CHAPTER FIVE

Correlation

5.1 Introduction

Correlation is a statistical technique that is used to determine if there is a **relationship** between two variables. For example, it may assess the degree and magnitude of the relationship between students' high school GPA scores and their success in college. If there is a relationship, we say that the two variables are *correlated*.

Correlation techniques measure both the strength and direction of the relationship that exist between two variables; so that a single value will tell us how any two variables are related. This single value is called the *correlation coefficient*, ***r***. When we which to predict scores from one variable, knowing the scores of another, we use another statistical technique called **regression**. So correlation tells us if a relationship exists and regression enables us to use this relationship to predict one variable score, given the score of the other. Table 5.1.1 shows some assessment scores for five individual players.

The correlation coefficient, *r*, ranges from values of -1 to +1. An *r* value of +1 suggests that the two variables are strongly related positively; that is, as the scores of one variable increases, the other also increases (See Figure 5.1.1: Ability and Speed). An *r* value of -1 suggests that the two variables are strongly related negatively; as the scores of one variable increases, the other decreases (See Figure 5.1.2: Ability and GPA). When the *r* value is 0, there is no relationship or no correlation (See Figure 5.1.3).

Table 5.1.1 *Correlation Example Table*

Player	Ability	GPA	Speed	Index
A	1	5	1	3
B	2	4	2	2
C	3	3	3	4
D	4	2	4	2
E	5	1	5	3

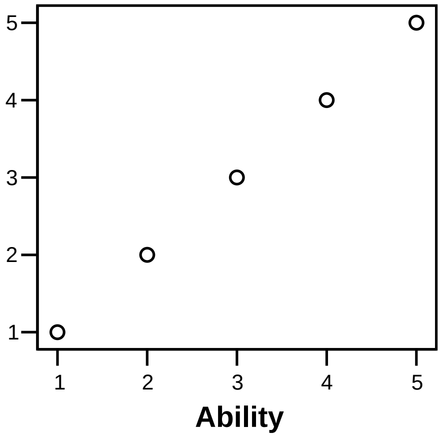


Figure 5.1.1 Positive correlation: Ability vs. Speed.

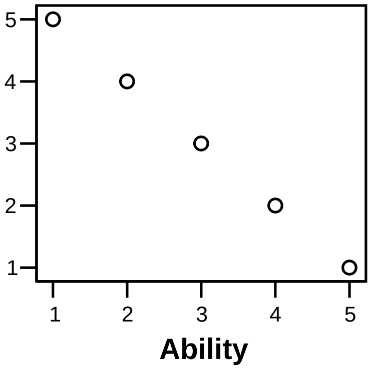


Figure 5.1.2 Negative correlation: Ability vs. GPA.

Figures 5.1.1 through 5.1.3 are examples of **scatterplots**. In a scatterplot we show the *X* and *Y* coordinates of the two variables being examine for their degree of association. We often can often make very crude assumptions, just by looking at the scatterplot, about the direction and degree of the association between variables. The scatterplot also allows us to spot or locate pair of points that maybe outliers (extreme points). If the association between variables is not linear, the scatterplot may provide an early clue of this association. In preparing a scatterplot, we represent the predictor variable (independent variable) on the *X* axis or the *abscissa* and the criterion or dependent variable on the *Y* axis or *ordinate*. We may use regression technique, in later chapter, to draw a "best fit" straight line (the regression line) through the center of the scattered points on the scatterplot.

Correlation is concern about the relationship or association between variables.

Correlation coefficient, ***r*** is a measure of the degree and magnitude of the relationship between variable.

Scatterplot is a figure in which pairs of individual data points are plotted on the *XY* coordinates graph.

Predictor variable is the independent variable from which predictions are made.

Criterion variable is the dependent variable to be predicted.

Regression line is a "best fit" straight line drawn on a scatterplot.

