables for calculating Pearson's r . However, researchers often do not have access to interval data. For example, researchers might be limited to discussing hospital rankings as a way to discuss quality care. If this is the case, these rank-ordered data are ordinal data, and you can make no assumption that the interval between the rankings is equal.

If you use ordinal data, you can use a variation of correlation called Spearman's r . Also called Spearman's rank order correlation, or Spearman's Rho, this procedure calculates a correlation *using ordinal data or interval data that have been ranked*.

Spearman's Rho results vary between -1 and $+1$ as with Pearson's r . You can obtain a negative value, which is interpreted as it is with Pearson's r . However, since this not a parametric statistic, you cannot square the r_s value to arrive at a "variance explained" figure as in Pearson's r . You can perform the hypothesis test to see whether the value is significant.

The manual calculation for Spearman's Rho is as follows:

$$r_s = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

NONPARAMETRIC STATISTICS: SPEARMAN'S RANK ORDER CORRELATION (r_s)

359

The formula appears somewhat strange in comparison to the other procedures I have presented. Like some nonparametric tests, the Spearman's Rho formula includes a *constant* in the formula (a constant is a number that does not change). The "6" appears mysterious, but the nonparametric formula is created to yield a value that fits the $-1/+1$ range, and the value of 6 allows the computed values to fall in this range. There is a longer explanation, but this gives you some insight into nonparametric procedures.

The preceding formula is used with two sets of ranks like the hypothetical example of data in Table 11.4.

If you use the formula I included earlier, the following is the calculation for Spearman's Rho:

$$\begin{aligned}
 r_s &= 1 - \frac{6 \sum d^2}{N(N^2 - 1)} \\
 &= 1 - \frac{6(8)}{5(25 - 1)} \\
 &= 1 - \frac{48}{120} \\
 &= 0.60
 \end{aligned}$$

Spearman's Rho calculations can be part of hypotheses tests using the same process as the hypothesis tests in the parametric procedures. The null and alternative hypotheses state that the correlation will be 0 (or not 0, respectively) in the population the sample data represent. Spearman's Rho is calculated with the previous formula, and the researcher can compare the calculated value to a critical test value.

Spearman's Rho has a special table of comparison values to use for hypothesis tests. In this table, *the researcher does not use degrees of freedom* since no parameters are estimated. However, if a table of values is not available (I do not include such a table in this book), the researcher can use the *T* table of values with the same formula I noted earlier for Pearson's *r*. *When you use the T distribution table for identifying a comparison value, you can use degrees of freedom as you did with Pearson's r* since

TABLE 11.4 Hypothetical Correlation of Hospital Rankings by Patient Satisfaction Ranking

Hospital Ranking	Patient Satisfaction Ranking	d	d ²
1	2	-1	1
2	1	1	1
3	5	-2	4
4	3	1	1
5	4	1	1
Totals		0	8

the table is created for a parametric comparison:

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

$$\begin{aligned}
 t &= \frac{r_s}{\sqrt{1 - r_s^2}} \\
 &= \frac{0.6\sqrt{3}}{\sqrt{0.64}} \\
 &= 1.30
 \end{aligned}$$

Using the T table of values, I note that the critical value of rejection at the 0.05 (two-tailed) level with $df = 3$ ($5 - 2$) is 3.182. Therefore, since the calculated t value (from Spearman's Rho) is 1.30, you cannot reject the null hypothesis.

The one-tailed 0.05 exclusion value using the Spearman's table of critical values for 5 data pairs is 0.90. (Spearman's table does not provide a two-tailed value with $N = 5$.) You would therefore need a 0.90 correlation or higher to reject the null hypothesis. Using either table would therefore indicate the same result: you cannot reject the null hypothesis with this brief example.

Variations of the Spearman's Rho Formula: Tied Ranks

Researchers oftentimes use rank order procedures when they have data that are interval, but the data are not normally distributed, as I mentioned earlier when I discussed skewed variables. Situations in which this is common involve using financial data, like housing values, physician salaries, or hospital funding. I will provide an extended example of this later.

