Chapter 12
Correlation

Pearson Correlation

Sir Francis Galton in Great Britain was interested in studying individual differences based on the work of his cousin, Charles Darwin. In 1869, Sir Francis Galton demonstrated that the mathematics scores of students at Cambridge University and the admissions exam scores at the Royal Military College were normally distributed. In 1889, Francis Galton published an essay suggesting the idea for examining how two traits varied together (covaried). This effort resulted in the first use of the term "regression." Karl Pearson in 1898, based on the suggestions made by Sir Francis Galton, investigated the development of a statistical formula that would capture the relationship between two variables.

The idea was to determine the degree to which two things went together, i.e., how two things varied together. The concept was simple enough in principle, take measurements on two variables, order the measurements of the two variables, and determine if one set of measurements increased along with the second set of measurements. In some cases, maybe the measurements of one variable decreased while the other increased. The basic assumption Karl Pearson made was that the measurements needed to be linear or continuous. He quickly determined that how two things covaried divided by how they individually varied would yield a statistic that was bounded by +1 and −1, depending on the relationship of the two measurements. The conceptual formula he developed, which took into account the covariance between two variables divided by the variance of the two variables, was defined as:

$$r = \frac{\text{Covariance XY}}{(\text{Var X})(\text{Var Y})}$$

In 1927, after L. L. Thurstone developed the concept of a standard score (z-score) as the deviation of a raw score from the mean, divided by the standard deviation, the correlation formula was further defined as the average product of standard scores:

R. Schumacker and S. Tomek, *Understanding Statistics Using R*,
DOI 10.1007/978-1-4614-6227-9_12, © Springer Science+Business Media New York 2013

$$r = \frac{\Sigma z_x z_y}{N}$$

An example of the relationship between two continuous variables will better illustrate how the bivariate (two variable) correlated relationship is established. A typical research question for a group of students can be stated as "Is there a significant relationship between the amount of time spent studying and exam scores?" The data for these two continuous variables, ordered by time spent studying, is listed below.

Time spent	Exam Score
1 h	75
1 h	80
2 h	75
3 h	90
3 h	85
4 h	95
4 h	85
5 h	90
5 h	95
6 h	90

A computational version of the correlation formula makes the calculation easier and uses the following summary values:

$\Sigma X = 34$
$\Sigma X^2 = 142$
$\Sigma Y = 860$
$\Sigma Y^2 = 56800$
$\Sigma XY = 3015$

The computational correlation coefficient formula is:

$$r = \frac{SP}{\sqrt{SS_x SS_y}}$$

The expression, SP, is defined as the sum of cross products for X and Y. The expression, SS_x, is the sum of squares X, and the expression Ss_y is the sum of squares Y.

These values are computed for each expression in the correlation coefficient formula as:

$$SP = \sum XY - \frac{(\sum X)(\sum Y)}{N} = 3015 - \frac{(34)(860)}{10} = 91$$

$$SS_X = \sum X^2 - \frac{(\sum X)^2}{N} = 142 - \frac{(34)^2}{10} = 26.40$$

$$SS_Y = \sum Y^2 - \frac{(\sum Y)^2}{N} = 56800 - \frac{(860)^2}{10} = 490$$

These values are substituted in the Pearson correlation coefficient:

$$r = \frac{SP}{\sqrt{SS_x SS_y}} = \frac{91}{\sqrt{26.4(490)}} = +.80$$

The value of r = +.80 indicates a positive relationship between the two variables implying that as the amount of study time increases, exam scores increase. The correlation coefficient also indicates the magnitude of the relationship since the r-value is approaching +1.0, which would indicate a perfect relationship.

Interpretation of Pearson Correlation

The correlation coefficient can be interpreted in several different ways. First we can test it for significance using tabled correlation values for different sample sizes (degrees of freedom). Second, we can square the correlation coefficient to obtain a variance accounted for interpretation. Third, we can graph the relationship between the data points in a scatter plot to visually see the trend of the relationship.

To test the significance of the correlation coefficient, we use our standard hypothesis testing approach:

Step 1: State the Null and Alternative Hypothesis using Population Parameters.
H_0: $\rho = 0$ (no correlation)
H_A: $\rho \neq 0$ (correlation exists)

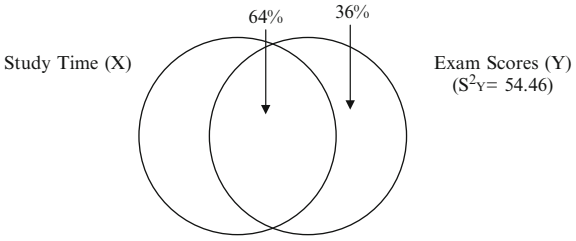
Step 2: Choose the appropriate statistic and state the sample size obtained.
Pearson correlation coefficient for continuous variables
Sample Size, N = 10