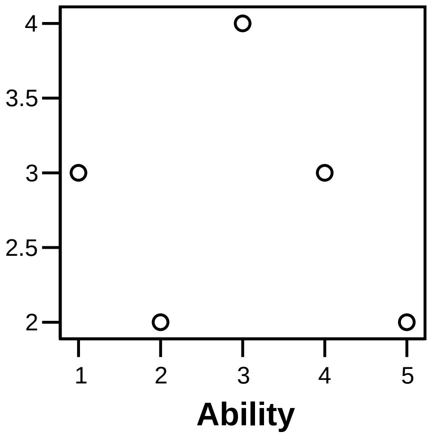


Figure 5.1.3 No correlation: Ability vs. Index.

If the relationship between the variables is linear and/or the variables are interval or ratio scales, we may use the correlation coefficient determined by the Pearson *r*. For nonlinear relationships and/or the variables are ordinal scales, we may use Spearman rank correlation coefficient, r_s.

Pearson Correlation

The most common correlation measurement is the *Pearson correlation* (or Pearson product-moment correlation). The Pearson correlation measures the degree and direction of the linear relationship between two variables.

Three assumptions are made about the Pearson's correlation coefficient, *r*: 1. it requires interval or ratio data, 2. the relationship between variables must be linear, and 3. the technique requires pairs of data values.

The **Pearson correlation** (or **product-moment correlation**) measures the degree and direction of the linear relationship between two variables.

Table 5.1.2 shows the Pearson *r* for the data in Table 5.1.1. This display is often called a correlation coefficient matrix because is shows *r* for more than two variables.

Table 5.1.2 *Correlation Matrix for Correlation Example*

Variable	Ability	GPA	Speed	Index
Ability	1	-1**	1**	0
GPA		1	-1**	0
Speed			1	0
Index				1

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

Covariance

The correlation coefficient is based on a statistics called the covariance. The **covariance** is a statistics that describes the degree which two variables vary together. When high scores of one variable tend to pair with high scores on the other, the covariance will be large and positive (Figure 5.1.1). When high scores of one variable tend to pair with low scores on the other, the covariance will be large and negative (Figure 5.1.2). Finally, when high scores on one variable are paired about the same with both high and low scores on the other, the covariance will be near zero (Figure 5.1.3).

The **covariance** is a statistics that represents the degree that two variables vary together.

Mathematically, the covariance is average sum of the product deviations of the *X* and *Y* variables from their mean. It is given by the formula:

$$\text{cov}_{xy} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n-1} \quad \text{(definition formula)}$$

The computational formula for the covariance is:

$$\text{cov}_{xy} = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{n-1} \quad \text{(computational formula)}$$

Example 5.1.1: Find the covariance of the Verbal and Quant variables in Table 5.1.3. From Table 5.1.3 we compute: ∑XY = 133891, ∑X = 1144, and ∑Y = 1162, so

$$\text{cov}_{xy} = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{n-1} = \frac{133891 - \frac{(1144)(1162)}{10}}{9} = 106.47$$

Table 5.1.3 *Covariance Example*

VERBAL (X)	***QUANT (Y)***	***XY***
108.00	*111.00*	*11988*
133.00	*132.00*	*17556*
109.00	*114.00*	*12426*
118.00	*110.00*	*12980*
94.00	*98.00*	*9212*
111.00	*103.00*	*11433*
107.00	*116.00*	*12412*
125.00	*130.00*	*16250*
120.00	*122.00*	*14640*
119.00	*126.00*	*14994*
$\sum X = 1162$	$\sum Y = 1144$	$\sum XY = 133891$

The magnitude of the covariance is dependent upon the units of measurements for *X* and *Y*. A measurement of either of these variables in inches will give a larger covariance than if the measurements were in feet. The correlation coefficient, however, is not affected by the units of measurement. Conceptually, the Pearson correlation is the covariance of *X* and *Y* divided by the variability of *X* and *Y* separately or the degree to which *X* and *Y* vary together divided by the degree to which they vary separately. This leads to a formula for Pearson *r* as:

$$r = \frac{\text{covariance of X and Y}}{\text{variability of X and Y separately}} = \frac{\text{cov}_{xy}}{s_x s_y} \quad \text{(definition formula)}$$

where s_x and s_y are the standard deviations of *X* and *Y* respectively

So the correlation coefficient (Pearson *r*) for the Example 5.1.1 is **0.86** when we use the covariance and individual standard deviations of *X* and *Y*.

