Chapter 11

Regression and Correlation Methods

Introduction

Correlation Coefficient

In this chapter, we will discuss

1. Scatter plot.
2. Pearson correlation coefficient (Section 11.7)
2. Testing Pearson Correlation coefficient (Section 11.8).
3. Confidence intervals for Pearson Correlation coefficient (Section 11.8).

Scatter Plot

- A scatter plot is a graphical representation that displays individual data points that correspond to values of two variables, one on the x-axis and the other on the y-axis.
- The primary purpose of a scatter plot is to visually examine the relationship between two variables. Plotting the data points allows us to identify patterns, trends, or correlations between the variables.

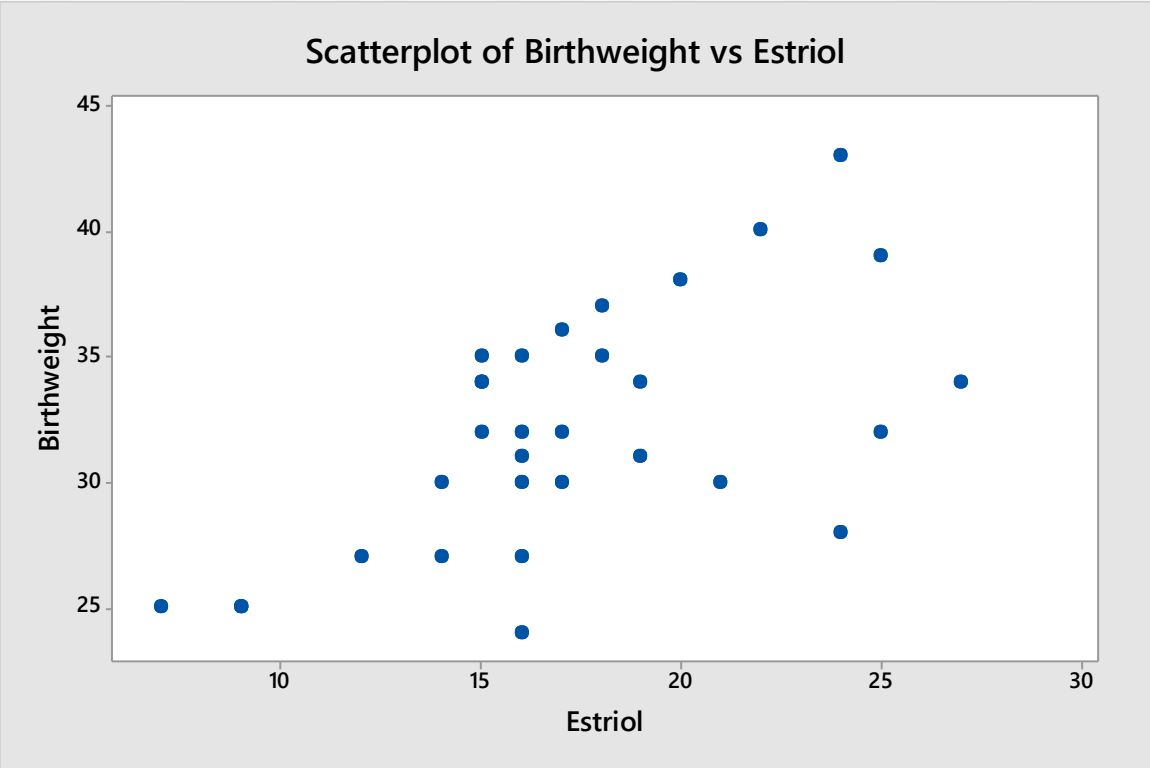
Example:

Table 11.1 **Sample data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term**

