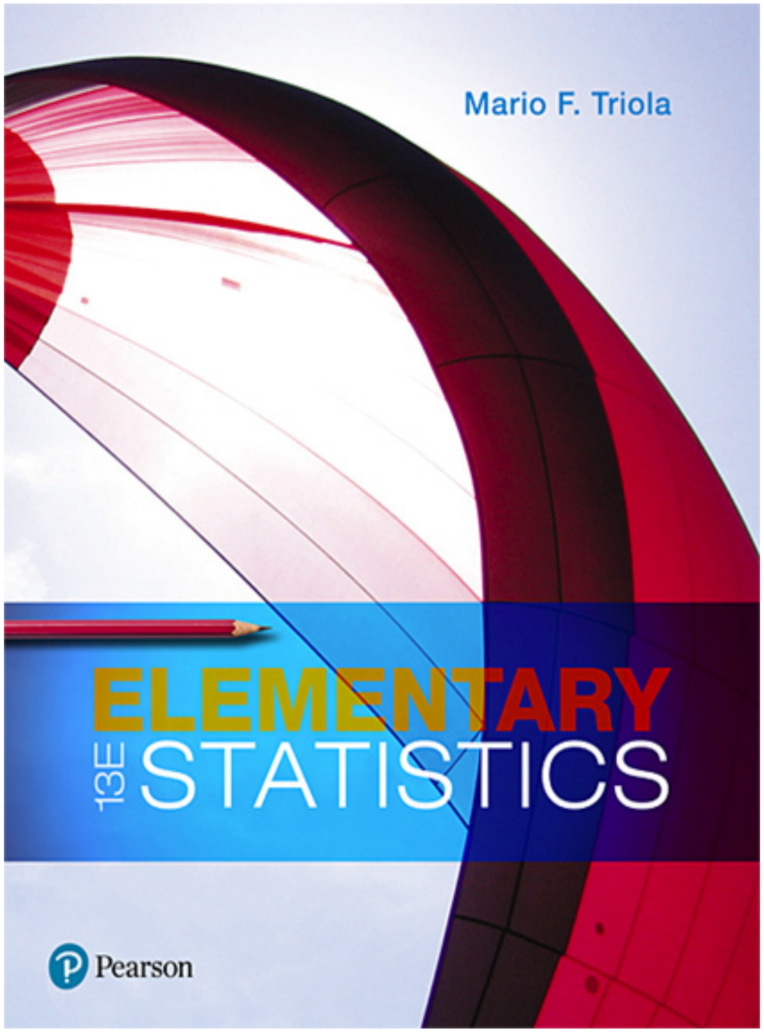
Elementary Statistics

Thirteenth Edition



Chapter 10

Correlation and Regression

Correlation and Regression

10-1 Correlation

10-2 Regression

10-3 Prediction Intervals and Variation

10-4 Multiple Regression

10-5 Nonlinear Regression

Key Concept

We introduce the **linear correlation coefficient** *r*, which is a number that measures how well paired sample data fit a straight-line pattern when graphed. We use the sample of paired data (sometimes called **bivariate data**) to find the value of *r*, and then we use *r* to decide whether there is a linear correlation between the two variables.

We consider only **linear** relationships, which means that when graphed in a scatterplot, the points approximate a **straight-line** pattern. Then, we discuss methods for conducting a formal hypothesis test that can be used to decide whether there is a linear correlation between all population values for the two variables.