$$r = \frac{\text{cov}_{xy}}{s_x s_y} = \frac{106.47}{10.94(11.28)} = 0.86$$

The computational formula for the Pearson *r* is:

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

The computational Table 5.1.4 shows how to use the computational formula above to compute the correlation coefficient.

Table 5.1. 4 *Pearson r Computational 1*

X	Y	XY	X^2	Y^2
108	*111*	11988	11664	12321
133	*132*	17556	17689	17424
109	*114*	12426	11881	12996
118	*110*	12980	13924	12100
94	*98*	9212	8836	9604
111	*103*	11433	12321	10609
107	*116*	12412	11449	13456
125	*130*	16250	15625	16900
120	*122*	14640	14400	14884
119	*126*	14994	14161	15876
∑X = ***1144***	**∑Y =** ***1162***	**∑XY =** ***133891***	**∑X^2 =** ***131950***	**∑Y^2 =** ***136170***
(∑X)2 = ***1308736***	**(∑Y)2 =** ***1350244***			

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$r = \frac{10(133891) - (1144)(1162)}{\sqrt{[10(131950) - (1308736][10(136170) - 1350244]}} = 0.86$$

Before we introduce another useful computation formula for the Pearson *r*, let us examine a few concepts. The **sum of product deviation**, ***SP***, is a similar concept to sum of square deviation, *SS*. The sum of product can be computed from either the definitional or computation formula shown below.

$$SP = \Sigma(X - M_X)(Y - M_Y) \text{ (definition formula)}$$

$$SP = \Sigma XY - \frac{\Sigma X \Sigma Y}{n} \text{ (computational formula)}$$

Using the sum of square deviations for each variable,

$$SS_X = \Sigma(X - M_X)^2$$

$$SS_Y = \Sigma(Y - M_Y)^2$$

Pearson r is computed by the formula and Table 5.1.5, so *r* = **0.86**. Figure 5.1.4 shows the scatterplot for the *X* and *Y* variables with a best fit the regression line.

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{958.2}{\sqrt{(1076.4)(1145.6)}} = 0.86$$

Table 5.1.5 *Pearson r Computation 2*

Scores		Deviations		Squared Deviations		Products
X	***Y***	***X-M$_X$***	***Y-M$_Y$***	***(X-M$_X$)^2***	***(Y-M$_Y$)^2***	***(5)(6)***
108	***111***	-6.4	-5.2	40.96	27.04	33.28
133	***132***	18.6	15.8	345.96	249.64	293.88
109	***114***	-5.4	-2.2	29.16	4.84	11.88
118	***110***	3.6	-6.2	12.96	38.44	-22.32
94	***98***	-20.4	-18.2	416.16	331.24	371.28
111	***103***	-3.4	-13.2	11.56	174.24	44.88
107	***116***	-7.4	-0.2	54.76	0.04	1.48
125	***130***	10.6	13.8	112.36	190.44	146.28
120	***122***	5.6	5.8	31.36	33.64	32.48
119	***126***	4.6	9.8	21.16	96.04	45.08
				SS$_X$=1076.4	*SS$_Y$*=1145.6	*SP=958.2*