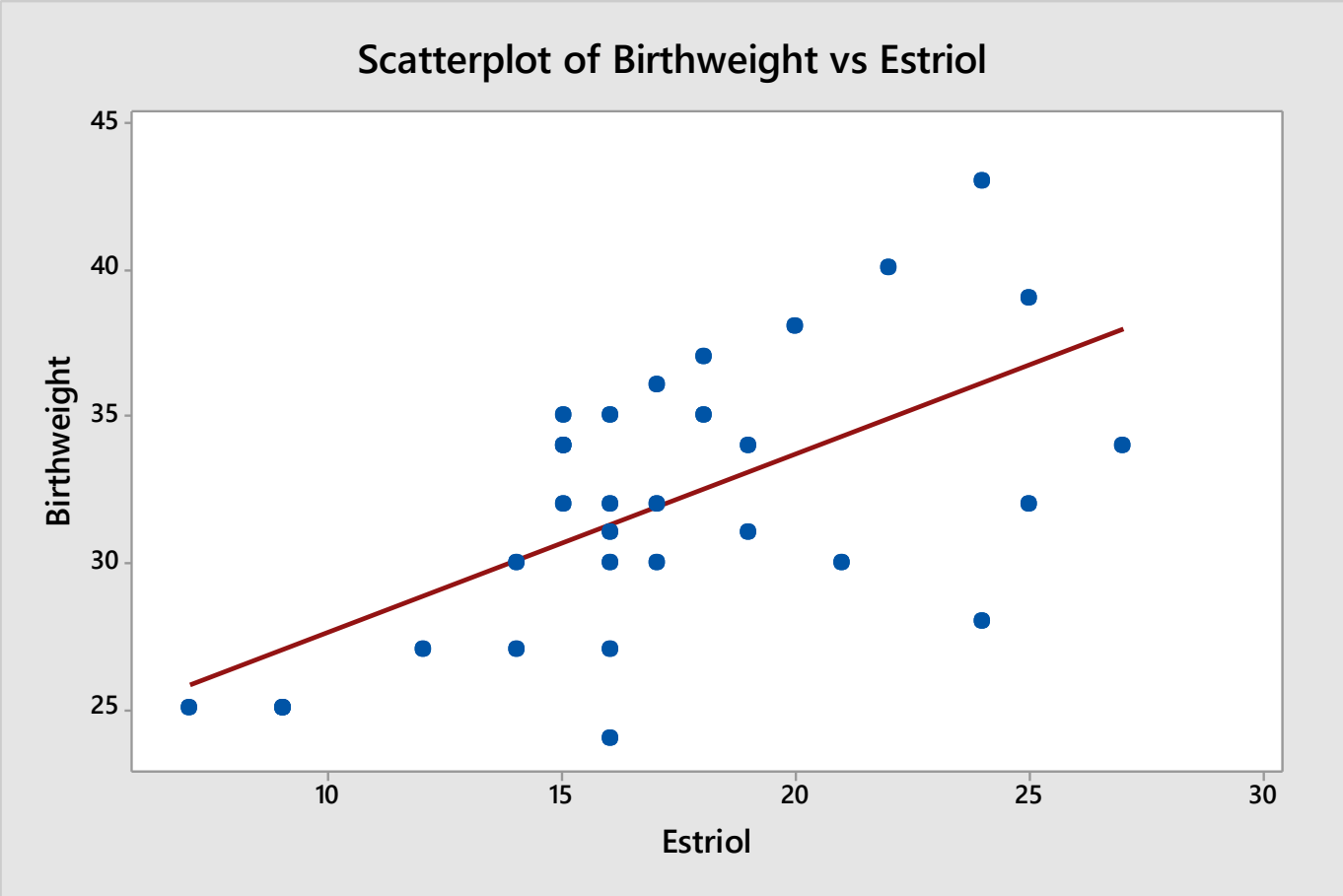
i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

Source: Reprinted with permission of the *American Journal of Obstetrics and Gynecology*, *85*(1), 1–9, 1963.

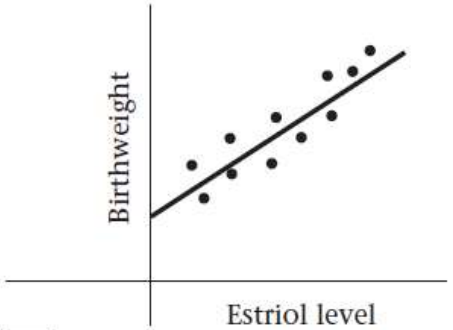
Example: Scatter plot of birthweight and Estriol Level



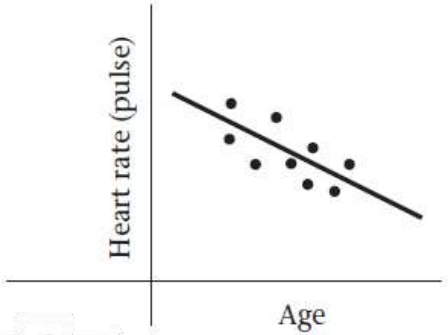
Example: Scatter plot of birthweight and Estriol Level along with a line (called regression line) representing the linear relationship between the two variables.



