Chapter 11

Correlation Methods

Fundamentals of Biostatistics

Prof. Dr. Moustafa Omar Ahmed Abu-Shawiesh

Professor of Statistics

11.1 Introduction

To quantify the association between two continuous variables, we can use the correlation coefficient. In this chapter (chapter 11), we consider hypothesis-testing methods for correlation coefficients to describe association among two continuous variables in the same sample.

EXAMPLE 11.2

Hypertension Much discussion has taken place in the literature concerning the familial aggregation of blood pressure. In general, children whose parents have high blood pressure tend to have higher blood pressure than their peers. One way of expressing this relationship is by computing a correlation coefficient relating the blood pressure of parents and children over a large collection of families.

11.7 The Correlation Coefficient

In this section, we will introduce the concept of a correlation coefficient which will be used when we are interested in investigating whether or not there is a relationship (*association*) between two variables, a dependent variable (y) and an independent variable (x).

EXAMPLE 11.1

Obstetrics Obstetricians sometimes order tests to measure estriol levels from 24-hour urine specimens taken from pregnant women who are near term because level of estriol has been found to be related to infant birthweight. Therefore, the relationship between estriol level and birthweight relates the two variables. Birthweight is the dependent variable and estriol is the independent variable because estriol levels are being used to try to predict birthweight.

EXAMPLE 11.26

Cardiovascular Disease Serum cholesterol is an important risk factor in the etiology of cardiovascular disease. Much research has been devoted to understanding the environmental factors that cause elevated cholesterol levels. For this purpose, cholesterol levels were measured on 100 genetically unrelated spouse pairs. We are interested in a quantitative measure of the relationship between their levels. We will use the correlation coefficient for this purpose.

First, we discuss the related concept of covariance. The covariance is a measure used to quantify the relationship between two random variables.

DEFINITION 11.15

The **covariance** between two random variables X and Y is denoted by $Cov(X, Y)$ and is defined by

$$Cov(X,Y) = E\left[(X - \mu_x)(Y - \mu_y)\right]$$

which can also be written as $E(XY) - \mu_x\mu_y$, where μ_x is the average value of X, μ_y is the average value of Y, and $E(XY)$ = average value of the product of X and Y.

Notations

- It can be shown that if the random variables X and Y are independent, then the covariance between them is 0.
- If large values of X and Y tend to occur among the same subjects (as well as small values of X and Y), then the covariance is positive.
- If large values of X and small values of Y (or conversely, small values of X and large values of Y) tend to occur among the same subjects, then the covariance is negative.

One issue is that, the covariance between two random variables X and Y is in the units of X multiplied by the units of Y. Thus, it is difficult to interpret the strength of association between two variables from the magnitude of the covariance. To obtain a measure of relatedness independent of the units of X and Y, we consider the correlation coefficient.

DEFINITION 11.16

The **correlation coefficient** between two random variables X and Y is denoted by $Corr(X, Y)$ or ρ and is defined by

$$\rho = Corr(X,Y) = Cov(X,Y)/(\sigma_x\sigma_y)$$

where σ_x and σ_y are the standard deviations of X and Y, respectively.

Notations

- Unlike the covariance, the correlation coefficient is:
  (1) A dimensionless quantity that is independent of the units of X and Y, and
  (2) Ranges between −1 and 1.

- For random variables that are approximately linearly related, a correlation coefficient of 0 implies independence.

- A correlation coefficient close to 1 implies nearly perfect positive dependence with large values of X corresponding to large values of Y and small values of X corresponding to small values of Y.

Example

(1) A strong positive correlation is between forced expiratory volume (FEV), a measure of pulmonary function, and height.

(2) A somewhat weaker positive correlation exists between serum cholesterol and dietary intake of cholesterol.

- A correlation coefficient close to −1 implies ≈ perfect negative dependence, with large values of X corresponding to small values of Y and vice versa.

Example

The relationship between resting pulse rate and age in children under the age of 10. A somewhat weaker negative correlation exists between FEV and number of cigarettes smoked per day in children.

- For variables that are not linearly related, it is difficult to infer independence or dependence from a correlation coefficient.

- It would be a mistake to assume that the random variables X and Y are independent if the correlation coefficient between them is 0, that is, $Corr(X, Y) = 0$.

11.7.1 Scatter Plot

Many research projects are correlational studies because they investigate the relationships that may exist between variables. Prior to investigating the relationship between two quantitative variables, it is always helpful to create a graphical representation that includes both of these variables. Such a graphical representation is called a scatterplot.

Notation

- It is the most useful graph for displaying the relationship between two quantitative variables.
- The purpose of a scatterplot is to provide a general illustration of the relationship between the two variables.