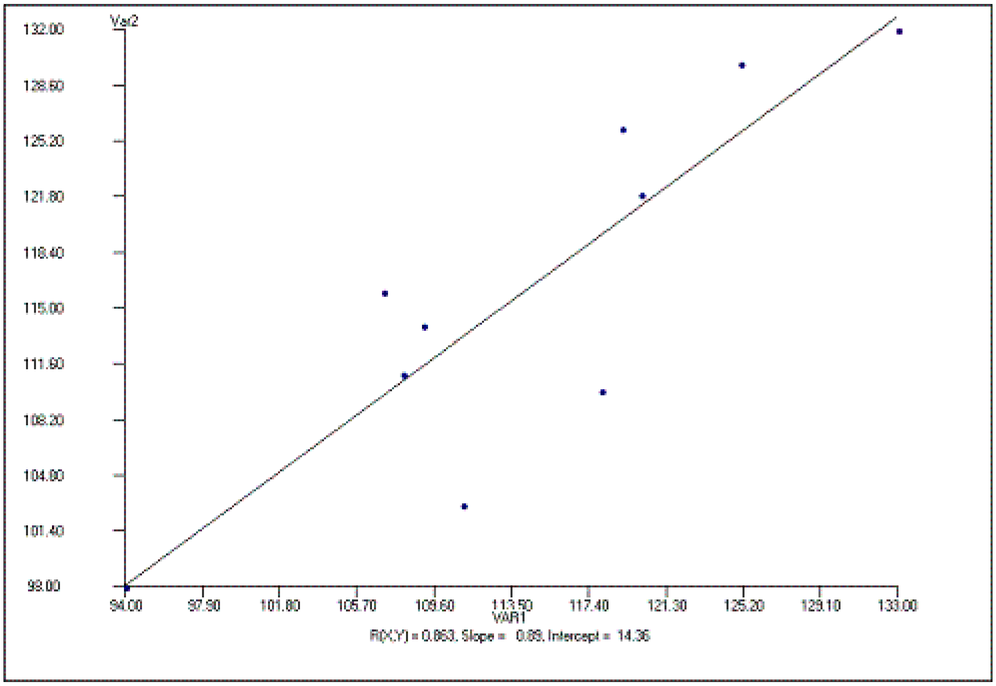


Figure 5.1.4 Scatterplot for the Correlation Example

The **significance of the correlation coefficient**, *r* is dependent upon the sample size and the level of confidence one wishes to have for the correlation coefficient. In the SPSS correlation computational output, this is related to the *p-value* or the significance level. If the **significance level**, *p-value*, is very small (less than 0.05, for 95% confidence), then the correlation is significant and the two variables are linearly related (especially so for the Pearson *r*). If the significance level, *p-value* is very large (or *p* > 0.50) the correlation is *not* significant, and the two variables are *not* linearly related. Testing the significance of the correlation coefficient will be discussed in later chapters. However, most textbooks use the following scheme in Table 5.1.6 to interpret the value of the correlation coefficient as follows:

Table 5.1.6 *Interpretations for Correlation Coefficient*

Correlation Coefficient value	**Interpretation**
>= 0.80	Very Strong
0.60 to 0.80	Strong
0.40 to 0.60	Moderate
0.20 to 0.40	Low
=< 0.20	Very Low