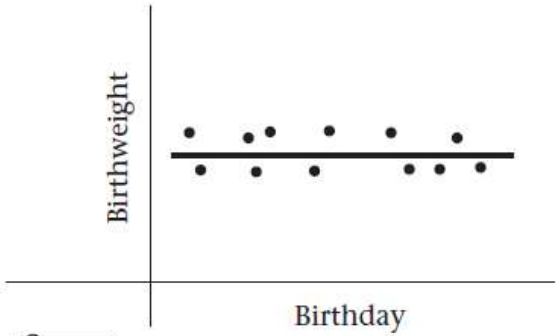
If ρ is positive, it indicates a positive linear relationship. As one variable increases, the other tends to increase as well. If ρ is negative, it indicates a negative linear relationship. As one variable increases, the other tends to decrease.



$\rho > 0$



$\rho < 0$

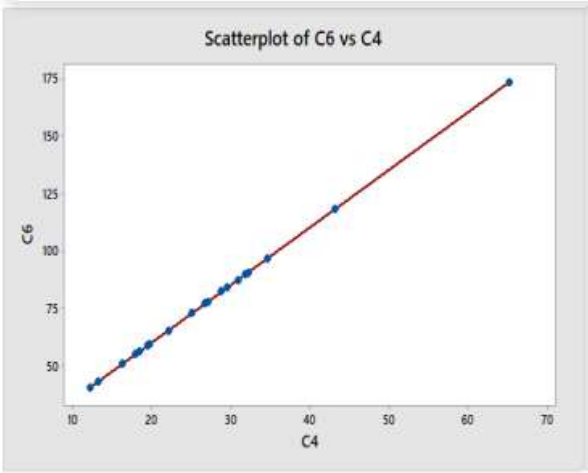


$\rho = 0$

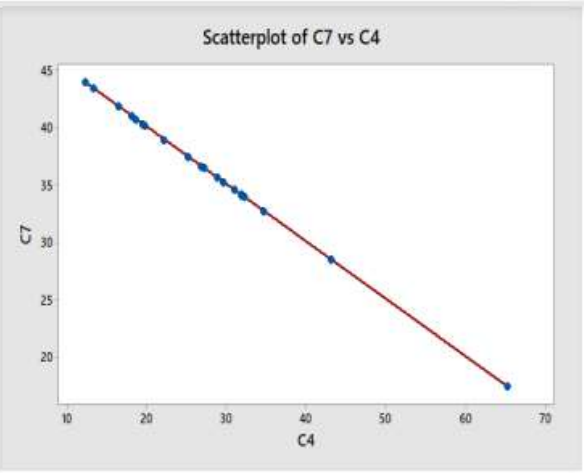
If ρ=0, it indicates no linear relationship between the two variables.

If ρ=1, it indicates a perfect positive linear relationship.

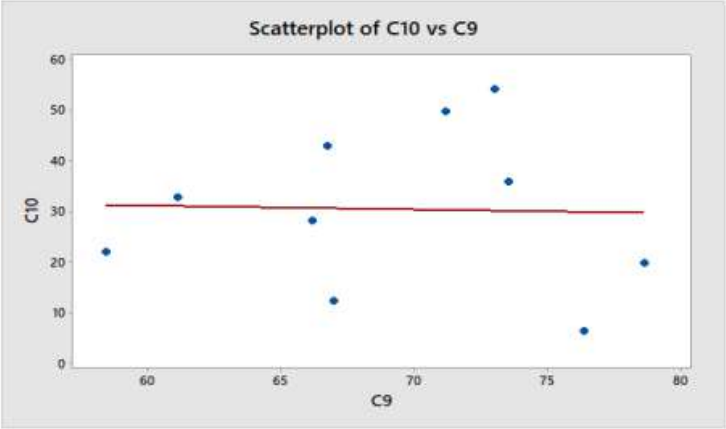
ρ=-1, indicates a perfect negative linear relationship.



