Chapter 9

Correlation and Regression

Chapter Outline

- 9.1 Correlation
- 9.2 Linear Regression
- 9.3 Measures of Regression and Prediction Intervals
- 9.4 Multiple Regression

Section 9.1

Correlation

Section 9.1 Objectives

- Introduce linear correlation, independent and dependent variables, and the types of correlation
- Find a correlation coefficient
- Test a population correlation coefficient ρ using a table
- Perform a hypothesis test for a population correlation coefficient ρ
- Distinguish between correlation and causation

Correlation

Correlation

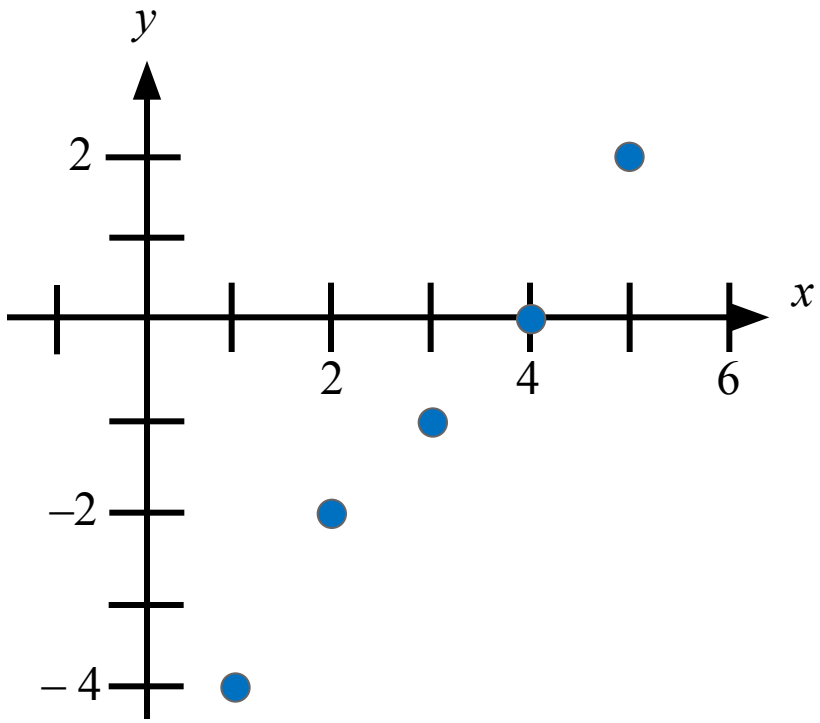
- A relationship between two variables.
- The data can be represented by ordered pairs (x, y)
  - x is the **independent** (or **explanatory**) **variable**
  - y is the **dependent** (or **response**) **variable**

Correlation

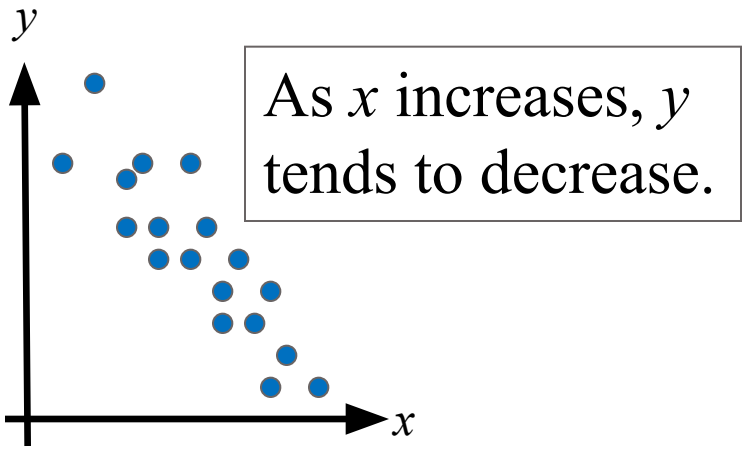
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