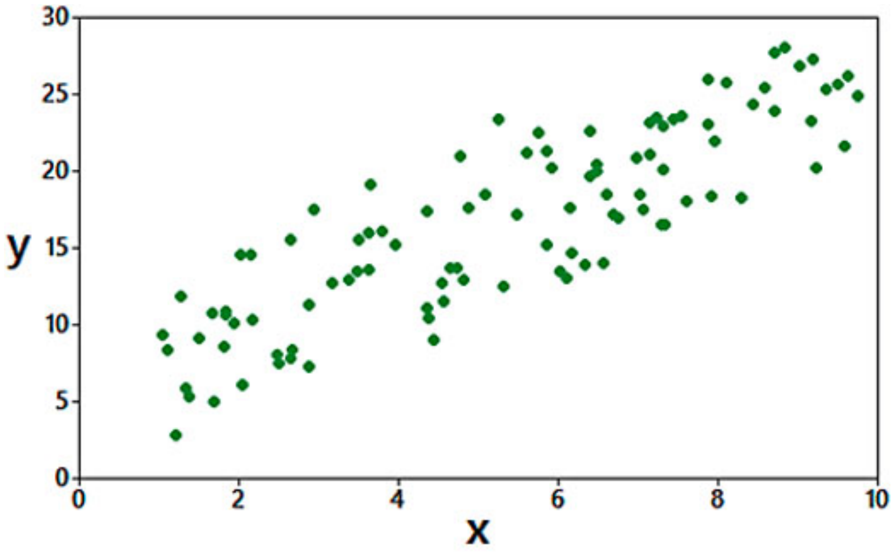
Correlation

- Correlation
  - A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

Linear Correlation

- Linear Correlation
  - A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

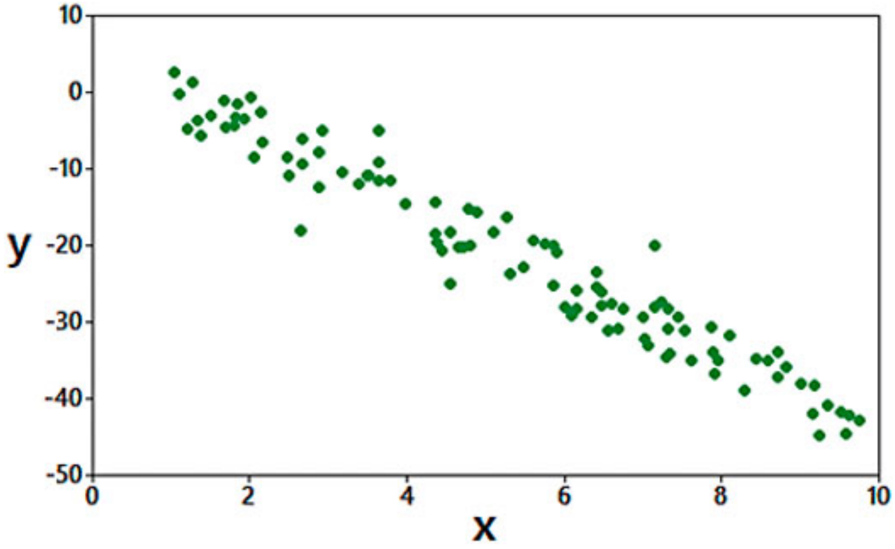
Interpreting Scatterplots (1 of 4)



(a) Positive correlation: $r = 0.859$

Distinct straight-line, or linear, pattern. We say that there is a **positive** linear correlation between *x* and *y*, since as the *x* values increase, the corresponding *y* values also increase.

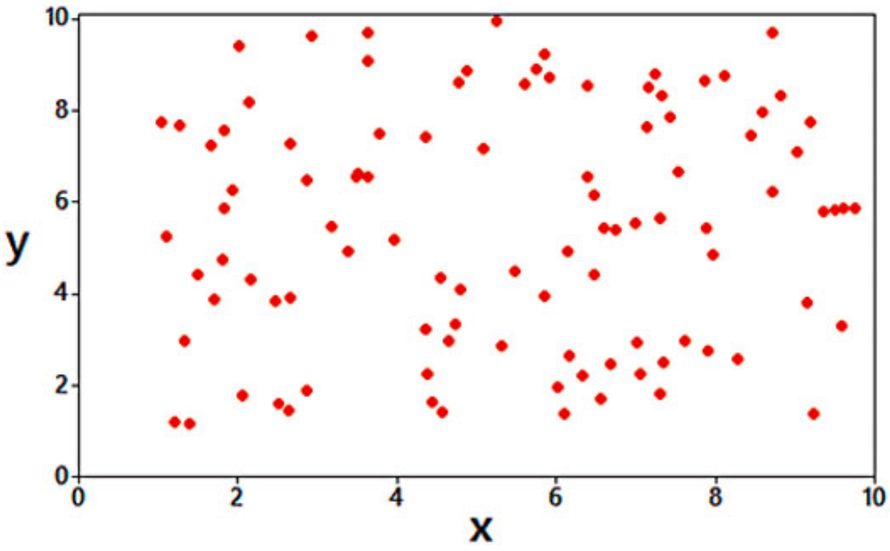
Interpreting Scatterplots (2 of 4)



(b) Negative correlation: $r = -0.971$

Distinct straight-line, or linear pattern. We say that there is a **negative** linear correlation between *x* and *y*, since as the *x* values increase, the corresponding *y* values decrease.

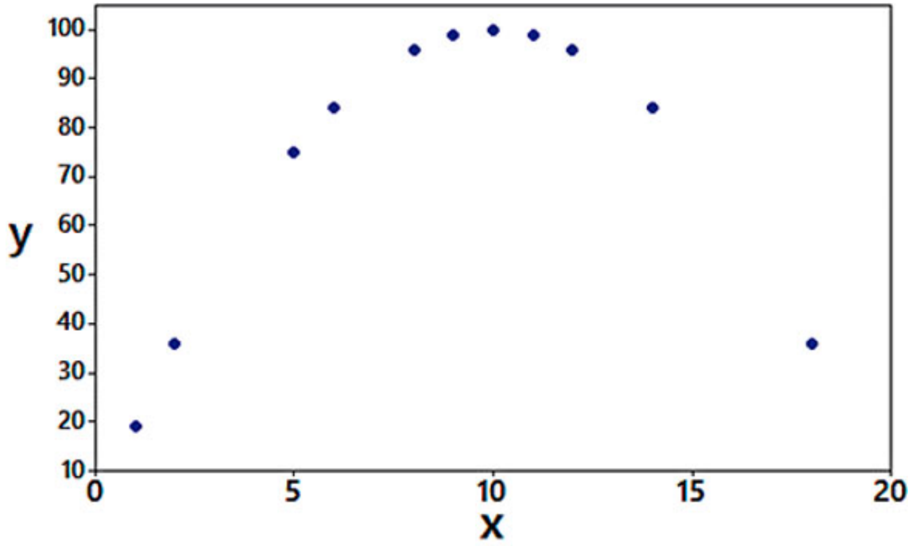
Interpreting Scatterplots (3 of 4)



(c) No correlation: $r = 0.074$

No distinct pattern, which suggests that there is no correlation between *x* and *y*.