$\rho = 1$



$\rho = -1$



$\rho = 0$

The following notation is needed

The **raw sum of squares for *x*** is defined by $\sum_{i=1}^{n} x_i^2$

The **corrected sum of squares for *x*** is denoted by L_{xx} and defined by $\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2/n$

It represents the sum of squares of the deviations of the x_i from the mean. Similarly, the **raw sum of squares for *y*** is defined by $\sum_{i=1}^{n} y_i^2$

The **corrected sum of squares for *y*** is denoted by L_{yy} and defined by $\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2/n$

The raw sum of cross products is defined by $\sum_{i=1}^{n} x_i y_i$

The corrected sum of cross products is defined by which is denoted by L_{xy}. $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

It can be shown that a short form for the corrected sum of cross products is given by

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

$$a = \bar{y} - b\bar{x} = \left(\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \right) \Big/ n$$

Correlation Coefficient

The **sample (Pearson) correlation coefficient (r)** is defined by $L_{xy}/\sqrt{L_{xx}L_{yy}}$. The correlation is not affected by changes in location or scale in either variable and must lie between -1 and +1. It is a useful tool for quantifying the relationship between variables.

Interpretation of the sample correlation coefficient

- If the correlation is greater than 0, 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.
- If the correlation is less than 0, 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.
- If the correlation is exactly 0, then the variables are said to be uncorrelated. Two variables (*x*,*y*) are uncorrelated if there is no linear relationship between *x* and *y*.

The correlation coefficient provides a *quantitative* measure 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.

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.

The sample correlation coefficient (r) is a point estimator of the population correlation coefficient (ρ)

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Where $\sigma_{xy} = E(XY) - E(X)E(Y)$ is the (population) covariance between the two variables X and Y.

The sample correlation coefficient is given by:

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

$$= \frac{s_{xy}}{s_x s_y}$$

$$= \frac{\text{sample covariance between } x \text{ and } y}{(\text{ sample standard deviation of } x)(\text{ sample standard deviation of } y)}$$

Where $s_x^2 = L_{xx}/(\mathrm{n}-1)$ and $s_y^2 = L_{yy}/(\mathrm{n}-1)$

Statistical Inference for Correlation Coefficients

One sample *t* test for a correlation coefficient

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

Compute the sample correlation coefficient r

Compute the test statistic $t = \frac{r\sqrt{n-2}}{1-r^2}$. Which under H_o follows a *t* distribution with *n*-2 *df*.

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_o.

If $-t_{n-2,1-\alpha/2} \leq t \leq t_{n-2,1-\alpha/2}$, then accept H_o.

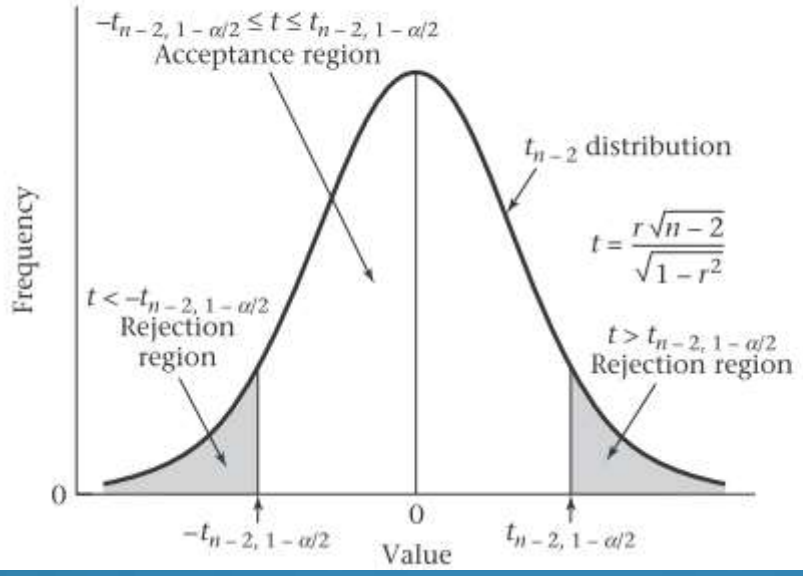
The *p*-value is given by

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

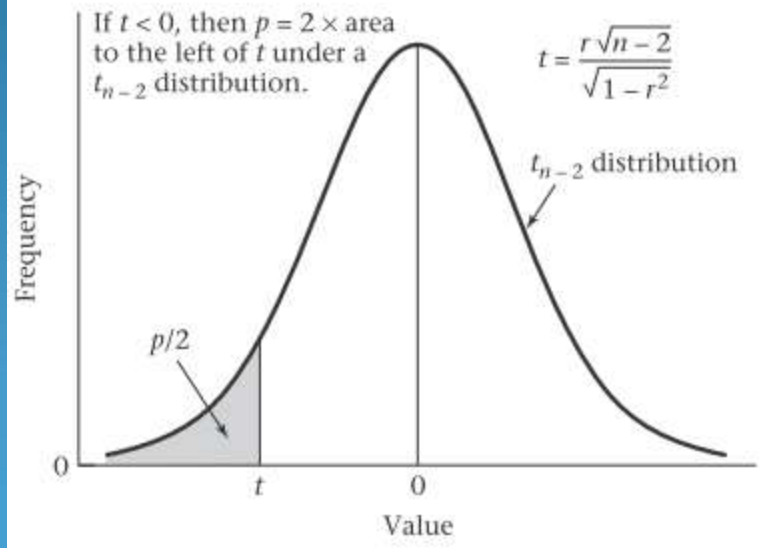
$p = 2 \times$ (area to the right of *t* under a t_{n-2} distribution) if $t \geq 0$