More commonly, researchers have studies in which they may have *both* ordinal- and interval-level variables. In either case, however, the simple Spearman's Rho formula earlier will most likely not be appropriate as I will discuss later. The reason is that larger studies with more cases or studies that involve ranking of interval variables usually result in *tied ranks*, which are problematic to the "basic" Spearman's Rho formula.

If skewed interval-level variables are used along with ranked variables, *the skewed interval data must first be ranked and then correlated with variables already ranked*. In these cases, the Spearman's Rho formula earlier should not be used.

In the fictitious study earlier, I used hospital ranking as one of the ordinal (ranked) variables. What if I had used the number of positive citations the hospitals had received *instead* of the ranking variable? I would then have an interval-level variable. In such a case, I would need to rank order the interval data so that both of the study variables would be ordinal (remember, the other variable, patient satisfaction ranking was already an ordinal variable). Table 11.5 shows the comparison of these variables and how the interval-level variable can be changed to rankings.

I placed the original ranked variable in the first column. The second column is the interval-level variable used in place of the original ordinal variable (hospital ranking) in the study *instead* of the ranked variable. The problem with these data

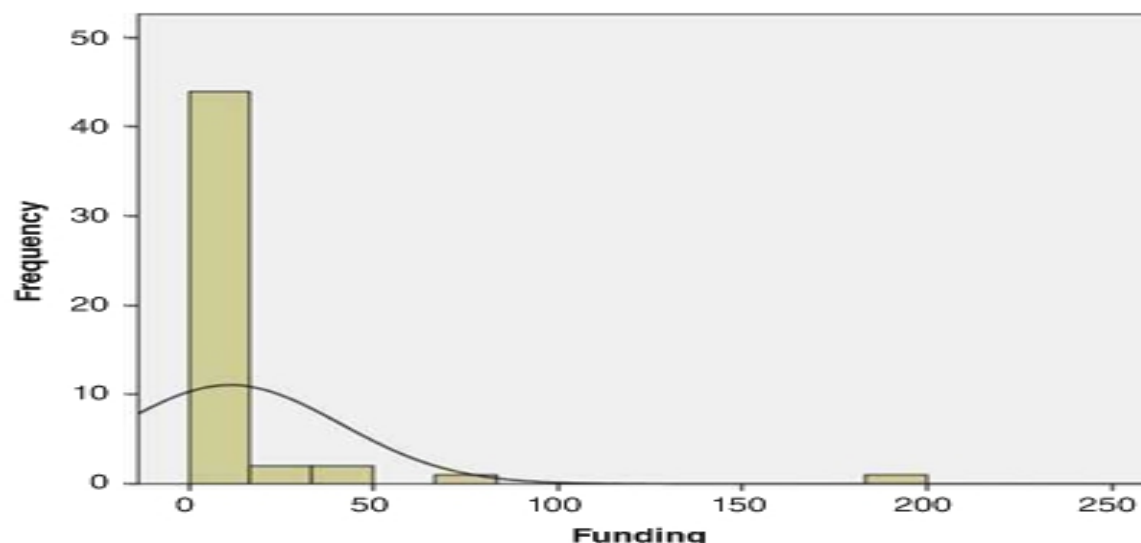


Figure 11.26 Funding for study hospitals.

whether measured by Pearson's r or Spearman's Rho, can only establish relationship, not causality. One of the reasons for this potential problem in interpretation is that a great many factors (in addition to funding) might have an impact on health outcomes. More complex associational designs, like the ones discussed in the next two chapters, would help the researcher to speak more precisely about the specific contribution of funding to health outcomes.

As I noted earlier, any variable measuring income, funding, or wealth is always suspect for being normally distributed because in real life funding is not normally distributed across society. This presents a problem for measuring correlation using Pearson's r . In such cases, researchers might use Spearman's Rho because this procedure can be used effectively with skewed data. Creating ranks with these kinds of variables would be very tedious, especially with large data sets. Fortunately, SPSS includes Spearman's procedure that can be very helpful.