Interpreting Scatterplots (4 of 4)

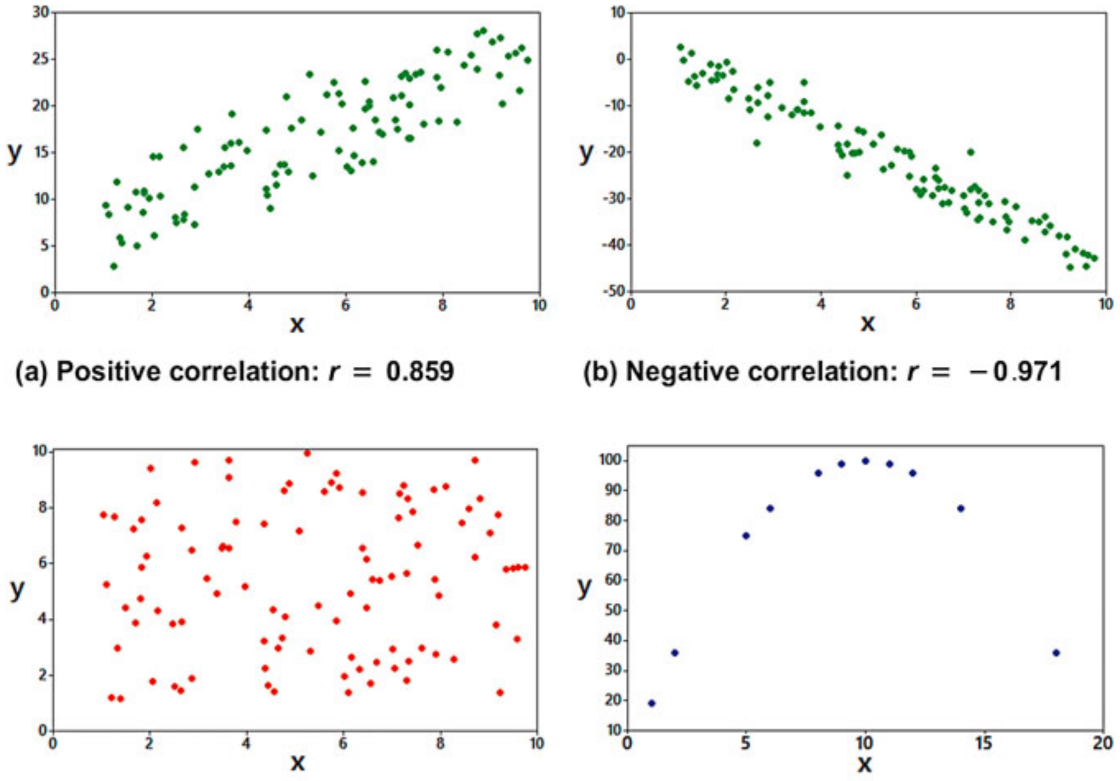


(d) Nonlinear relationship: $r = 0.330$

Distinct pattern suggesting a correlation between *x* and *y*, but the pattern is not that of a straight line.

Measure the Strength of the Linear Correlation with *r*

We use the linear correlation coefficient *r*, which is a number that measures the strength of the linear association between the two variables.



(a) Positive correlation: $r = 0.859$

(b) Negative correlation: $r = -0.971$

(c) No correlation: $r = 0.074$

(d) Nonlinear relationship: $r = 0.330$

Linear Correlation Coefficient *r*

- Linear Correlation Coefficient *r*
  - The **linear correlation coefficient *r*** measures the strength of the linear correlation between the paired quantitative *x* values and *y* values in a **sample**.

Calculating and Interpreting the Linear Correlation Coefficient *r*: Objective

Determine whether there is a linear correlation between two variables.

Calculating and Interpreting the Linear Correlation Coefficient *r*: Notation for the Linear Correlation Coefficient (1 of 2)

n number of **pairs** of sample data.

$\sum$ denotes addition of the items indicated.

$\sum x$ sum of all x values.

$\sum x^2$ indicates that each x value should be squared and then those squares added.

$(\sum x)^2$ indicates that the x values should be added and the total then squared. Avoid confusing $\sum x^2$ and $(\sum x)^2$.