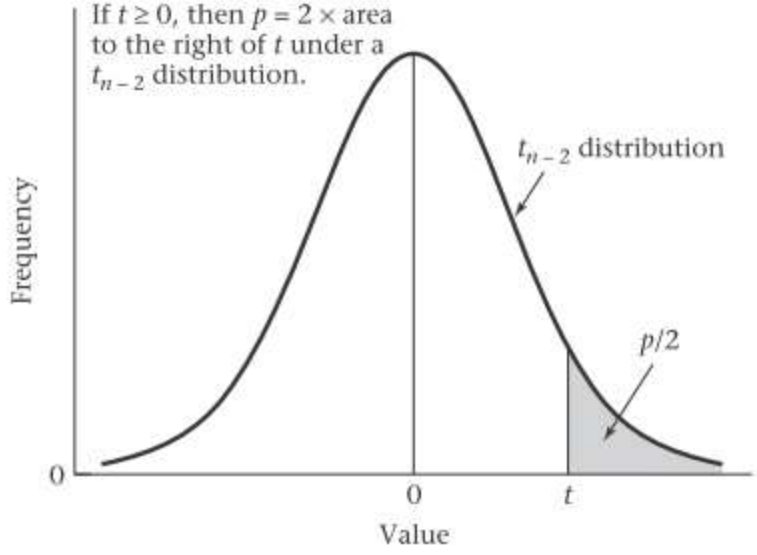
We assume an underlying normal distribution for each of the random variables used to compute *r*.

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



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





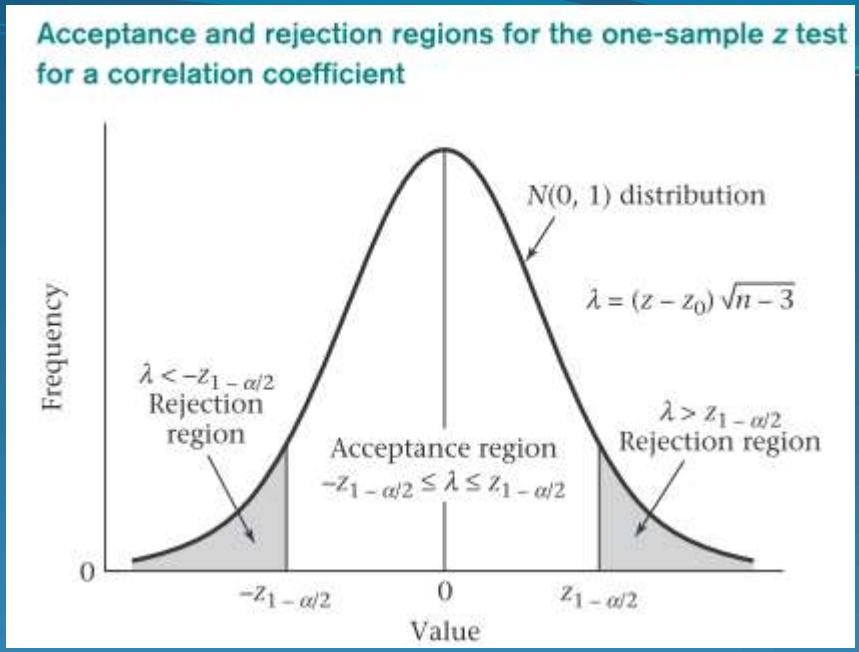
Fisher's z transformation of the sample correlation coefficient r

The z transformation of r is $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ which is approximately normally distributed under H_o with mean $z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$ and variance $\frac{1}{n-3}$.

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.