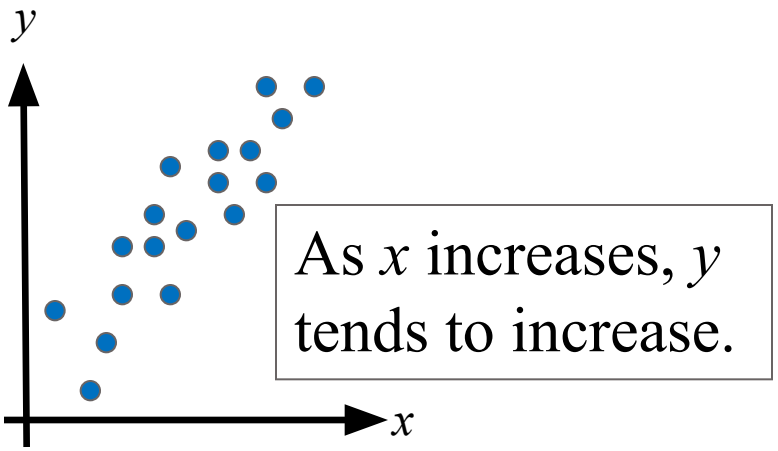
x	1	2	3	4	5
y	-4	-2	-1	0	2



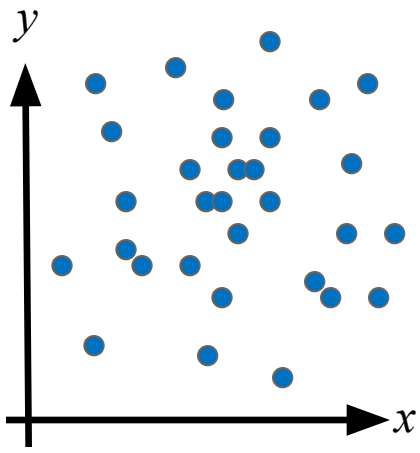
Types of Correlation

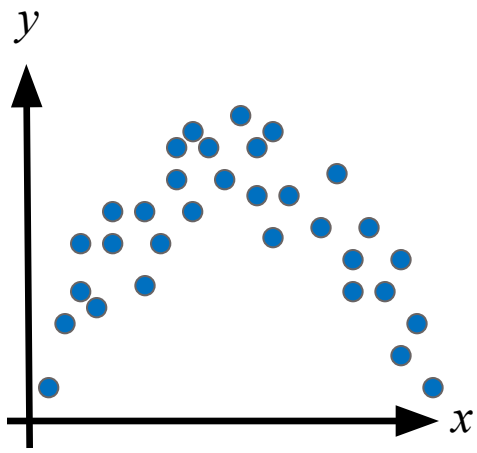


Negative Linear Correlation



Positive Linear Correlation





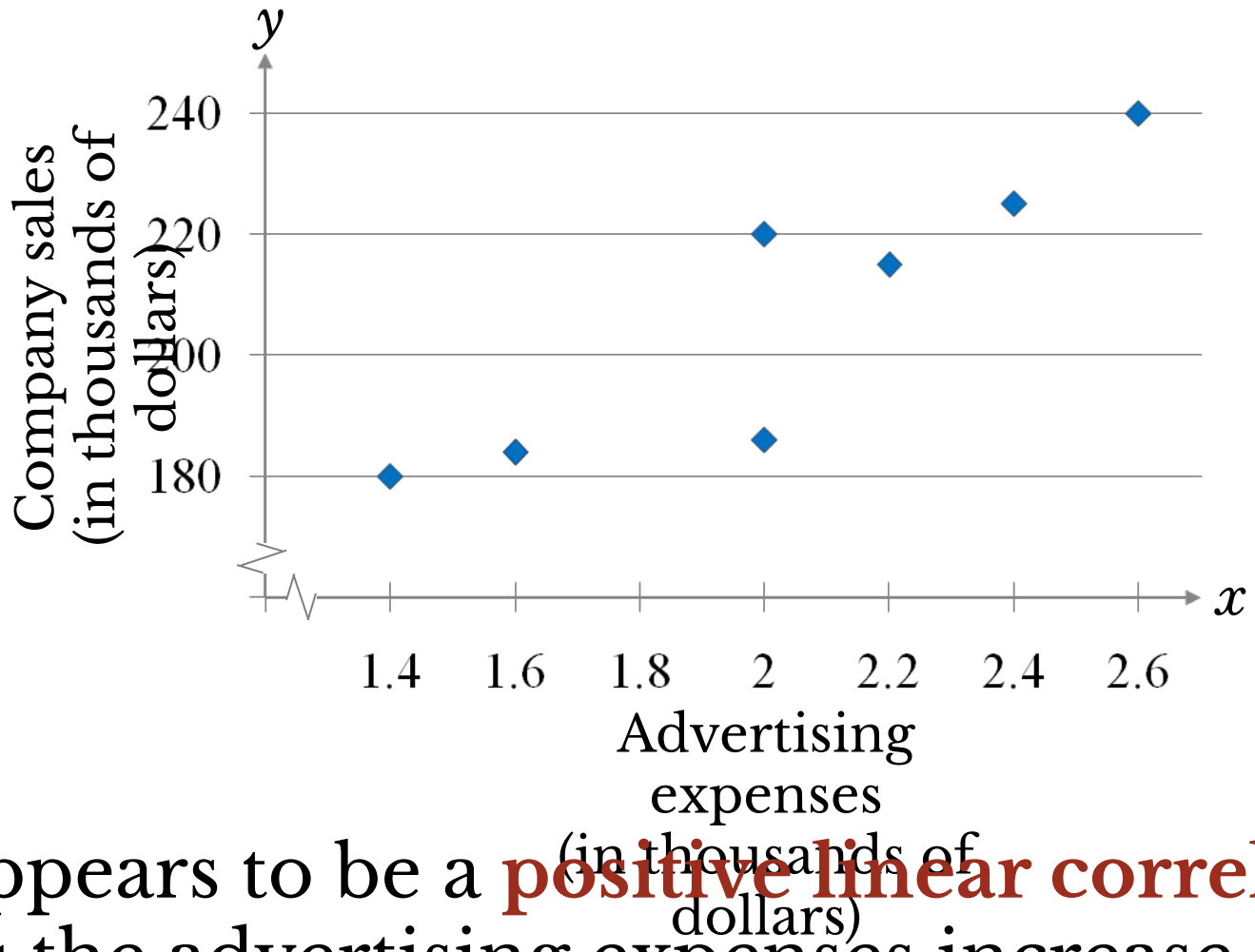
Nonlinear Correlation

Example: Constructing a Scatter Plot

A marketing manager conducted a study to determine whether there is a linear relationship between money spent on advertising and company sales. The data are shown in the table. Display the data in a scatter plot and determine whether there appears to be a positive or negative linear correlation or no linear correlation.

Advertising expenses, ($1000), x	Company sales ($1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

Solution: Constructing a Scatter Plot



Appears to be a **positive linear correlation**. As the advertising expenses increase, the sales tend to increase.