Step 3: State the level of significance, direction of alternative hypothesis, and region of rejection.
Level of Significance (α) = .05
Alternative Hypothesis: Non-directional (Two-tailed test)
For N = 10, df = N − 1 = 9
R: $rta_{bled} > .602$

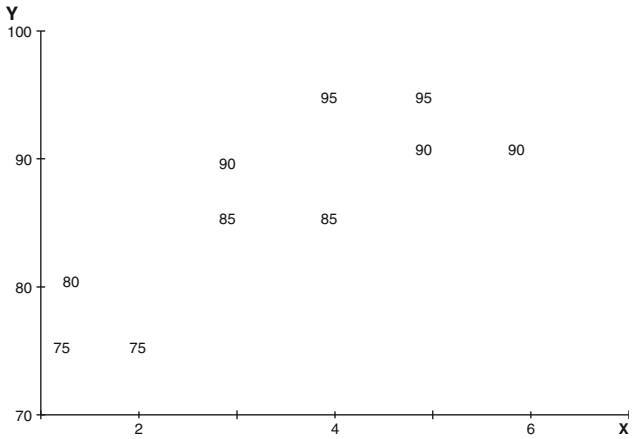
Step 4: Collect Data and Calculate Sample Correlation Statistic.
Continuous Variables: Time spent studying and exam scores
N = 10 pairs of data
r = .80

Step 5: Test statistical Hypothesis, make decision, and interpret results.
Since the computed r = .80 is greater than the tabled r of .602 at the .05 level of significance for a two-tailed test, reject the null hypothesis and accept the alternative hypothesis. There is a statistically significant relationship between the amount of time spent studying and exam scores.

The second approach to interpreting the Pearson correlation coefficient is to square the sample correlation value. The r^2 value is $(.80)^2 = .64$. This implies that 64% of the variability in exam scores can be explained by knowledge of how much time a student spent studying. This also implies that 36% of the variability in the exam scores is due to other variable relationships or unexplained variance. The average number of hours spent studying was 3.4 h with a standard deviation of 1.71. The average exam score was 86 with a standard deviation of 7.38. The interpretation is linked to the variance of the exam scores, hence $(7.38)^2 = 54.46$. We would state that 64% of 54.46 is explained variability and 36% of 54.46 is unexplained variability given knowledge of how much time a student spent studying. We can also depict this relationship using a *Venn* or *Ballentine* diagram.



The third approach to interpreting the correlation coefficient obtained from sample data is to graph the data points of the two variables. The scatter plot is used for this purpose. We draw a Y-axis for the exam scores and an X-axis for the amount of time spent studying. We label and scale these two axes to provide a grid such that the pairs of data points can be graphed. A visual look at the "trend" of the pairs of data points helps in interpreting whether the positive direction of the correlation coefficient exists. Scatter plots can display an upward trend (positive relationship), downward trend (negative relationship), or a curvilinear trend (one-half positive and the other half negative). If a curvilinear relationship exists, one-half cancels the other half out, so the correlation coefficient would be zero and the interpretation of the correlation coefficient meaningless. This is why Karl Pearson made the assumption of linear data. A scatter plot of the data points visually reveals the positive upward trend expected from r = +.80.



The Pearson correlation coefficient in determining whether or not there is a linear relationship between two continuous variables provides both a measure of the strength of the relationship, as well as, the direction of the relationship. In our example, the number of hours a student spent studying was related positively to the exam score. The strength of the relationship was indicated by a value close to 1.0 and the direction of the relationship by the positive sign. We are also able to explain the variability in the exam scores by squaring the correlation coefficient. In other words, why didn't all the students get the same score, because some students studied more! This can be presented in a diagram and depicted as a percent of the variance of the exam scores that can be explained. A scatter plot is the best visual aid to understanding the trend in the pairs of scores in regards to both magnitude and direction.

Karl Pearson's correlation coefficient was one of the most important discoveries in the field of statistics because numerous other statistical techniques, such as multiple regression, path analysis, factor analysis, cluster analysis, discriminant analysis, canonical correlation, and structural equation modeling, are based on this coefficient and interpretative understanding. Over one hundred years later, the examination of variable relationships is the singular most important analysis conducted in education, psychology, business, medicine, and numerous other disciplines. The correlation approach assumes that both the X and the Y variables are random, and have a distribution known as the bivariate normal distribution. In the bivariate normal distribution, for any given X value the Y values have a normal distribution in which the mean and the standard deviation depend on the value of X and the strength of the relationship between X and Y. The strength of the relationship between X and Y is measured by a population parameter ρ (pronounced "rho"), which can range between -1 and 1 inclusive. If $\rho = 0$, there is either no relationship between X and Y, or a curvilinear relationship which the Pearson correlation coefficient doesn't detect. If $\rho = 1$, there is a perfect *positive* linear relationship between the two variables, and if $\rho = -1$, there is a perfect *negative* linear relationship between the variables. A value of ρ close to zero indicates a weak relationship (assuming linear data), and a value close to either $+1$ or -1 indicates a strong relationship. Consequently, $r = -.90$ is a stronger relationship than $r = +.50$ for the same sample size. Because the Pearson correlation values form an ordinal scale, we do not directly compare the distances between two correlation values. For example, if a correlation of $r = +.50$ was obtained in one sample and a correlation of $r = +.60$ obtained in a second sample, you *would not* indicate that the second correlation was .10 higher than the first because the correlation is an ordinal scale!