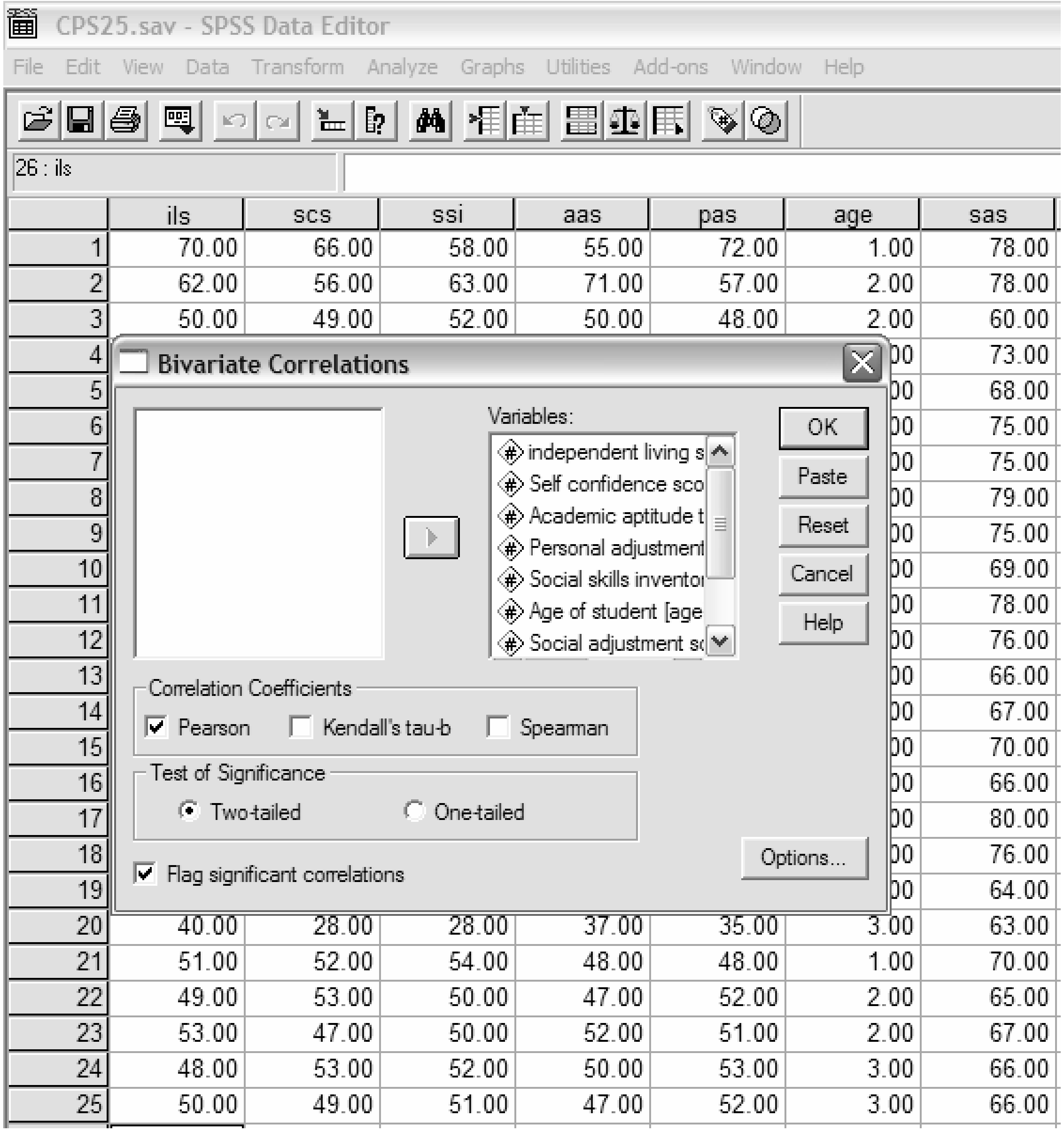
Table 5.1.7 shows a correlation matrix for several variables of the CPS50 database. Various asterisk shows whether there is a relationship between variables and the significant level of each relationship; this will be explain later. Figure 5.1.5 shows the SPSS procedure for computing the Pearson correlation coefficient for two or more variables.

Table 5.1.7

Correlation Matrix for First 25 Data for CPS50

Variables	independent living scale	Self confidence score	Academic aptitude test	Personal adjustment scale	Social skills inventory	Age of student
independent living scale	1	0.774**	0.37	0.623**	0.707**	-0.687**
Self confidence score		1	0.559**	0.764**	0.863**	-0.720**
Academic aptitude test			1	0.552**	0.422*	-0.405*
Personal adjustment scale				1	0.669**	-0.31
Social skills inventory					1	-0.618**
Age of student						1

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).


