Correlation

Colby Community College

Definition

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

Definition

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

Definition

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.

Definition

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

Definition

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.

Definition

A **scatterplot** is a plot of paired (x, y) quantitative data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first variable (x), and the vertical axis for the second variable (y).

Definition

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

Definition

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.

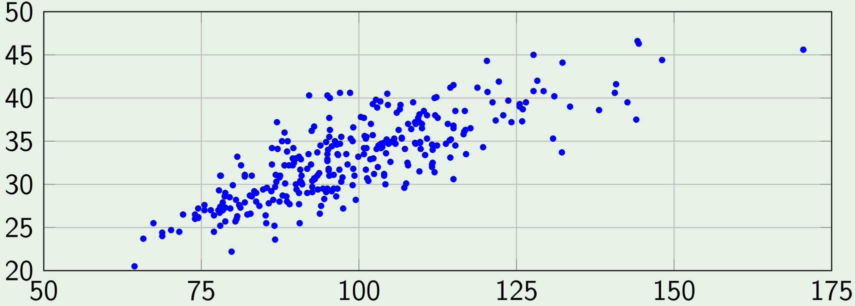
Definition

A **scatterplot** is a plot of paired (x, y) quantitative data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first variable (x), and the vertical axis for the second variable (y).

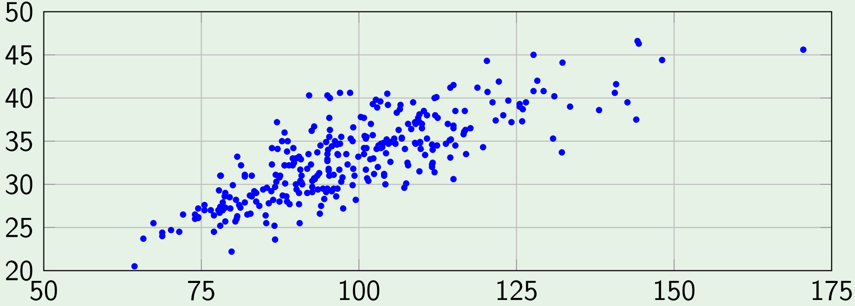
Note

Because conclusions based on visual examinations of scatterplots are largely subjective, we need more objective measures. We use the linear correlation coefficient r, which is a number that measures the strength of the linear association between the two variables.

Example 1

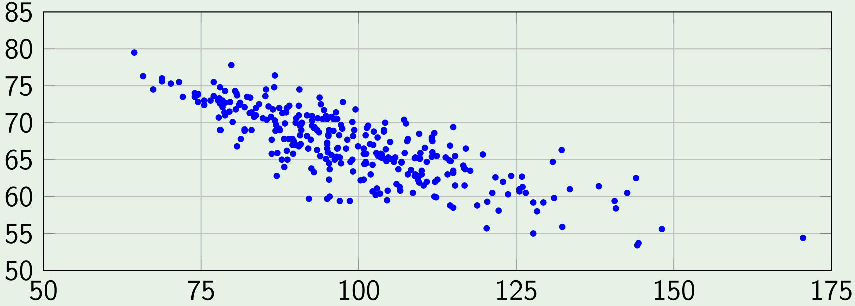


Example 1

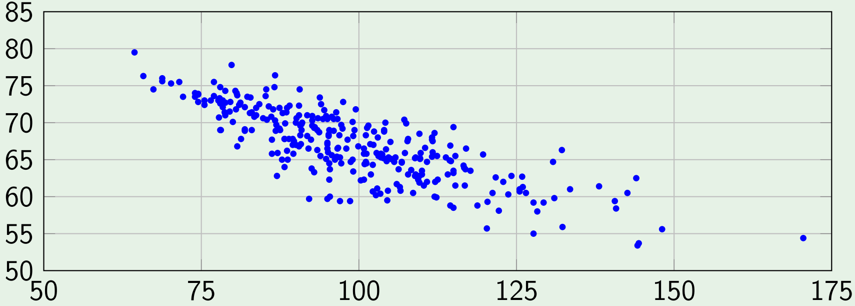


Distinct straight-line, or linear, pattern. We say that there is a positive linear correlation ($r = 0.80241$) between x and y, since as the x values increase, the corresponding y values also increase.