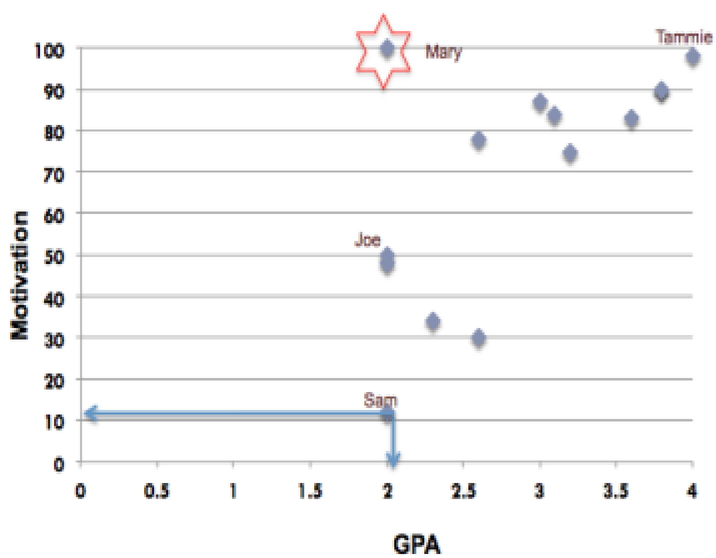
Definition

A **scatterplot** shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.

Scatter Plot Example

The scatter plot given below show the relationship between students' achievement motivation and GPA:

Student	Student GPA	Motivation
Joe	2.0	50
Lisa	2.0	48
Mary	2.0	100
Sam	2.0	12
Deana	2.3	34
Sarah	2.6	30
Jennifer	2.6	78
Gregory	3.0	87
Thomas	3.1	84
Cindy	3.2	75
Martha	3.6	83
Steve	3.8	90
Jamell	3.8	90
Tammie	4.0	98



- The image given above is an example of a scatter plot and displays the data from the table. GPA scores are displayed on the horizontal axis (x) and motivation scores are displayed on the vertical axis (y).

- Each dot on the scatter plot represents one individual from the data set. The location of each point on the graph depends on both the GPA and motivation scores.

➢ Scatter plots are not meant to be used in great detail because there are usually hundreds of individuals in a data set.

Important Notation

In Definition 11.16, we defined the population correlation coefficient ρ. In general, ρ is unknown and we have to estimate ρ by the sample correlation coefficient r.

DEFINITION 11.17

The sample correlation coefficient (*Pearson's correlation coefficient*), usually we refer to it by r, of the data pairs $(x_i\,, y_i)$, $i \;=\; 1, \dots, n$ is defined by

$$r = \frac{L_{xy}}{\sqrt{L_{xx}\,L_{yy}}}$$

where the following formulas are needed to calculate the value of the sample correlation coefficient (r):

(1) L_{xy} is the **corrected sum of cross products** defined by:

$$L_{xy} \quad = \quad \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \quad = \quad \sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)\Big/n$$

(2) L_{xx} is the **corrected sum of squares for x** defined by:

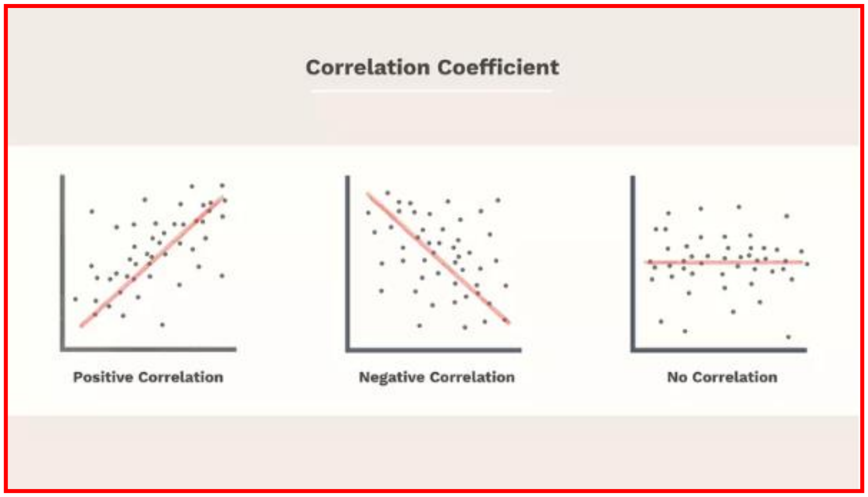
$$L_{xx} \quad = \quad \sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\Big/n$$

(3) L_{yy} is the **corrected sum of squares for y** defined by:

$$L_{yy} \quad = \quad \sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\Big/n$$

Notation

The correlation is not affected by changes in location or scale in either variable and must lie between −1 and +1, that is, $-1 \le r \le +1$.