In this chapter, the Pearson correlation coefficient will be calculated using the bivariate normal distribution. Other correlation coefficients have been developed since 1898 to establish the relationship between nominal and ordinal data, but are not presented in this chapter, i.e., phi-coefficient (nominal data), Spearman-coefficient (ordinal data). R has functions for these correlation coefficients making individual calculations straightforward.

In a bivariate normal distribution, both X and Y are random variables. The Pearson correlation coefficient indicates the linear relationship between two continuous variables. The Pearson correlation coefficient indicates both the magnitude

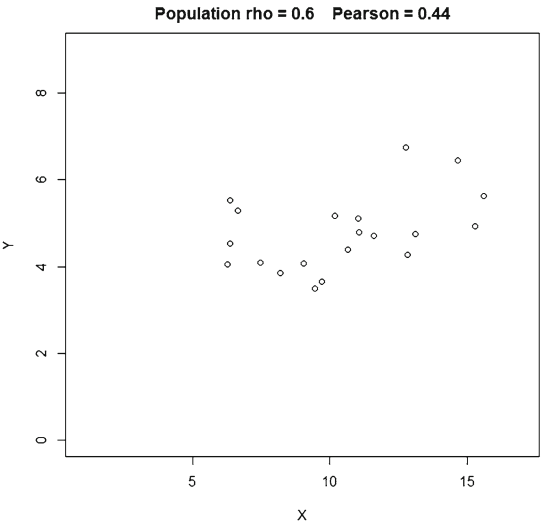
and direction of the relationship between two linear continuous variables. A Pearson correlation coefficient of r = 0 indicates no linear relationship between two variables. The correlation coefficient can be interpreted in three ways: test of significance, variance accounted for, and a diagram of trend in the paired data points. The sample correlation coefficient, r, is an estimate of the population correlation coefficient rho. If rho = 0, then the sample data points will have a random scatter plot, and the least squares line will be horizontal. As rho approaches +1 or −1, the sample data points are closer to a straight line, either upward for positive correlations, or downward for negative correlations. If rho = +1 or −1, then the sample points will lie directly in a line.

CORRELATION R Program

The CORRELATION program specifies the value of rho (correlation) in the population. The program then selects variable X at random from a normal distribution with mean of 10 and standard deviation of 3. Next, a random variable Y is selected from the normal distribution that is determined by the random X variable, given rho selected and the mean and standard deviation of Y, which are 5 and 1, respectively. A scatter plot is drawn for the pairs of X and Y scores. By varying rho in the CORRELATION program, you can observe the different scatter plots of data points that arise when rho has different values in the population. The **cor** function is used to compute the sample Pearson correlation, which is placed into a label to be used later along with a label for the *rho* value. The limits of the X and Y axes are set before plotting the pairs of data points so that no plotted points fall outside the axes. Finally, the sample data points are plotted in a scatter plot with the labels for the sample correlation and *rho*. The last program line prints out the sample correlation.

CORRELATION Program Output

```
Population rho=0.6
Sample Size=20
Pearson r=0.44
```