Example 2

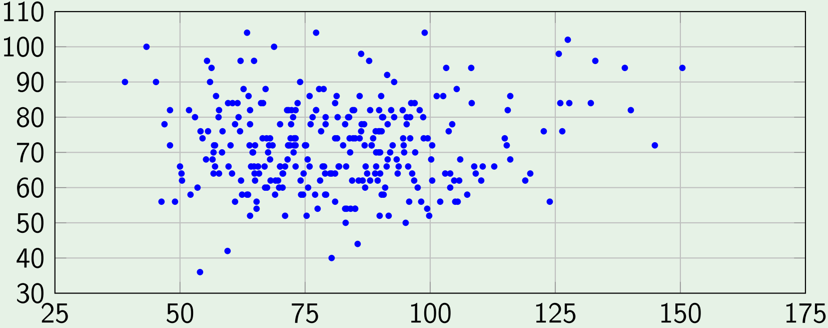


Example 2

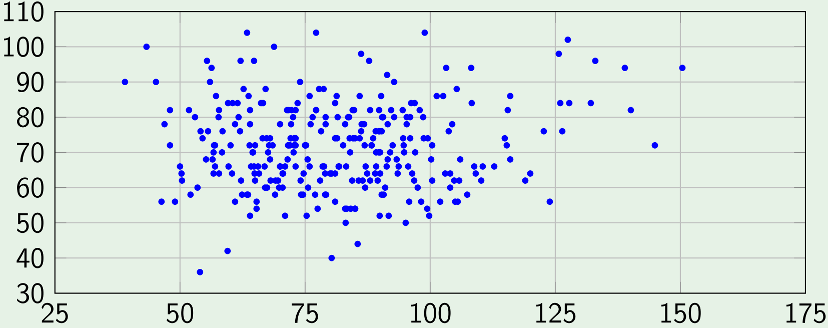


Distinct straight-line, or linear, pattern. We say that there is a positive linear correlation ($r = -0.80241$) between x and y, since as the x values increase, the corresponding y values also increase.

Example 3



Example 3



The points do not show any obvious pattern ($r = 0.08161$), and this lack of a pattern suggests that there is no relationship between the variables.

Definition

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

Definition

The **linear correlation coefficient** $\boldsymbol{r}$ measures the strength of the linear correlation between the paired quantitative x values and y values in a sample.

Requirements

The following 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 data is a simple random of quantitative data.

Definition

The **linear correlation coefficient** $\boldsymbol{r}$ measures the strength of the linear correlation between the paired quantitative x values and y values in a sample.

Requirements

The following 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 data is a simple random of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.

Definition

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

Requirements

The following 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 data is a simple random of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. Because the results can be strongly affected by outliers, any outliers mist be removed if they are known to be errors.

Caution

A linear correlation coefficient r can always be calculated, wether or not it applies.

Formula

The formula for r is:

$$r = \frac{\sum(z_x z_y)}{n-1}$$

where z_x and z_y are the z-scores for the sample values x and y, respectively.

Formula

The formula for r is:

$$r = \frac{\sum(z_x z_y)}{n-1}$$

where z_x and z_y are the z-scores for the sample values x and y, respectively.

Note

Technology is almost always used to generate r.

Formula

The formula for r is:

$$r = \frac{\sum(z_x z_y)}{n-1}$$

where z_x and z_y are the z-scores for the sample values x and y, respectively.

Note

Technology is almost always used to generate r.

Alternative Formula

A formula for r that is better for hand calculations is:

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

Properties of the Linear Correlation Coefficient

- $-1 \leq r \leq 1$

Properties of the Linear Correlation Coefficient

- $-1 \leq r \leq 1$
- If all values of either variable are converted to a different scale, the value of r does not change.

Properties of the Linear Correlation Coefficient

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

Properties of the Linear Correlation Coefficient

- $-1 \leq r \leq 1$
- If all values of either variable are converted to a different scale, the value of r does not change.
- The value of r is not affected by the choice of x or y. Interchange all x values and all y values, and the value of r will not change.
- r measures the strength of a linear relationship. It is not designed to measure the strength of a relationship that is nor linear.

Properties of the Linear Correlation Coefficient

- $-1 \leq r \leq 1$
- If all values of either variable are converted to a different scale, the value of r does not change.
- The value of r is not affected by the choice of x or y. Interchange all x values and all y values, and the value of r will not change.
- r measures the strength of a linear relationship. It is not designed to measure the strength of a relationship that is nor linear.
- r is very sensitive to outliers in the sense that a single outlier could dramatically affect its value.

Is There a Linear Correlation?