Calculating and Interpreting the Linear Correlation Coefficient *r*: Notation for the Linear Correlation Coefficient (2 of 2)

$\sum xy$ indicates that each *x* value should first be multiplied by its corresponding *y* value. After obtaining all such products, find their sum.

r linear correlation coefficient for **sample** data

p linear correlation coefficient for a **population** of paired data

Calculating and Interpreting the Linear Correlation Coefficient *r*: Requirements (1 of 2)

Given any collection of sample paired quantitative data, the linear correlation coefficient *r* can always be computed, but the following requirements should be satisfied when using the sample paired data to make a conclusion about linear correlation in the corresponding population of paired data.

1. The sample of paired (*x*, *y*) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.

Calculating and Interpreting the Linear Correlation Coefficient *r*: Requirements (2 of 2)

3. Because results can be strongly affected by the presence of outliers, any outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating *r* with and without the outliers included*.

*Requirements 2 and 3 above are simplified attempts at checking this formal requirement: The pairs of (*x*, *y*) data must have a bivariate normal distribution.

Calculating and Interpreting the Linear Correlation Coefficient *r*: Formulas for Calculating *r* (1 of 2)

FORMULA 10-1

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

(Good format for calculations)

Calculating and Interpreting the Linear Correlation Coefficient *r*: Formulas for Calculating *r* (2 of 2)

FORMULA 10-2

$$r = \frac{\sum \left(z_x z_y \right)}{n-1}$$ (Good format for understanding)

where z_x denotes the z score for an individual sample value x and z_y is the z score for the corresponding sample value y.

Calculating and Interpreting the Linear Correlation Coefficient *r*: Rounding the Linear Correlation Coefficient *r*

Round the linear correlation coefficient *r* to three decimal places so that its value can be directly compared to critical values.

Calculating and Interpreting the Linear Correlation Coefficient *r*: Interpreting the Linear Correlation Coefficient *r* (1 of 2)

Using *P*-Value from Technology to Interpret *r*: Use the *P*-value and significance level α as follows:

P-value $\leq \alpha$: Supports the claim of a linear correlation.

P-value $> \alpha$: Does not support the claim of a linear correlation.

Calculating and Interpreting the Linear Correlation Coefficient *r*: Interpreting the Linear Correlation Coefficient *r* (2 of 2)

Using Table A-6 to Interpret *r*: Consider critical values from Table A-6 or technology as being both positive and negative:

- **Correlation** If $|r| \geq$ critical value, conclude that there is sufficient evidence to support the claim of a linear correlation.
- **No Correlation** If $|r| <$ critical value, conclude that there is not sufficient evidence to support the claim of a linear correlation.

Properties of the Linear Correlation Coefficient *r* (1 of 2)

1. The value of *r* is always between −1 and 1 inclusive. That is, $-1 \leq r \leq 1$.
2. If all values of either variable are converted to a different scale, the value of *r* does not change.
3. The value of *r* is not affected by the choice of *x* or *y*. Interchange all *x* values and *y* values, and the value of *r* will not change.

Properties of the Linear Correlation Coefficient *r* (2 of 2)

4. *r* measures the strength of a linear relationship. It is not designed to measure the strength of a relationship that is not linear.
5. *r* is very sensitive to outliers in the sense that a single outlier could dramatically affect its value.

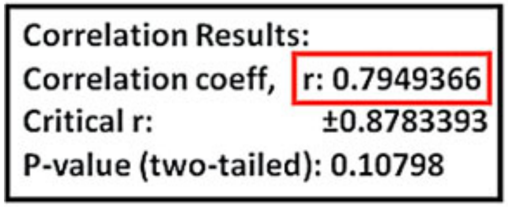
Example: Finding *r* Using Technology (1 of 2)

The table lists five paired data values. Use technology to find the value of the correlation coefficient *r* for the data.

Chocolate	5	6	4	4	5
Nobel	6	9	3	2	11

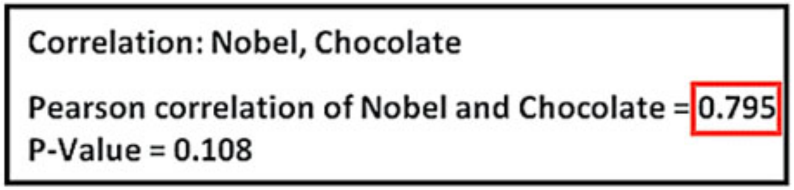
Example: Finding *r* Using Technology (2 of 2)

Statdisk



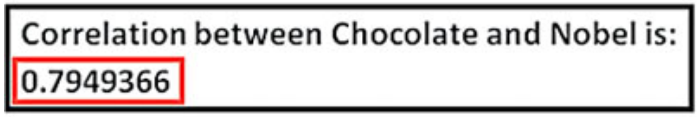
```
Correlation Results:
Correlation coeff,    r: 0.7949366
Critical r:          ±0.8783393
P-value (two-tailed): 0.10798
```

Minitab



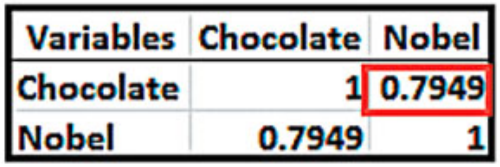
```
Correlation: Nobel, Chocolate

Pearson correlation of Nobel and Chocolate = 0.795
P-Value = 0.108
```