The sample correlation coefficient can be interpreted in a similar manner to the population correlation coefficient (ρ) as in Equation 11.15 given bellow:

EQUATION 11.15

Interpretation of the Sample Correlation Coefficient

(1) If the correlation is greater than 0, such as for birthweight and estriol, then the variables are said to be **positively correlated.** Two variables (x, y) are positively correlated if as x increases, y tends to increase, whereas as x decreases, y tends to decrease.

(2) If the correlation is less than 0, such as for pulse rate and age, then the variables are said to be **negatively correlated**. Two variables (x, y) are negatively correlated if as x increases, y tends to decrease, whereas as x decreases, y tends to increase.

(3) If the correlation is exactly 0, such as for birthweight and birthday, then the variables are said to be **uncorrelated.** Two variables (x, y) are uncorrelated if there is no linear relationship between x and y.

Thus the sample correlation coefficient provides a *quantitative* estimate of the dependence between two variables: the closer $|r|$ is to 1, the more closely related the variables are; if $|r| = 1$, then one variable can be predicted exactly from the other.

As was the case for the population correlation coefficient (ρ), interpreting the sample correlation coefficient (r) in terms of degree of dependence is only correct if the variables x and y are normally distributed and in certain other special cases. If the variables are not normally distributed, then the interpretation may not be correct.

EXAMPLE

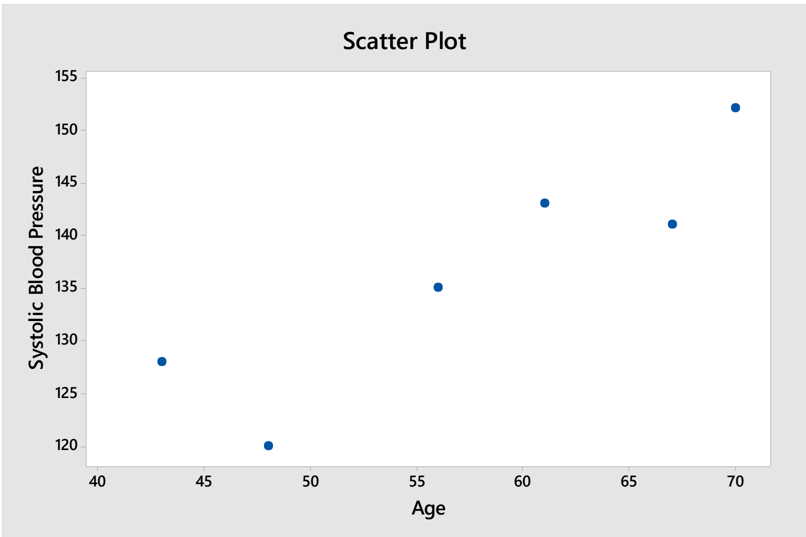
The data shown in the table below obtained in a study of age (x), $in\ years$, and systolic blood pressure (y), *in mm Hg*, (*indicates how much pressure your blood is exerting against your artery walls when the heart contracts*) for a random sample of six patients selected from the emergency room of Jordan University Hospital (JUH) in a given day:

Age (x) $(years)$	Systolic Blood Pressure (y) *(mm Hg)*
43	128
48	120
56	135
61	143
67	141
70	152

Answer the following:

(a) Construct a scatter plot for the data? Conclusion?

Solution



Conclusion

From the scatter plot we can conclude that there is a strong positive linear relationship between the age and systolic blood pressure.

(b) Calculate the value of the correlation coefficient for the data? Conclusion?

Solution

Step (1): Make a worktable as shown below:

	Age (x) ($years$)	Systolic Blood Pressure (y) (*mm Hg*)	x^2	y^2	xy
	43	128	**1849**	**16384**	**5504**
	48	120	**2304**	**14400**	**5760**
	56	135	**3136**	**18225**	**7560**
	61	143	**3721**	**30449**	**8723**
	67	141	**4489**	**19881**	**9447**
	70	152	**4900**	**23104**	**10640**
Total (Sum)	**345**	**819**	**20399**	**112443**	**47634**

That is:

$$n = 6 \quad ; \quad \sum_{i=1}^{6} x_i y_i = 47634$$

$$\sum_{i=1}^{6} x_i = 345 \quad ; \quad \sum_{i=1}^{6} y_i = 819$$

$$\sum_{i=1}^{6} x_i^2 = 20399 \quad ; \quad \sum_{i=1}^{6} y_i^2 = 112443$$

$$\bar{x} = \sum_{i=1}^{6} x_i/6 = 345/6 = 57.5 \quad ; \quad \bar{y} = \sum_{i=1}^{6} y_i/6 = 819/6 = 136.5$$

Step (2): The value of the correlation coefficient (r) can be calculated by using the formula as follows:

$$r = \frac{L_{xy}}{\sqrt{L_{xx}\, L_{yy}}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n\right)\left(\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n\right)}}$$

$$= \frac{[47634 - ((345)(819)/6)]}{\sqrt{[20399 - ((345)^2/6)][112443 - ((819)^2/6)]}}$$

$$= \frac{541.5}{\sqrt{(561.5)(649.5)}}$$

$$= 0.897$$

Conclusion

From the sign and value of the Pearson's correlation coefficient (r) we can conclude that there is a strong positive linear relationship between the age (x) and systolic blood pressure (y).