Figure 11.26 shows the SPSS histogram for funding received by the study hospitals, indicating that this variable is markedly skewed. Under these circumstances, with the example study question, the researcher might seek an alternative to Pearson's correlation despite the fact that the data are interval (even ratio) level data. If the researcher has interval data that depart this far from the assumptions of parametric tests to the extent they cannot be reasonably used, then Spearman's Rho can be used instead of Pearson's r .

Figure 11.27 shows the Spearman's r results for this research question. As you can see, the correlation (-0.324) is significant at $p = 0.022$. (Figure 11.21 shows how to obtain this SPSS output. There is simply a box to check for Spearman's correlation

in the same specification window in which you call for Pearson's r .)

TERMS AND CONCEPTS

363

Correlations			Funding	HealthOutcome
Spearman's Rho	Funding	Correlation Coefficient	1.000	-.324 [*]
		Sig. (2-tailed)	.	.022
		N	50	50
	HealthOutcome	Correlation Coefficient	-.324 [*]	1.000
		Sig. (2-tailed)	.022	.
		N	50	50

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

Figure 11.27 The Spearman's Rho correlation between the study variables.

Correlations			
		Funding	HealthOutcome
Funding	Pearson Correlation	1	-.135
	Sig. (2-tailed)		.350
	N	50	50
HealthOutcome	Pearson Correlation	-.135	1
	Sig. (2-tailed)	.350	
	N	50	50

Figure 11.28 The Pearson's r correlation between the study variables.

Figure 11.28 shows the same correlation using Pearson's r . As you see, $r = -0.135$ and is not significant ($p = 0.350$). In this example, a more meaningful outcome may be achieved by using the appropriate statistical procedure (Spearman's Rho). Hospitals with greater funding correlate inversely with the number of complaints about the funded procedures. Again, the researcher cannot make causal attributions from these findings, but the outcome generally speaks to at least one facet of the wealth–health relationship using Spearman's Rho.

TERMS AND CONCEPTS

Aggregate data:

Aggregate data are those data gathered on large-scale social units such as states, counties, industries, and the like, which combine data from smaller analytical levels (e.g., individual opinions, individual salaries, etc.).

364

CORRELATION

Coefficient of determination:

The squared value of r , the Pearson correlation coefficient, which represents the proportion of variance in the outcome variable accounted for by the predictor variable of a study.

Dichotomized variables:

Continuous scores that have been transformed into variables with only two values (e.g., high or low). This is related to but not the same as “dichotomous variables” which are those variables that naturally have two categories (e.g., male or female).

Heteroscedasticity:

A violation of the assumption of equal variation. In this condition, the variance of one study variable is not equal at each value of the other study variable.

Homoscedasticity:

The assumption of equal variance in a study in which the variance of one variable is equal at values of the other variable in a study.

Linear relationship:

The relationship among study variables that forms a straight line if plotted on a graph. Violations of linearity might take the form of “curvilinear” relationships in which the graphed line is not straight but curved.

Outliers:

Extreme scores in a distribution that may result in a distortion of values of the total set of scores.

Proportional reduction in error (PRE) measures:

Procedures that measure how much variation is reduced or explained by knowing the relationship among study variables. In correlation, r^2 is such a measure in that it describes the proportion of variance in the outcome variable that is accounted for by the predictor variable.

Restricted range:

A problem in correlation studies in which the entire set of scores are not used in an association,

Scattergram:

but rather a selected group of the scores is used that that do not represent the total variability. A graph that shows the relationship between study variables when they are plotted together. Also called “scatter diagram,” “scatterplot,” or “scatter graph.”

DATA LAB AND EXAMPLES (WITH SOLUTIONS)**Problem 1**

A research professor wanted to check the relationship between the score on a quiz and the minutes it took students to complete the quiz. Does time matter to test scores? Table 11.6 presents the data for this problem.