StatCrunch



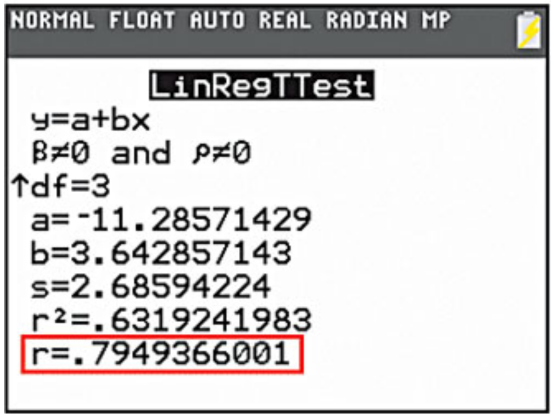
```
Correlation between Chocolate and Nobel is:
0.7949366
```

XLSTAT



Variables	Chocolate	Nobel
Chocolate	1	0.7949
Nobel	0.7949	1

TI-83/84 Plus



```
NORMAL FLOAT AUTO REAL RADIAN MP

        LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑df=3
 a=⁻11.28571429
 b=3.642857143
 s=2.68594224
 r²=.6319241983
 r=.7949366001
```

Solution

The value of *r* will be automatically calculated with software or a calculator. $r = 0.795$ (rounded).

Example: Finding *r* Using Formula 10-1 (1 of 3)

Use Formula 10-1 to find the value of the linear correlation coefficient *r* for the five pairs of chocolate/Nobel data listed in the table.

Chocolate	5	6	4	4	5
Nobel	6	9	3	2	11

Example: Finding *r* Using Formula 10-1 (2 of 3)

Solution

The value of *r* is calculated as shown below. Here, the variable *x* is used for the chocolate values, and the variable *y* is used for the Nobel values. Because there are five pairs of data, *n* = 5.

x **(Chocolate)**	***y*** **(Nobel)**	x^2	y^2	xy
5	6	25	36	30
6	9	36	81	54
4	3	16	9	12
4	2	16	4	8
5	11	25	121	55
$\sum x = 24$	$\sum y = 31$	$\sum x^2 = 118$	$\sum y^2 = 251$	$\sum xy = 159$

Example: Finding *r* Using Formula 10-1 (3 of 3)

Solution

$$r = \frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right)-\left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right)-\left(\sum y\right)^2}}$$

$$= \frac{5(159)-(24)(31)}{\sqrt{5(118)-(24)^2}\sqrt{5(251)-(31)^2}}$$

$$= \frac{51}{\sqrt{14}\sqrt{294}} = 0.795$$

Example: Finding *r* Using Formula 10-2 (1 of 5)

Use Formula 10-2 to find the value of the linear correlation coefficient *r* for the five pairs of chocolate/Nobel data listed in the table.

Chocolate	5	6	4	4	5
Nobel	6	9	3	2	11

Example: Finding *r* Using Formula 10-2 (2 of 5)

Solution

Formula 10-2 has the advantage of making it easier to **understand** how *r* works. The variable *x* is used for the chocolate values, and the variable *y* is used for the Nobel values. Each sample value is replaced by its corresponding *z* score.

Example: Finding *r* Using Formula 10-2 (3 of 5)

Solution

For example, using unrounded numbers, the chocolate values have a mean of $\bar{x} = 4.8$ and a standard deviation of $s_x = 0.836660$, so the first chocolate value of 5 is converted to a *z* score of 0.239046 as shown here:

$$z_x = \frac{x - \bar{x}}{s_x} = \frac{5 - 4.8}{0.836660} = 0.239046$$

Example: Finding *r* Using Formula 10-2 (4 of 5)

Solution

The *z* scores for all of the chocolate values (see the third column) and the *z* scores for all of the Nobel values (see the fourth column) are below. The last column lists the products $z_x \cdot z_y$.

x **(Chocolate)**	***y*** **(Nobel)**	z_x	z_y	$z_x \cdot z_y$
5	6	0.239046	−0.052164	−0.012470
6	9	1.434274	0.730297	1.047446
4	3	−0.956183	−0.834625	0.798054
4	2	−0.956183	−1.095445	1.047446
5	11	0.239046	1.251937	0.299270
				$\sum (z_x \cdot z_y) = 3.179746$

Example: Finding *r* Using Formula 10-2 (5 of 5)

Solution

Using $\sum(z_x \cdot z_y) = 3.179746$, the value of *r* is calculated as shown below.

$$r = \frac{\Sigma(z_x \cdot z_y)}{n-1} = \frac{3.179746}{4} = 0.795$$

Example: Is There a Linear Correlation? (1 of 12)

Using the value of $r = 0.801$ for the 23 pairs of data shown on the next slide, and using a significance level of 0.05, is there sufficient evidence to support a claim that there is a linear correlation between chocolate consumption and numbers of Nobel Laureates?