11.7.2 The Relationship Between the Sample Correlation Coefficient (r) and the Population Correlation Coefficient (ρ)

We can relate the sample correlation coefficient (r) and the population correlation coefficient (ρ) more clearly by dividing the numerator and denominator of sample correlation coefficient (r) by $(n-1)$ in Definition 11.17, where by:

Equation 11.16

$$r = \frac{L_{xy} / (n-1)}{\sqrt{\left(\frac{L_{xx}}{n-1}\right)\left(\frac{L_{yy}}{n-1}\right)}}$$

We note that:

$$\boxed{s_x^2 = L_{xx} / (n-1)} \quad \text{and} \quad \boxed{s_y^2 = L_{yy} / (n-1)}$$

Furthermore, if we define the *sample covariance* by:

$$\boxed{s_{xy} = L_{xy} / (n-1)}$$

Then we can re-express Equation 11.16 in the following form:

Equation 11.17

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\text{sample covariance between } x \text{ and } y}{(\text{sample standard deviation of } x)(\text{sample standard deviation of } y)}$$

This is completely analogous to the definition of the population correlation coefficient (ρ) given in Definition 11.16 with the population quantities, Cov(X, Y), σ_x , and σ_y replaced by their sample estimates s_{xy}, s_x , and s_y.

Notations

- The sample correlation coefficient (r) will be unchanged by a change in the units of x or y (or even by which variable is designated as x and which is designated as y).

- Based on Equation 11.17, if every unit in the reference population could be sampled, then the sample correlation coefficient (r) would be the same as the population correlation coefficient, denoted by ρ, which was introduced in Definition 11.16 (on p. 486).

- The correlation coefficient is used when we simply want to describe the linear relationship (*association*) between two variables but are not interested in predicting one variable from another.

11.8 Statistical Inference for Correlation Coefficients

In the previous section, we defined the sample correlation coefficient (r). In this section, we discuss various hypothesis tests concerning correlation coefficients. That is, we will use r, which is computed from finite samples, to test various hypotheses concerning ρ.

11.8.1 One-Sample t Test for a Correlation Coefficient (ρ)

In this section, we want to test the hypothesis $H_0: \rho = 0$ vs $H_1: \rho \neq 0$, d, then the best procedure for testing the hypothesis is given as follows:

Equation 11.20

One-Sample *t* Test for a Correlation Coefficient

To test the hypothesis H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$, use the following procedure:

(1) Compute the sample correlation coefficient r.

(2) Compute the test statistic

$$t = r(n-2)^{1/2}/(1-r^2)^{1/2}$$

which under H_0 follows a t distribution with $n - 2$ df.

(3) For a two-sided level α test,

if $t > t_{n-2,1-\alpha/2}$ or $t < -t_{n-2,1-\alpha/2}$

then reject H_0.

If $-t_{n-2,1-\alpha/2} \le t \le t_{n-2,1-\alpha/2}$

then accept H_0.

(4) The p-value is given by

$$p = 2 \times (\text{area to the left of } t \text{ under a } t_{n-2} \text{ distribution}) \quad \text{if } t < 0$$

$$p = 2 \times (\text{area to the right of } t \text{ under a } t_{n-2} \text{ distribution}) \quad \text{if } t \ge 0$$

(5) We assume an underlying normal distribution for each of the random variables used to compute r.

The acceptance and rejection regions for this test are shown in Figure 11.14. Computation of the p-value is illustrated in Figure 11.15.