DATA LAB: SOLUTIONS**365**

TABLE 11.6 Test Taking Minutes and Test Score Data

Minutes	Test Score
16	19
3	17
6	18
12	19
7	16
16	20
12	19
15	20
8	17
9	19

Problem 2

Earlier in this book I introduced data from a partial housing survey I conducted some years ago that focused on neighborhood characteristics. (Table 3.4 presents partial data from this survey.) In this problem, I present a sample of residents’ survey responses to two neighborhood characteristics: job opportunities and community involvement. Are these variables related?

Table 11.7 presents the data for this problem. Conduct a correlation analysis and

summarize your responses using SPSS. Be sure to discuss whether the assumptions were met for this procedure.

DATA LAB: SOLUTIONS

Problem 1

Findings

The minutes taking the test and the test score are positively correlated with these data. The longer the time taking the test, the higher the score:

- Pearson's $r = 0.81$
- Critical value = 0.632
- Reject H_0 , $p < 0.05$

Manual calculations

The SS_{XY}

$$SS_{XY} = \sum XY - \frac{(\sum X)(\sum Y)}{N}$$

$$SS_{\text{Minutes} \times \text{Test Scores}} = 1958 - \frac{(104)(184)}{10}$$

$$SS_{\text{Min.} \times \text{Scores}} = 44.4$$

TABLE 11.7 The Job Opportunity–Community Involvement Study Data

Job Opportunities	Community Involvement
1	1
1	3
1	1
1	1
2	4
2	1
2	1
2	2
2	4
2	3
3	2

3	4
3	5
3	4
4	4
4	3
4	4
4	3
4	4
4	6
5	5
5	5
5	7
6	5
7	7
7	7
7	6
7	7

The SS_X

$$SS_X = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$SS_{\text{Minutes}} = 1264 - 1081.6$$

$$= 182.4$$

DATA LAB: SOLUTIONS

367

The SS_Y

$$SS_Y = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

$$SS_{\text{Scores}} = 3402 - 3385.6$$

$$= 16.4$$

Pearson's r

$$r_{vv} = \frac{SS_{XY}}{SS_X}$$

$$\begin{aligned}
 &= \frac{44.4}{\sqrt{(182.4)(16.4)}} \\
 &= \frac{44.4}{54.69} \\
 &= \mathbf{0.81}
 \end{aligned}$$

Problem 2

The elements of the solution for this problem include an examination of assumptions, the hypothesis test, and effect size. The database I used is only to illustrate the correlation procedure and should not be taken to indicate real-world findings.

Assumptions

1. Randomly chosen sample. I chose the respondents from the database (of about $N = 140$) to illustrate the correlation procedure.
2. Variables are interval level. Data are from respondent survey that I treat as interval.
3. Variables are independent of one another. Respondent data were not dependent in this study.
4. Variables are normally distributed.
The SPSS outcome shown in Figure 11.29 (from “Analyze – Compare Means – Means” menus) indicates that the two variables are normally distributed based on skewness and kurtosis. The histograms in Figure 11.30 (Community Involvement) and Figure 11.31 (Job Opportunities) show some skewness in Job Opportunities but “somewhat” normal distribution in Community Involvement, findings indicated by the Shapiro–Wilk test of normality shown in Figure 11.32. Despite minor irregularities in the normal assumption, Pearson’s r is robust and can provide meaningful results.
5. Variances are equal. Pearson’s r is robust for these violations unless one or both variables are significantly skewed. Although the results indicate some skewness in

Report

	Job	Community
--	-----	-----------

	Job Opportunities	Community Involvement
Mean	3.61	3.89
N	28	28
Std. Deviation	1.950	1.969
Kurtosis	-.806	-.912
Std. Error of Kurtosis	.858	.858
Skewness	.439	.035
Std. Error of Skewness	.441	.441

Figure 11.29 Descriptive findings for study variables.