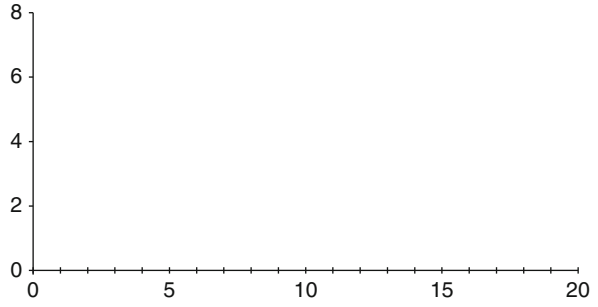
Correlation Exercises

1. Run CORRELATION program for the following values of rho and sample size. Record the Pearson correlation coefficient.

rho = + .5

Sample Size = 20

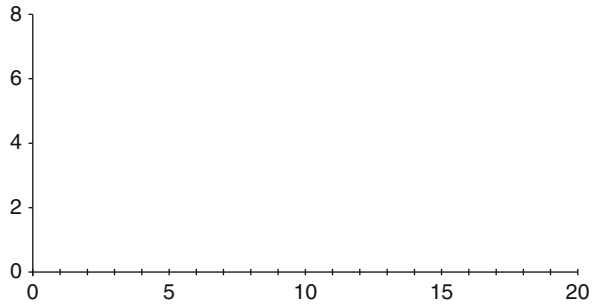
Pearson r = _____



rho = – .5

Sample Size = 20

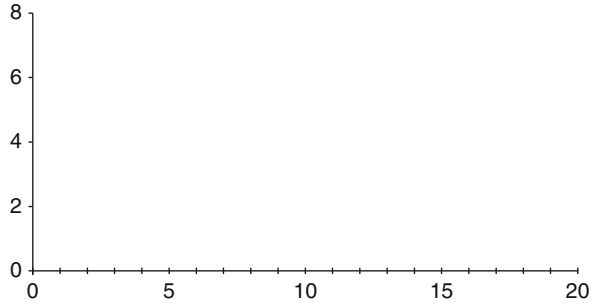
Pearson r = _____



rho = 0.0

Sample Size = 20

Pearson r = _____



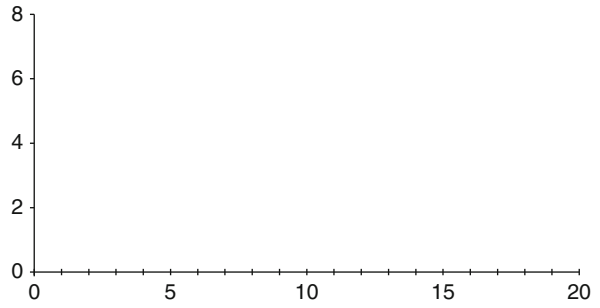
a. For rho = +.5, is there an upward trend to the points? YES _____ NO _____
b. For rho = –.5, is there a downward trend to the points? YES _____ NO _____
c. For rho = 0.0, does there appear to be no upward or downward trend to the points? YES _____ NO _____

2. Run CORRELATION program for the following values of rho. Record the correlation coefficient, r. Plot the data values and draw a line over the points on the graph.

rho = +1.0

Sample Size = 20

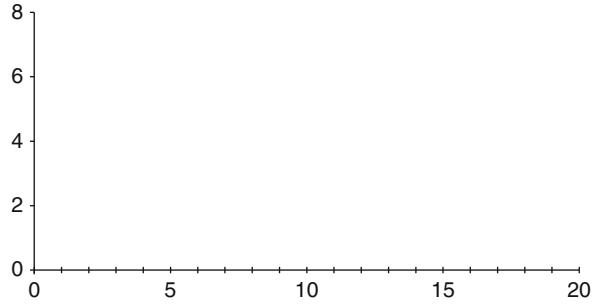
Pearson r = _____



rho = –1.0

Sample Size = 20

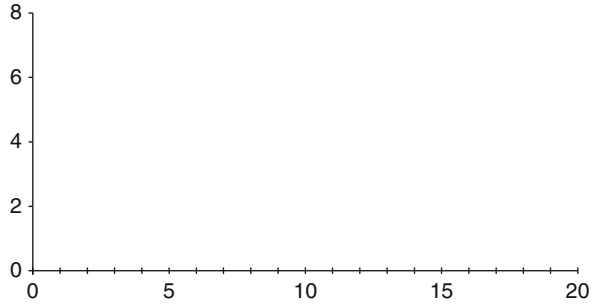
Pearson r = _____



rho = 0

Sample Size = 20

Pearson r = _____



a. Describe the scatter plot for the data points when rho = 1.

_____

_____

b. Describe the scatter plot for the data points when rho = −1.

_____

_____

c. Describe the scatter plot for the data points when rho = 0.

_____

_____

3. Input sample size = 100, then run Correlation program for rho = .6 to compute Pearson r.
   rho = .4, sample size = 20, Pearson r = .42
   rho = .4, sample size = 100, Pearson r = _____
   Does sample size effect Pearson r? Yes_____ No _____

True or False Questions

Pearson Correlation

T F a. The Pearson correlation coefficient indicates the relationship between two linear continuous variables.

T F b. A correlation of r = .60 is greater than a correlation of r = −.80.

T F c. A correlation of r = 0 implies no relationship between two variables.

T F d. If rho is −.5, the scatter plot of data points will indicate an upward trend.

T F e. If r = .80, then 64% of the variability in one variable is explained by knowledge of the other variable.

T F f. If rho = 0, then r = 0.

T F g. Rho is a parameter, and r is a statistic.

T F h. Data points will fall on a straight line when r = 1.0.