Notation

The test statistic (t) given in step (2) of the test procedure (Equation 11.20) can be re-expressed as follows:

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

FIGURE 11.14 Acceptance and rejection regions for the one-sample *t* test for a correlation coefficient

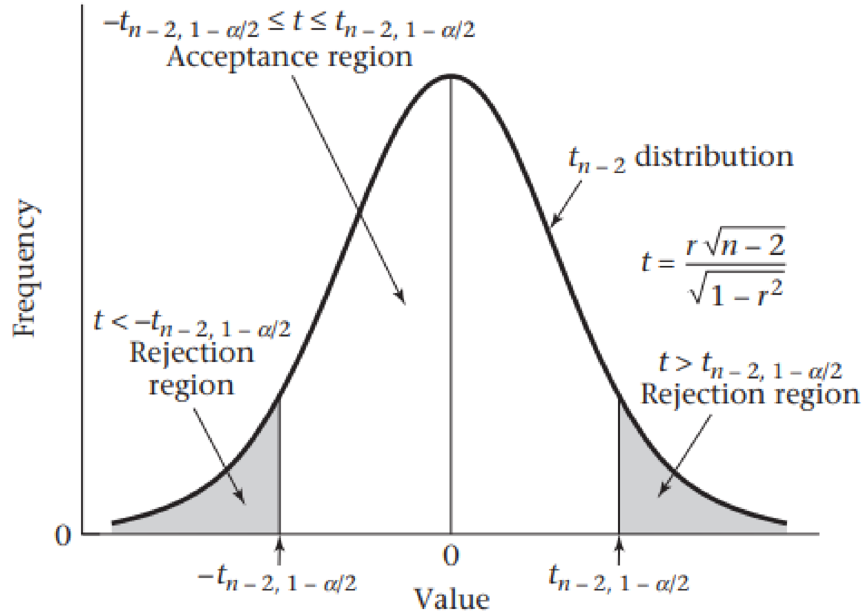
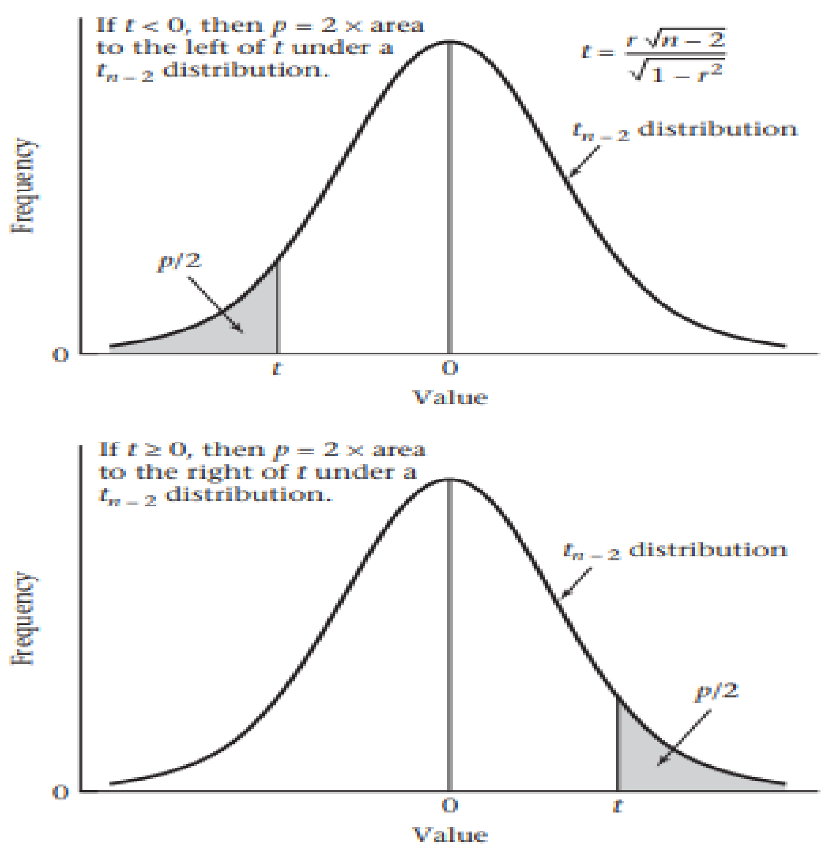


FIGURE 11.15 Computation of the *p*-value for the one-sample *t* test for a correlation coefficient



EXAMPLE 11.31

Cardiovascular Disease Suppose serum-cholesterol levels in spouse pairs are measured to determine whether there is a correlation between cholesterol levels in spouses. Specifically, we wish to test the hypothesis:

$$H_0: \rho = 0 \ vs\ H_1: \rho \neq 0$$

Suppose that $r = 0.897$ based on $n = 6$ spouse pairs. Is this evidence enough to warrant rejecting H_0? Perform a test of significance for the data in this Example? Use $\alpha = 0.05$?