CPS25.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
26 : ils
Bivariate Correlations
OK
Paste
Reset
Cancel
Help
Options...

	ils	scs	ssi	aas	pas	age	sas
1	70.00	66.00	58.00	55.00	72.00	1.00	78.00
2	62.00	56.00	63.00	71.00	57.00	2.00	78.00
3	50.00	49.00	52.00	50.00	48.00	2.00	60.00
4						00	73.00
5						00	68.00
6						00	75.00
7						00	75.00
8						00	79.00
9						00	75.00
10						00	69.00
11						00	78.00
12						00	76.00
13						00	66.00
14						00	67.00
15						00	70.00
16						00	66.00
17						00	80.00
18						00	76.00
19						00	64.00
20	40.00	28.00	28.00	37.00	35.00	3.00	63.00
21	51.00	52.00	54.00	48.00	48.00	1.00	70.00
22	49.00	53.00	50.00	47.00	52.00	2.00	65.00
23	53.00	47.00	50.00	52.00	51.00	2.00	67.00
24	48.00	53.00	52.00	50.00	53.00	3.00	66.00
25	50.00	49.00	51.00	47.00	52.00	3.00	66.00

Variables:
- independent living s
- Self confidence sco
- Academic aptitude t
- Personal adjustment
- Social skills inventor
- Age of student [age
- Social adjustment sc

Correlation Coefficients

☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance

● Two-tailed ○ One-tailed

☑ Flag significant correlations

Figure 5.1.5 *SPSS correlation procedure: Analyze -> Correlate (Pearson).*

Factors Affecting Correlation

There are many factors to consider when evaluating the correlation between variables. These are some factors that may affect the correlation coefficient: range restriction, nonlinearity of scores, heterogeneous sub-samples, extreme scores, and sample size.

1. Not all the range of variables *X* and *Y* are applicable for all comparisons. It might be more appropriate, for example, to restrict a correlation of height and age to an age range of 4 to 17 to obtain a more meaningful correlation. So when a researcher restricts the range of some variables, the correlation may be restricted to that range limitation.
2. When the distribution is not linear the correlation will be biased, because the premise of Pearson product-moment analysis is based on the data being linear. There are other techniques that have been developed to handle non-linear relationships between variables.
3. Heterogeneous sub-samples is a data distribution that can be subdivided into two or more distinct distributions based on sub-categories within a variable or different variables that have been combined into one distribution. For example, if you combine scores for male and female into one distribution of scores for height, you might get a different correlation if the gender scores are compared against height separately. The combination of gender into one distribution could be an example of a heterogeneous sub-sample.
4. Extreme values or outliers are scores that are too large or small. These large of small scores may increase or decrease the correlations coefficient. A scatterplot or other statistical techniques can help identify these extreme scores that often should not be included in the correlation analysis.

5. The sample size should be large enough to provide a meaningful basis on which to make your comparison between variables; typically the Spearman rank correlation analysis is used instead of the Pearson correlation analysis when the sample size is small.

There are many other types of correlation techniques developed to handle various restrictions on other data types and the nature of the variables being compared. Many of these techniques are modifications of the Pearson product-moment technique. The Spearman rank correlation, for example, calculates the product-moment correlation on ordinal measurements or ranked scores. The Spearman rank correlation will be discussed in the next section of this chapter. Table 5.1.8 shows some other correlation techniques used for various situations involving the type and nature of variables being evaluated.

Table 5.1. 8 *Other Correlation Techniques*

Point-biserial r	One dichotomous variable (yes/no; male/female) and one interval or ratio variable
Biserial r	One variable forced into a dichotomy (grade distribution dichotomized to “pass” and “fail”) and one interval or ratio variable
Phi coefficient	Both variables are dichotomous on a nominal scale (male/female vs. high school graduate/dropout)
Tetrachoric r	Both variables are dichotomous with underlying normal distributions (pass/fail on a test vs. tall/short in height)
Correlation ratio	There is a curvilinear rather than linear relationship between the variables (also called the eta coefficient)
Partial correlation	The relationship between two variables is influenced by a third variable (e.g., mental age and height, which is influenced by chronological age)
Multiple R	The maximum correlation between a dependent variable and a combination of independent variables (a college freshman’s GPA as predicted by his high school grades in Math, chemistry, history, and English)