Example: Is There a Linear Correlation? (2 of 12)

Chocolate	Nobel
4.5	5.5
10.2	24.3
4.4	8.6
2.9	0.1
3.9	6.1
0.7	0.1
8.5	25.3
7.3	7.6
6.3	9.0
11.6	12.7
2.5	1.9
8.8	12.7

Example: Is There a Linear Correlation? (3 of 12)

Chocolate	Nobel
3.7	5.5
1.8	1.5
4.5	11.4
9.4	25.5
3.6	3.1
2.0	1.9
3.6	1.7
6.4	31.9
11.9	31.5
9.7	18.9
5.3	10.8

Example: Is There a Linear Correlation? (4 of 12)

Solution

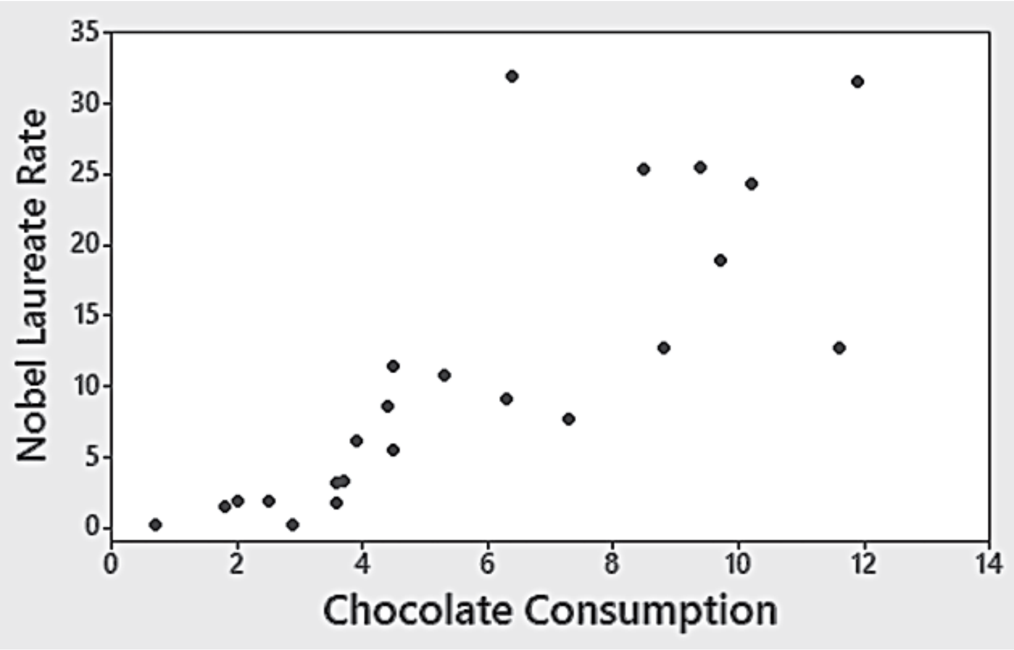
Requirement Check The first requirement of a simple random sample of quantitative data is questionable. The data are quantitative, but examining the original data, it does not appear that they are randomly selected pairs. Because this requirement is not satisfied, the results obtained may not be valid.

Example: Is There a Linear Correlation? (5 of 12)

Solution

Requirement Check

The second requirement of a scatterplot showing a straight-line pattern is satisfied. The scatterplot shows that the third requirement of no outliers is satisfied.



Example: Is There a Linear Correlation? (6 of 12)

Solution

Using *P*-Value from Technology to Interpret *r*: Use the *P*-value and significance level α as follows:

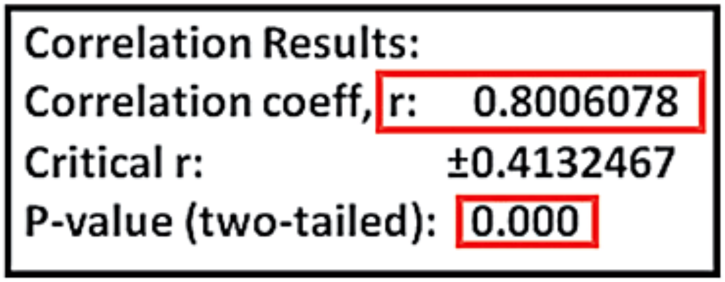
- *P*-value $\le \alpha$: Supports the claim of a linear correlation.
- *P*-value $> \alpha$: Does not support the claim of a linear correlation.

Example: Is There a Linear Correlation? (7 of 12)

Solution

The Stat disk display shows that the *P*-value is 0.000 when rounded. Because the *P*-value is less than or equal to the significance level of 0.05, we conclude there is sufficient evidence to support the conclusion that for countries, there is a linear correlation between chocolate consumption and numbers of Nobel Laureates.