Solution

Step (1): We have $r = 0.897$ based on $n = 6$. Thus, in this case, the value of the test statistic (t) can be calculated as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{(0.897)\sqrt{6-2}}{\sqrt{1-(0.897)^2}} = \frac{1.794}{0.442} = 4.056$$

Step (2): The critical value will be obtained from Table 5 in the Appendix as follows:

$$t_{(n-2,\ 1-\alpha/2)} = t_{(6-2,\ 1-0.05/2)} = t_{(4,\ 0.975)} = 2.776$$

Step (3): The decision will be to reject H_0 because we get:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 4.056 > t_{(n-2,\ 1-\alpha/2)} = t_{(4,\ 0.975)} = 2.776$$

Step (4): The p-value because ($t = 4.056 > 0$) can be calculated as follows:

$$\begin{aligned} p\text{-value} &= 2 \times P\left(t_{(4,\ 0.975)} > 4.056\right) \\ &= 2 \times \left[1 - P\left(t_{(4,\ 0.975)} \leq 4.056\right)\right] \\ &= 2 \times [1 - 0.99] \\ &= 2 \times [0.01] \\ &= 0.02 < \alpha = 0.05 \end{aligned}$$

Conclusion

We conclude there is a significant aggregation of cholesterol levels between spouses. This result is possibly due to common environmental factors such as diet. But it could also be due to the tendency for people of similar body build to marry each other, and their cholesterol levels may have been correlated at the time of marriage.

11.8.2 One-Sample Z Test for a Correlation Coefficient (ρ)

In the previous section, a test of the hypothesis:

$$H_0: \rho = 0 \ vs \ H_1: \rho \neq 0$$

was considered. Sometimes the correlation between two random variables is expected to be some quantity ρ_0 other than 0 and we want to test the hypothesis:

$$H_0: \rho = \rho_0 \ vs \ H_1: \rho \neq \rho_0$$

The problem with using the t test formation in Equation 11.20 is that the sample correlation coefficient (r) has a skewed distribution for nonzero ρ that cannot be easily approximated by a normal distribution. Fisher considered this problem and proposed the following transformation to better approximate a normal distribution:

Equation 11.21

Fisher's *z* Transformation of the Sample Correlation Coefficient *r*

The z transformation of r given by

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

is approximately normally distributed under H_0 with mean

$$z_0 = \frac{1}{2}\ln[(1+\rho_0)/(1-\rho_0)]$$

and variance $1/(n-3)$. The z transformation is very close to r for small values of r but tends to deviate substantially from r for larger values of r. A table of the z transformation is given in Table 12 in the Appendix.

EXAMPLE 11.35

Suppose the body weights of 100 fathers (x) and first-born sons (y) are measured and a sample correlation coefficient r of 0.38 is found. We might ask whether or not this sample correlation is compatible with an underlying correlation of 0.5 that might be expected on genetic grounds. Compute the z transformation of $r = 0.38$?

Solution

The z transformation can be computed from Equation 11.21 as follows:

$$z = \frac{1}{2}\ln\left(\frac{1+0.38}{1-0.38}\right) = \frac{1}{2}\ln\left(\frac{1.38}{0.62}\right) = \frac{1}{2}\ln(2.226) = \frac{1}{2}(0.800) = 0.400$$

Alternatively, we could refer to Table 12 (Page 887) in the Appendix with $r = 0.38$ to obtain the z transformation to be $\boxed{z = 0.400}$.

TABLE 12 Fisher's *z* transformation

r	*z*
.36	.377
.37	.388
.38	.400

The Fisher's z transformation can be used to conduct the hypothesis test procedure for a two-sided level α test as follows:

Equation 11.22

One-Sample *z* Test for a Correlation Coefficient

To test the hypothesis H_0: $\rho = \rho_0$ vs. H_1: $\rho \neq \rho_0$, use the following procedure:

(1) Compute the sample correlation coefficient r and the z transformation of r.

(2) Compute the test statistic

$$\lambda = (z - z_0)\sqrt{n-3}$$

(3) If $\lambda > z_{1-\alpha/2}$ or $\lambda < -z_{1-\alpha/2}$ reject H_0.

If $-z_{1-\alpha/2} \leq \lambda \leq z_{1-\alpha/2}$ accept H_0.

(4) The exact p-value is given by

$$p = 2 \times \Phi(\lambda) \quad \text{if } \lambda \leq 0$$
$$p = 2 \times [1 - \Phi(\lambda)] \quad \text{if } \lambda > 0$$

(5) Assume an underlying normal distribution for each of the random variables used to compute r and z.

The acceptance and rejection regions for this test are shown in Figure 11.16. Computation of the p-value is illustrated in Figure 11.17.