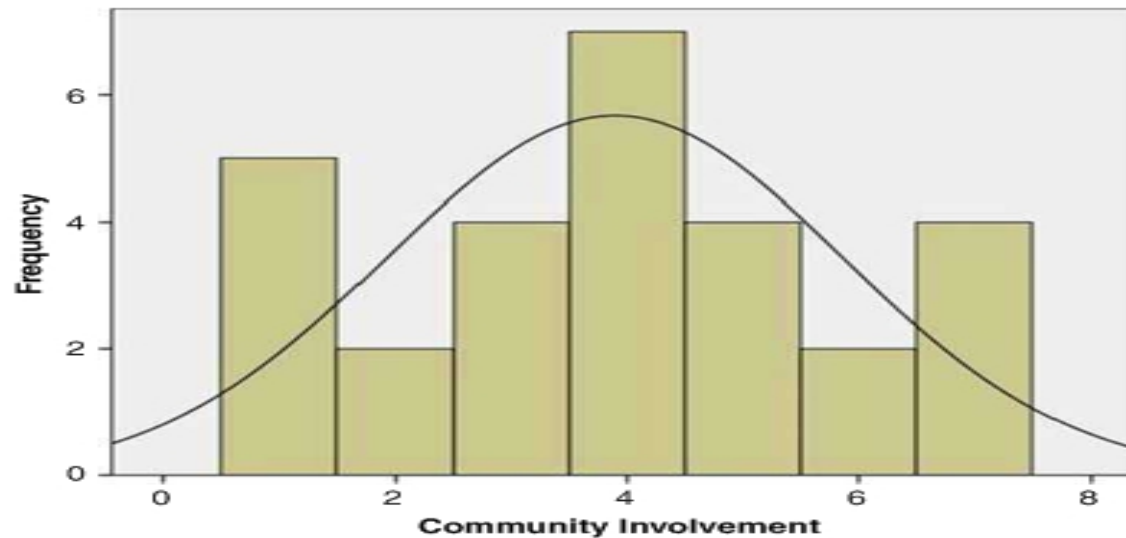


Figure 11.30 The histogram for community involvement.

one variable (Job Opportunities), I will proceed with this study problem. The scattergram shown in Figure 11.33 is partial evidence of bivariate normality, discussed in the next assumption.

- 6. Linear relationship.** The two variables must display a “straight line” when plotting their values. Violations of this assumption might include “curvilinear” relationships in which plotted data appear to be in the form of a “U” pattern. Figure 11.33 shows the scattergram for the study variables, the data of which appear to be uniformly distributed around the regression line. In a later chapter, I show how to detect “curvilinear” and other relationships through SPSS. Using this example, those tests and the visual examination of the scattergram provide evidence of linearity.

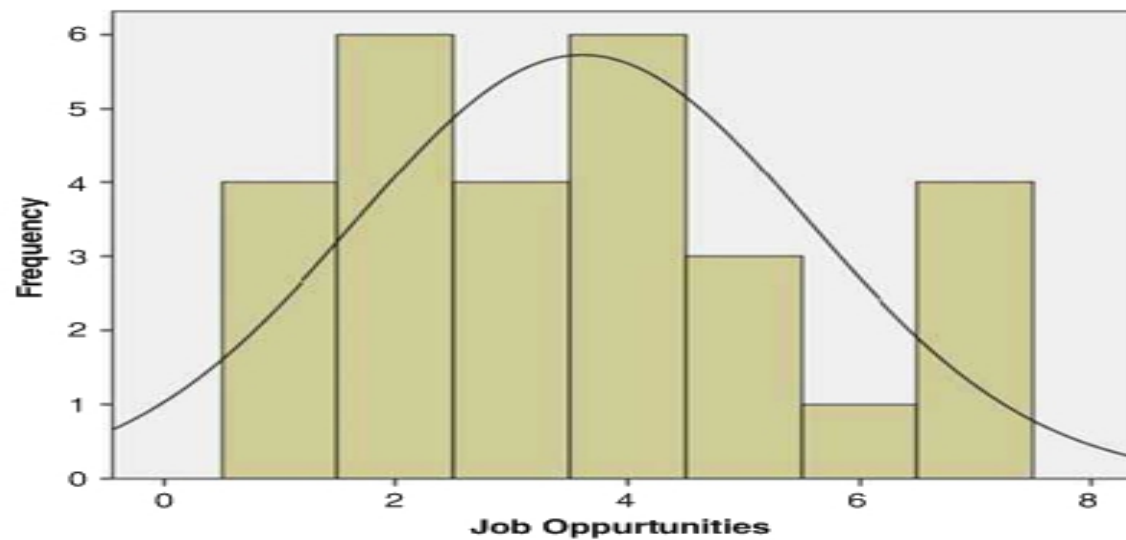


Figure 11.31 The histogram for job opportunities.