Technology will generate a P-value along with r. If we have a significance level α, then

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

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

Is There a Linear Correlation?

Technology will generate a P-value along with r. If we have a significance level α, then

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

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

Example 4

For Data Set 2 "Foot and Height," when we use technology to calculate the linear correlation between the foot length and age of 40 randomly selected prople we get:

$$r = 0.3591 \quad \text{and} \quad P\text{-value} = 0.02287$$

Is There a Linear Correlation?

Technology will generate a P-value along with r. If we have a significance level α, then

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

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

Example 4

For Data Set 2 "Foot and Height," when we use technology to calculate the linear correlation between the foot length and age of 40 randomly selected prople we get:

$$r = 0.3591 \quad \text{and} \quad P\text{-value} = 0.02287$$

If we have a significance level $\alpha = 0.05$, is there evidence of a linear correlation?

Is There a Linear Correlation?

Technology will generate a P-value along with r. If we have a significance level α, then

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

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

Example 4

For Data Set 2 "Foot and Height," when we use technology to calculate the linear correlation between the foot length and age of 40 randomly selected prople we get:

$$r = 0.3591 \quad \text{and} \quad P\text{-value} = 0.02287$$

If we have a significance level $\alpha = 0.05$, is there evidence of a linear correlation?

Because $0.02287 \leq 0.05$ we have evidence of a linear correlation.

Note

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

Note

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 5

From Example 4 we say that $r = 0.3591$ when comparing foot length to age.

Note

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 5

From Example 4 we say that $r = 0.3591$ when comparing foot length to age.

We can then calculate $r^2 = (0.3591)^2 = 0.129$.

Note

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 5

From Example 4 we say that $r = 0.3591$ when comparing foot length to age.

We can then calculate $r^2 = (0.3591)^2 = 0.129$.

We can conclude that about 12.9% of the variation in ages can be explained by the linear relationship between foot length and age.

Note

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 5

From Example 4 we say that $r = 0.3591$ when comparing foot length to age.

We can then calculate $r^2 = (0.3591)^2 = 0.129$.

We can conclude that about 12.9% of the variation in ages can be explained by the linear relationship between foot length and age.

this implies that about 87.1% of the variation in ages cannot be explained by rates of chocolate consumption.

Do Not Assume That Correlation Implies Causality!

For several years, the data suggested a linear correlations between the stork population in Copenhagen and the number of human births.

Do Not Assume That Correlation Implies Causality!

For several years, the data suggested a linear correlations between the stork population in Copenhagen and the number of human births.

Storks do not actually have anything to do with human births. This means both variables were affected by another factor.

Do Not Assume That Correlation Implies Causality!

For several years, the data suggested a linear correlations between the stork population in Copenhagen and the number of human births.

Storks do not actually have anything to do with human births. This means both variables were affected by another factor.

Definition

A **lurking variable** is one that affects the variables being studied but is nor included in the study.

Do Not Assume That Correlation Implies Causality!

For several years, the data suggested a linear correlations between the stork population in Copenhagen and the number of human births.

Storks do not actually have anything to do with human births. This means both variables were affected by another factor.

Definition

A **lurking variable** is one that affects the variables being studied but is nor included in the study.

Do Not Use Data Based on Averages

Averages suppress individual variation and may inflate the correlation coefficient.

Do Not Assume That Correlation Implies Causality!

For several years, the data suggested a linear correlations between the stork population in Copenhagen and the number of human births.

Storks do not actually have anything to do with human births. This means both variables were affected by another factor.

Definition

A **lurking variable** is one that affects the variables being studied but is nor included in the study.

Do Not Use Data Based on Averages

Averages suppress individual variation and may inflate the correlation coefficient.

Do Not Ignore The Possibility of a Nonlinear Relationship

If there is no linear correlation, there might be some correlation that is not linear.

Formal Hypothesis Test

If conducting a formal hypothesis test to determine whether there is a significant linear correlation between two variables, use:

$$H_0: \; p = 0 \text{ (There is no linear correlation.)}$$
$$H_A: \; p \neq 0 \text{ (There is a linear correlation.)}$$

Formal Hypothesis Test

If conducting a formal hypothesis test to determine whether there is a significant linear correlation between two variables, use:

$$H_0: \; p = 0 \text{ (There is no linear correlation.)}$$
$$H_A: \; p \neq 0 \text{ (There is a linear correlation.)}$$

The test statistic (with $n-2$ degrees of freedom) is

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

Example 6

Using Data Set 16 let us conduct a formal hypothesis test of the claim that there is a linear correlation between chocolate consumption in a country and how many Nobel Laureates are from that country using a 0.05 significance.