Example: Constructing a Scatter Plot Using Technology

Old Faithful, located in Yellowstone National Park, is the world's most famous geyser. The duration (in minutes) of several of Old Faithful's eruptions and the times (in minutes) until the next eruption are shown in the table. Using a TI-83/84, display the data in a scatter plot. Determine the type of correlation.

Duration x	Time, y	Duration x	Time, y
1.8	56	3.78	79
1.82	58	3.83	85
1.9	62	3.88	80
1.93	56	4.1	89
1.98	57	4.27	90
2.05	57	4.3	89
2.13	60	4.43	89
2.3	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.6	92
3.27	77	4.63	91
3.65	77		

Solution: Constructing a Scatter Plot Using Technology

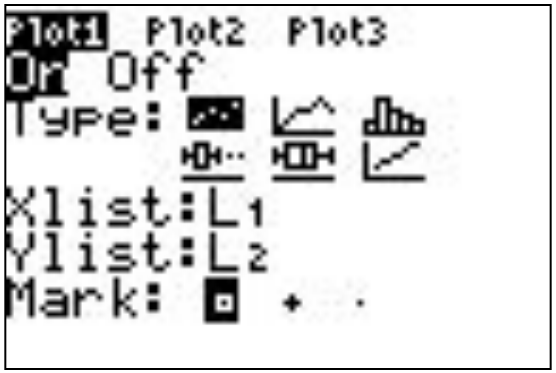
- Enter the x-values into list L1 and the y-values into list L2.
- Use *Stat Plot* to construct the scatter plot.