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Job Opportunities	.152	28	.095	.912	28	.022
Community Involvement	.129	28	.200 ^a	.924	28	.042

^a. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure 11.32 The tests of normality for the study variables.

Hypothesis Test

1. The null hypothesis (H_0): $\rho = 0$.
2. The alternative hypothesis (H_A): $\rho \neq 0$.
3. The critical value $r_{28(0.05)} = 0.374$.
4. The calculated value (0.874) (see Figure 11.34 for the correlation findings).
5. Statistical decision: Reject the null hypothesis. The calculated correlation falls into the exclusion area and is therefore unlikely a chance finding.

Effect Size

The effect size for this result is $r^2 = 0.76$ ($0.847^2 = 0.764$). The predictor variable (job opportunities) accounts for 76.4% of the variance in the outcome variable

370

CORRELATION

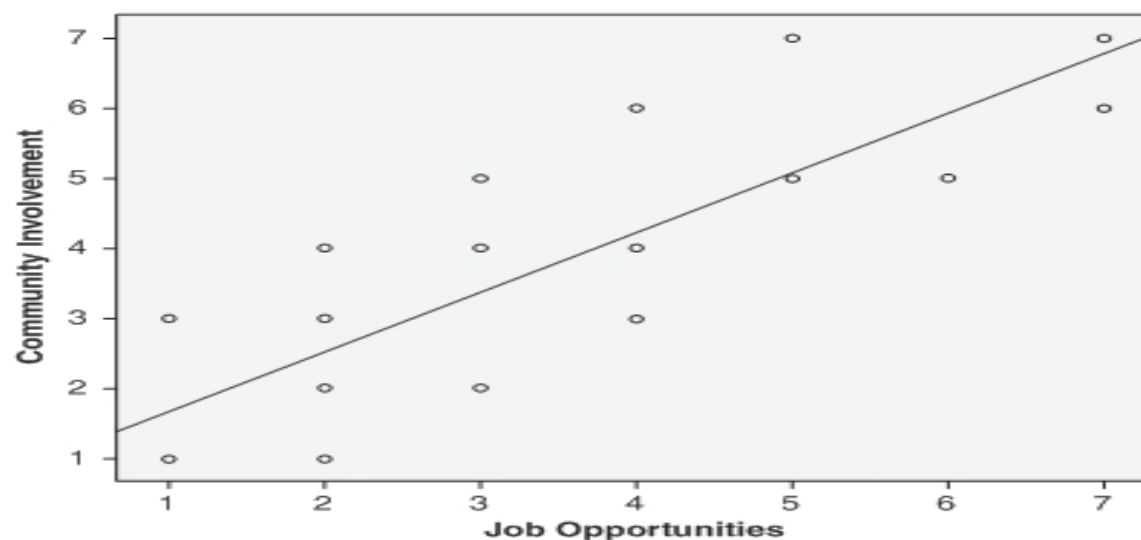


Figure 11.33 The scattergram for the study variables.

Correlations

		Job Opportunities	Community Involvement
Job Opportunities	Pearson Correlation	1	.847**
	Sig. (2-tailed)		.000
	N	28	28
Community Involvement	Pearson Correlation	.847**	1
	Sig. (2-tailed)	.000	
	N	28	28

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

Figure 11.34 The correlation findings for the study variables.

(community involvement). This amount is considered large by the Cohen (1988) criteria.

Discussion

For this example study, respondents indicated that job opportunities in a subsidized housing project were positively correlated to community involvement. The high correlation ($r = 0.847$) indicated that respondent's judged community involvement to be higher with increased job opportunities.