Example 6

Using Data Set 16 let us conduct a formal hypothesis test of the claim that there is a linear correlation between chocolate consumption in a country and how many Nobel Laureates are from that country using a 0.05 significance.

The hyoptheses we are testing are

$$H_0: \ p = 0 \text{ (There is no linear correlation.)}$$
$$H_A: \ p \neq 0 \text{ (There is a linear correlation.)}$$

Example 6

Using Data Set 16 let us conduct a formal hypothesis test of the claim that there is a linear correlation between chocolate consumption in a country and how many Nobel Laureates are from that country using a 0.05 significance.

The hyoptheses we are testing are

$$H_0: \ p = 0 \text{ (There is no linear correlation.)}$$
$$H_A: \ p \neq 0 \text{ (There is a linear correlation.)}$$

Technology give $r = 0.801$ and $n = 23$ which means the test statistic is

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

Example 6

Using Data Set 16 let us conduct a formal hypothesis test of the claim that there is a linear correlation between chocolate consumption in a country and how many Nobel Laureates are from that country using a 0.05 significance.

The hyoptheses we are testing are

$$H_0: \ p = 0 \text{ (There is no linear correlation.)}$$
$$H_A: \ p \neq 0 \text{ (There is a linear correlation.)}$$

Technology give $r = 0.801$ and $n = 23$ which means the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.891}{\sqrt{\frac{1-(0.801)^2}{23-2}}}$$

Example 6

Using Data Set 16 let us conduct a formal hypothesis test of the claim that there is a linear correlation between chocolate consumption in a country and how many Nobel Laureates are from that country using a 0.05 significance.

The hyoptheses we are testing are

$$H_0: \ p = 0 \text{ (There is no linear correlation.)}$$
$$H_A: \ p \neq 0 \text{ (There is a linear correlation.)}$$

Technology give $r = 0.801$ and $n = 23$ which means the test statistic is

$$t = \frac{r}{\sqrt{\dfrac{1-r^2}{n-2}}} = \frac{0.891}{\sqrt{\dfrac{1-(0.801)^2}{23-2}}} = 6.131$$

Example 6

Using Data Set 16 let us conduct a formal hypothesis test of the claim that there is a linear correlation between chocolate consumption in a country and how many Nobel Laureates are from that country using a 0.05 significance.

The hyoptheses we are testing are

$$H_0: \ p = 0 \text{ (There is no linear correlation.)}$$
$$H_A: \ p \neq 0 \text{ (There is a linear correlation.)}$$

Technology give $r = 0.801$ and $n = 23$ which means the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.891}{\sqrt{\frac{1-(0.801)^2}{23-2}}} = 6.131$$

We have $\mathrm{df} = 23 - 2 = 21$, technology gives a P-value of 0.000.

Example 6

Using Data Set 16 let us conduct a formal hypothesis test of the claim that there is a linear correlation between chocolate consumption in a country and how many Nobel Laureates are from that country using a 0.05 significance.

The hyoptheses we are testing are

$$H_0 : \; p = 0 \text{ (There is no linear correlation.)}$$
$$H_A : \; p \neq 0 \text{ (There is a linear correlation.)}$$

Technology give $r = 0.801$ and $n = 23$ which means the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.891}{\sqrt{\frac{1-(0.801)^2}{23-2}}} = 6.131$$

We have $\mathrm{df} = 23 - 2 = 21$, technology gives a P-value of 0.000.

Because the P-value is less than the significance level, we reject H_0.

Note

Hypotheses tests involving linear correlations are almost always two-tailed, but occasionally one-tailed tests can occur.

Note

Hypotheses tests involving linear correlations are almost always two-tailed, but occasionally one-tailed tests can occur.

Claim of Negative Correlation

When testing a claim of a negative linear correlation use

$$H_0: \ p = 0$$
$$H_A: \ p < 0$$

Note

Hypotheses tests involving linear correlations are almost always two-tailed, but occasionally one-tailed tests can occur.

Claim of Negative Correlation

When testing a claim of a negative linear correlation use

$$H_0: \; p = 0$$
$$H_A: \; p < 0$$

Claim of Positive Correlation

When testing a claim of a positive linear correlation use

$$H_0: \; p = 0$$
$$H_A: \; p > 0$$