Statdisk

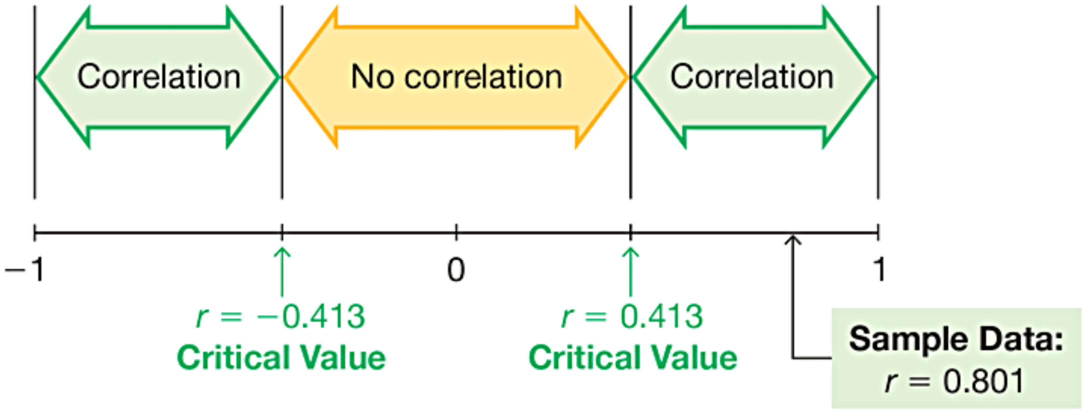


```
Correlation Results:
Correlation coeff, r:     0.8006078
Critical r:              ±0.4132467
P-value (two-tailed):  0.000
```

Example: Is There a Linear Correlation? (8 of 12)

Solution

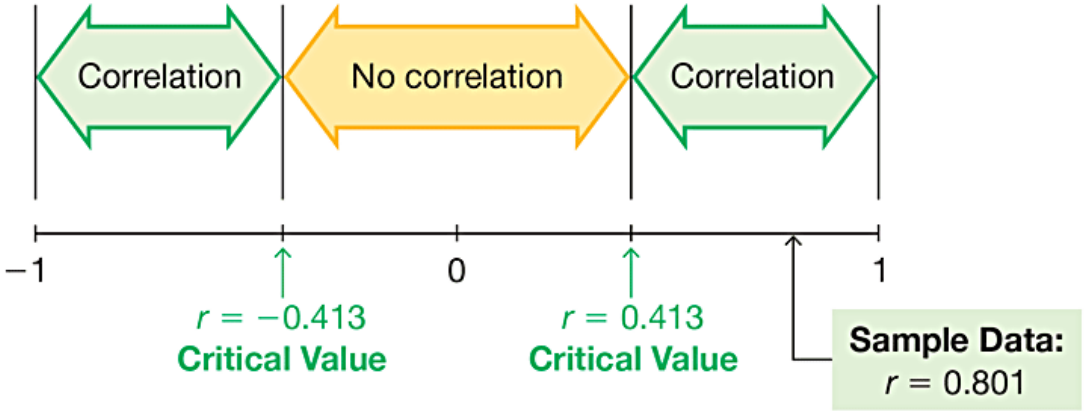
Using Table A-6 to Interpret *r*: Consider critical values from Table A-6 as being both positive and negative, and draw a graph similar to the figure shown.



Example: Is There a Linear Correlation? (9 of 12)

Solution

For the 23 pairs of data, Table A-6 yields a critical value that is between r = 0.396 and r = 0.444; technology yields a critical value of r = 0.413. We can now compare the computed value of r = 0.801 to the critical values of r = ±0.413, as shown.



Example: Is There a Linear Correlation? (10 of 12)

Solution

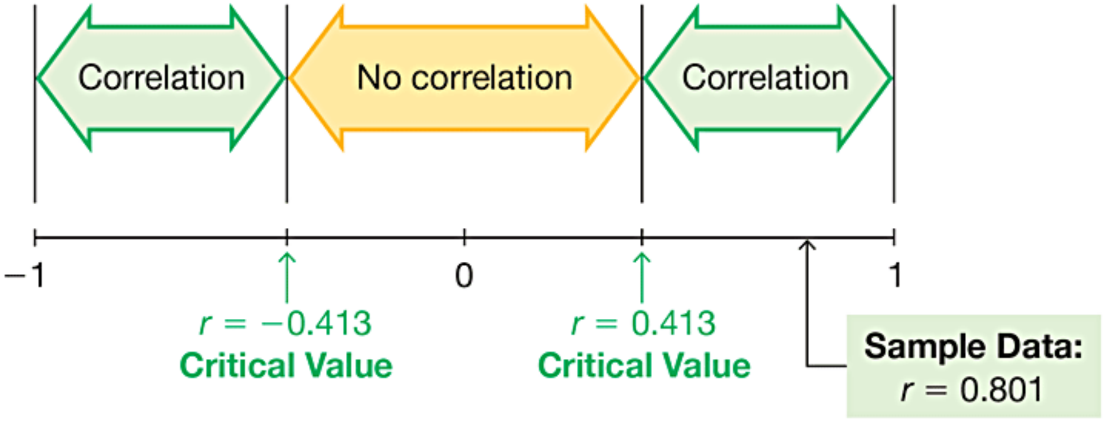
Correlation If the computed linear correlation coefficient *r* lies in the left or right tail region beyond the critical value for that tail, conclude that there is sufficient evidence to support the claim of a linear correlation.