FIGURE 11.16 Acceptance and rejection regions for the one-sample *z* test for a correlation coefficient

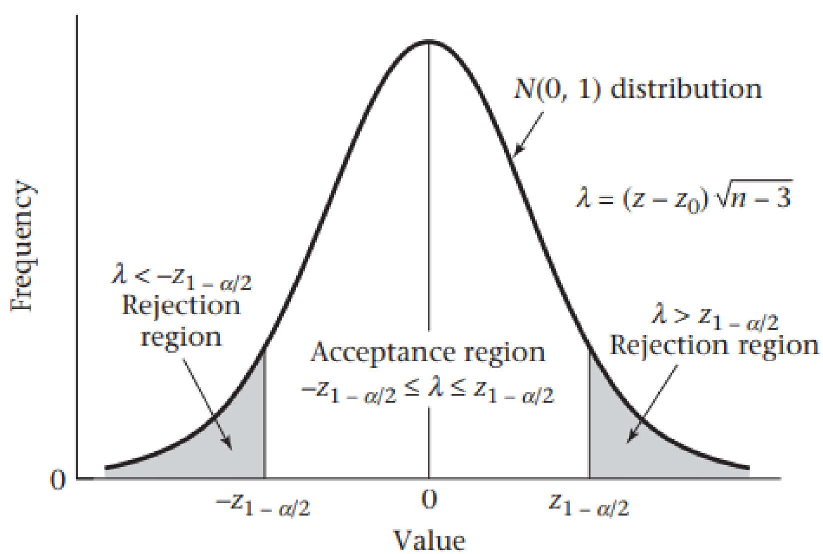
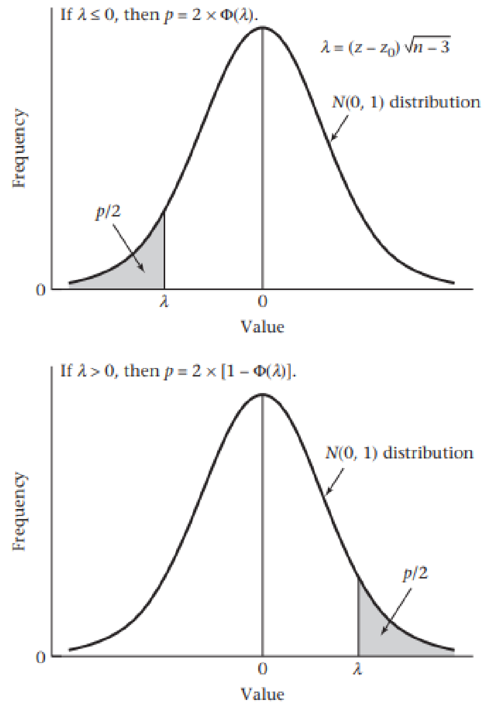


FIGURE 11.17 Computation of the p-value for the one-sample z test for a correlation coefficient



EXAMPLE 11.36

Perform a test of significance for the data in Example 11.35? Use $\alpha = 0.05$?

Solution

In this case $r = 0.38$, $n = 100$, $\rho_0 = 0.50$, then from Table 12 in the Appendix, or by using formula given in Equation 11.21 we get:

$$z_0 = \frac{1}{2}\ln\left(\frac{1+.5}{1-.5}\right) = .549 \qquad z = \frac{1}{2}\ln\left(\frac{1+.38}{1-.38}\right) = .400$$

Hence,

$$\lambda = (0.400 - 0.549)\sqrt{97} = (-0.149)(9.849) = -1.47 \sim N(0,1)$$

Now, because $\lambda = -1.47 < 0$, then the p-value can be calculated as follows:

$$\begin{aligned} p\text{-value} &= 2 \times \phi(\lambda) \\ &= 2 \times \phi(-1.47) \\ &= 2 \times [1 - \phi(1.47)] \\ &= 2 \times [1 - 0.9292] \\ &= 2 \times 0.0708 \\ &= 0.1416 > \alpha = 0.05 \end{aligned}$$

Decision and Conclusion

We accept H_0 that the sample estimate of correlation coefficient 0.38 is compatible with an underlying correlation of 0.50.

Notation

To sum up, the z test in Equation 11.22 is used to test hypotheses about nonzero null correlations, whereas the t test in Equation 11.20 is used to test hypotheses about null correlations of zero. The z test can also be used to test correlations of zero under the null hypothesis, but the t test is slightly more powerful in this case and is preferred. However, if $\rho_0 \neq 0$, then the one-sample z test is very sensitive to non-normality of either x or y.

11.8.3 Interval Estimation for Correlation Coefficients

In the previous sections, we learned how to estimate a correlation coefficient (ρ) and how to perform appropriate hypothesis tests concerning correlation coefficient (ρ). It is also of interest to obtain confidence limits (intervals) for the correlation coefficient (ρ). An easy method for obtaining confidence limits for correlation coefficient (ρ) can be derived based on the approximate normality of Fisher's z transformation of sample correlation coefficient (r). This method is given as follows:

Equation 11.23

Interval Estimation of a Correlation Coefficient (ρ)

Suppose we have a sample correlation coefficient r based on a sample of n pairs of observations. To obtain a two-sided $100\% \times (1-\alpha)$ confidence interval for the population correlation coefficient (ρ):

(1) Compute Fisher's z transformation of $r = z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$.