One sample z-test for a correlation coefficient

- To test the hypothesis $H_o: \rho = \rho_0$ vs. $H_1: \rho \neq \rho_0$, use the following procedure:
- Compute the sample correlation coefficient r and the z transformation of r.
- Compute the test statistic $\lambda = (z - z_o)\sqrt{n-3}$.
- 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_o.

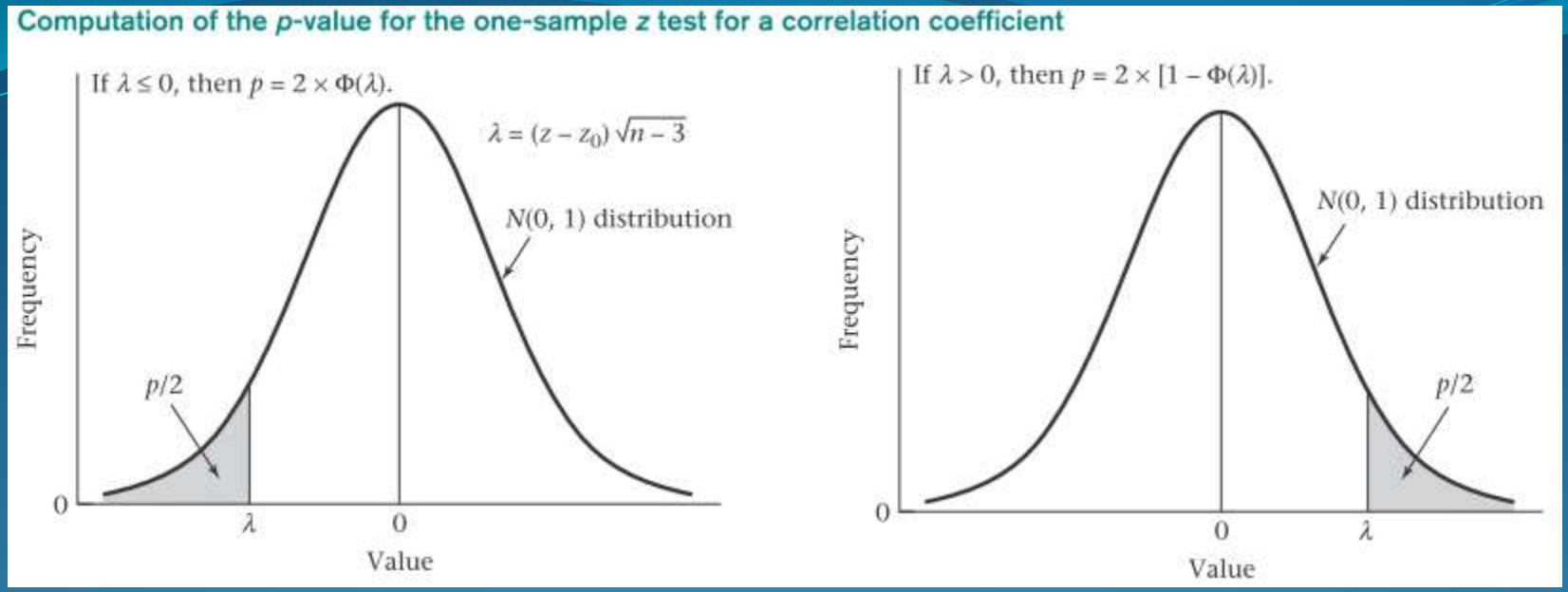


The exact *p*-value is given by

$p = 2 \times \Phi(\lambda)$ if $\lambda \le 0$

$p = 2 \times [1 - \Phi(\lambda)]$ if $\lambda > 0$

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



- The z test is used to test hypotheses about nonzero null correlations, whereas the t test 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 and is preferred.
- However, if $\rho_o \neq 0$, then the one-sample z test is very sensitive to non-normality of either x or y.

Confidence Interval for ρ:

To obtain a two-sided 100% × (1-α) confidence interval for the population correlation coefficient (ρ), we follow the following procedure:

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

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

A two-sided 100% ×(1-α) confidence interval is given for $z_\rho = (z_1, z_2)$ where

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

And $z_{1-\alpha/2}$ = 100% ×(1-α/2) percentile of an *N(0,1)* distribution.

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

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

Summary

In this chapter, we discussed

1. Scatter Plot

2. Pearson correlation methods are used to determine the association between two normally distributed variables without distinguishing between dependent and independent variables.

3. Statistical inference (testing and confidence intervals) methods for investigating the relationship between two or more variables.

The End