No Correlation If the computed linear correlation coefficient lies between the two critical values, conclude that there is not sufficient evidence to support the claim of a linear correlation.

Example: Is There a Linear Correlation? (11 of 12)

Solution

Because the figure shows that the computed value of $r = 0.801$ lies beyond the upper critical value, we conclude that there is sufficient evidence to support the claim of a linear correlation between chocolate consumption and number of Nobel Laureates for different countries.



Example: Is There a Linear Correlation? (12 of 12)

Interpretation

Although we have found a linear correlation, it would be absurd to think that eating more chocolate would help win a Nobel Prize. See the following discussion under “Interpreting *r* with Causation: Don’t Go There!” Also, the requirement of a simple random sample is not satisfied, so the conclusion of a linear correlation is questionable.

Interpreting *r*: Explained Variation

The value of r^2 is the proportion of the variation in *y* that is explained by the linear relationship between *x* and *y*.

Example: Explained Variation (1 of 3)

Using the 23 pairs of chocolate/Nobel data, we get $r = 0.801$. What proportion of the variation in numbers of Nobel Laureates can be explained by the variation in the consumption of chocolate?

Example: Explained Variation (2 of 3)

Solution

With $r = 0.801$ we get $r^2 = 0.642$.

Example: Explained Variation (3 of 3)

Interpretation

We conclude that 0.642 (or about 64%) of the variation in numbers of Nobel Laureates can be explained by the linear relationship between chocolate consumption and numbers of Nobel Laureates. This implies that about 36% of the variation in numbers of Nobel Laureates cannot be explained by rates of chocolate consumption.

Interpreting *r* with Causation: Don’t Go There!

Correlation does not imply causality!

We noted previously that we should use common sense when interpreting results. Clearly, it would be absurd to think that eating more chocolate would help win a Nobel Prize

Common Errors Involving Correlation

1. Assuming that correlation implies causality
2. Using data based on averages
3. Ignoring the possibility of a nonlinear relationship

Formal Hypothesis Test (1 of 2)

Hypotheses If conducting a formal hypothesis test to determine whether there is a significant linear correlation between two variables, use the following null and alternative hypotheses that use ρ to represent the linear correlation coefficient of the population:

Null Hypothesis H_0: $\rho = 0$ (No correlation)

Alternative Hypothesis H_1: $\rho \neq 0$ (Correlation)

Formal Hypothesis Test (2 of 2)

Test Statistic The same methods of Part 1 can be used with the test statistic *r*, or the *t* test statistic can be found using the following:

$$\textbf{Test Statistic } t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

(with $n - 2$ degrees of freedom)

Example: Hypothesis Test Using the *P*-Value from the *t* Test (1 of 5)

Use the paired chocolate/Nobel data to conduct a formal hypothesis test of the claim that there is a linear correlation between the two variables. Use a 0.05 significance level with the *P*-value method of testing hypotheses.

Example: Hypothesis Test Using the *P*-Value from the *t* Test (2 of 5)

Solution

REQUIREMENT CHECK The requirements were addressed. To claim that there is a linear correlation is to claim that the population linear correlation coefficient ρ is different from 0. We therefore have the following hypotheses:

H_0: $\rho = 0$ (No correlation)

H_1: $\rho \neq 0$ (Correlation)

Example: Hypothesis Test Using the *P*-Value from the *t* Test (3 of 5)

Solution

The linear correlation coefficient is $r = 0.801$ (from technology) and $n = 23$ (because there are 23 pairs of sample data), so the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.801}{\sqrt{\frac{1-0.801^2}{23-2}}} = 6.131$$

Example: Hypothesis Test Using the *P*-Value from the *t* Test (4 of 5)

Solution

Technologies use more precision to obtain the more accurate test statistic of $t = 6.123$. With $n - 2 = 21$ degrees of freedom, Table A-3 shows that the test statistic of $t = 6.123$ yields a *P*-value that is less than 0.01. Technologies show that the *P*-value is 0.000 when rounded. Because the *P*-value of 0.000 is less than the significance level of 0.05, we reject H_0. ("If the *P* is low, the null must go." The *P*-value of 0.000 is low.)

Example: Hypothesis Test Using the *P*-Value from the *t* Test (5 of 5)

Interpretation

We conclude that for countries, there is sufficient evidence to support the claim of a linear correlation between chocolate consumption and Nobel Laureates.

One-Tailed Tests

The examples and exercises in this section generally involve two-tailed tests, but one-tailed tests can occur with a claim of a positive linear correlation or a claim of a negative linear correlation. In such cases, the hypotheses will be as shown here.

Claim of Negative Correlation (Left-Tailed Test)	**Claim of Positive Correlation (Right-Tailed Test)**
$H_0 : \rho = 0$	$H_0 : \rho = 0$
$H_1 : \rho < 0$	$H_1 : \rho > 0$