(2) Let z_ρ = Fisher's z transformation of $\rho = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$.

A two-sided $100\% \times (1-\alpha)$ confidence interval for $z_\rho = (z_1, z_2)$ where

$$z_1 = z - z_{1-\alpha/2}\big/\sqrt{n-3}$$
$$z_2 = z + z_{1-\alpha/2}\big/\sqrt{n-3}$$

and $z_{1-\alpha/2} = 100\% \times (1-\alpha/2)$ percentile of an $N(0, 1)$ distribution

(3) A two-sided $100\% \times (1-\alpha)$ confidence interval for ρ is then given by (ρ_1, ρ_2) where

$$\rho_1 = \frac{e^{2z_1}-1}{e^{2z_1}+1}$$
$$\rho_2 = \frac{e^{2z_2}-1}{e^{2z_2}+1}$$

Note that: The interval (z_1, z_2) in Equation 11.23 can be derived in a similar manner to the confidence interval for the mean of a normal distribution with known variance which is given by:

Equation 11.24

$$(z_1, z_2) = z \pm z_{1-\alpha/2} / \sqrt{n-3}$$

We then solve Equation 11.23 for r in terms of z, whereby:

Equation 11.25

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

We now substitute the confidence limits for z_ρ —that is, (z_1, z_2) in Equation 11.24— into Equation 11.25 to obtain the corresponding confidence limits for the correlation coefficient (ρ) given by (ρ_1, ρ_2) in Equation 11.23. The transformation from z to r in Equation 11.25 is sometimes referred to as the *inverse Fisher's z transformation*.

EXAMPLE 11.37

Suppose that a sample correlation coefficient of $r = 0.38$ was obtained between the body weights of fathers (x) and first-born sons (y) of $n = 100$ pairs. Find the 95% confidence interval for the underlying correlation coefficient (ρ)?

Solution

Step (1): From Example 11.36, the z transformation of $r = 0.38$, is calculated as follows:

$$z = \frac{1}{2} \ln \left(\frac{1+.38}{1-.38} \right) = .400$$

Step (2): From step (2) of Equation 11.23, a two-sided $(1 - \alpha) \times 100\%$ confidence interval for (z_ρ) is (z_1, z_2) and given by:

A two-sided $100\% \times (1 - \alpha)$ confidence interval for $z_\rho = (z_1, z_2)$ where

$$z_1 = z - z_{1-\alpha/2} / \sqrt{n-3}$$

$$z_2 = z + z_{1-\alpha/2} / \sqrt{n-3}$$

Thus, a 95% confidence interval for (z_ρ) given by (z_1, z_2) can be calculated as follows:

$$z_1 = 0.400 - 1.96 / \sqrt{97} = 0.400 - 0.199 = 0.201$$
$$z_2 = 0.400 + 1.96 / \sqrt{97} = 0.400 + 0.199 = 0.599$$

That is, a 95% confidence interval for z = (0.201, 0.599).

Step (3): From step (3) of Equation 11.23, a two-sided $(1 - \alpha) \times 100\%$ confidence interval for (ρ) is (ρ_1, ρ_2) and given by:

A two-sided 100% × (1 – α) confidence interval for ρ is then given by (ρ_1, ρ_2) where

$$\rho_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$$
$$\rho_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$$

Thus, a 95% confidence interval for (ρ) given by (ρ_1, ρ_2) can be calculated as follows:

$$CI = \left(\begin{aligned} \rho_1 &= \frac{e^{2(0.201)} - 1}{e^{2(0.201)} + 1} \\ &= \frac{e^{.402} - 1}{e^{.402} + 1} \\ &= \frac{1.4950 - 1}{1.4950 + 1} \\ &= \frac{0.4950}{2.4950} = .198 \end{aligned} \quad , \quad \begin{aligned} \rho_2 &= \frac{e^{2(.599)} - 1}{e^{2(.599)} + 1} \\ &= \frac{e^{1.198} - 1}{e^{1.198} + 1} \\ &= \frac{2.3139}{4.3139} = .536 \end{aligned} \right)$$

That is, a 95% confidence interval for ρ = (0.198, 0.536).

Notice that

The confidence interval for z_ρ, given by (z_1, z_2) = (0.201, 0.599), is symmetric about z = 0.400. However, when the confidence limits are transformed back to the original scale (the scale of ρ) the corresponding confidence limits for ρ are given by (ρ_1, ρ_2) = (0.198, 0.536), which are not symmetric around r = 0.38. The reason for this is that Fisher's z transformation is a nonlinear function of r, which only becomes approximately linear when r is small (i.e., $|r| \le 0.2$).

Problems: 11.32 – 11.35.