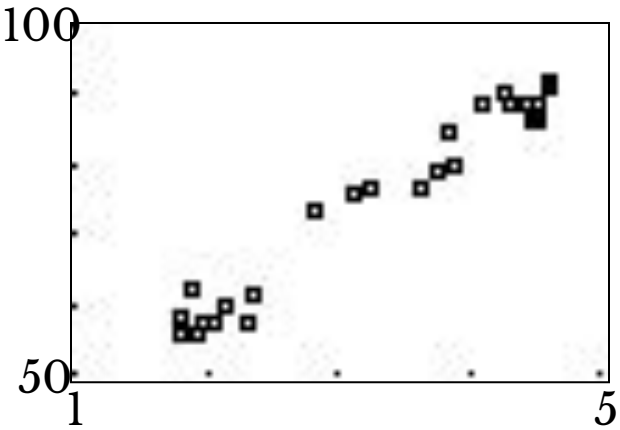
STAT > Edit...

L1	L2	L3 1
1.8	56	------
1.82	58	
1.9	62	
1.93	56	
1.98	57	
2.05	57	
2.13	60	

L1(1)=1.8

STATPLOT





From the scatter plot, it appears that the variables have a **positive linear correlation**.

Correlation Coefficient

Correlation coefficient

- A measure of the strength and the direction of a linear relationship between two variables.
- The symbol ***r*** represents the sample correlation coefficient.
- A formula for r is

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

n is the number of data pairs

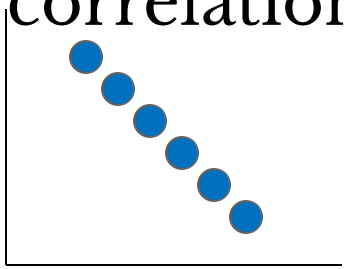
- The population correlation coefficient is represented by ρ (rho).

Correlation Coefficient

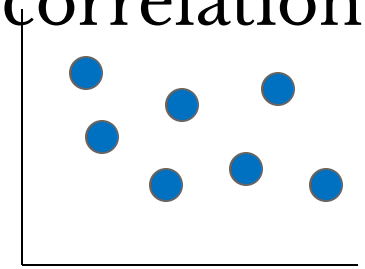
- The range of the correlation coefficient is -1 to 1.

-1 0 1

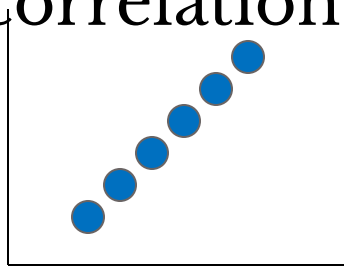
If r = -1 there is a perfect negative correlation



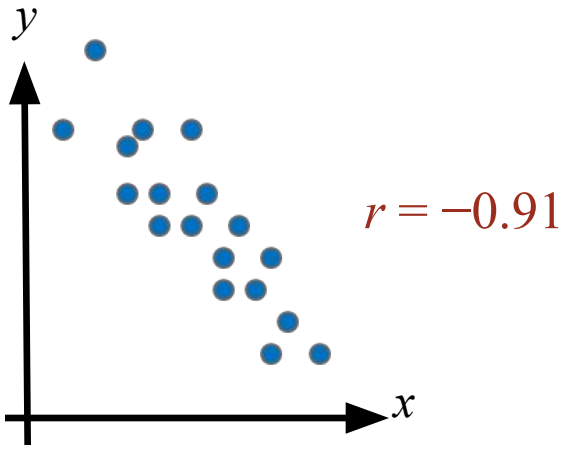
If r is close to 0 there is no linear correlation



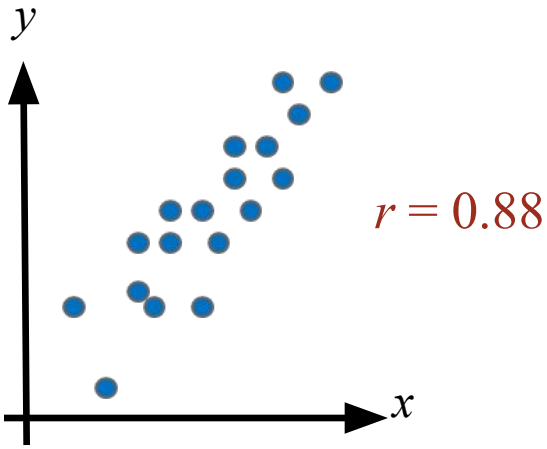
If r = 1 there is a perfect positive correlation



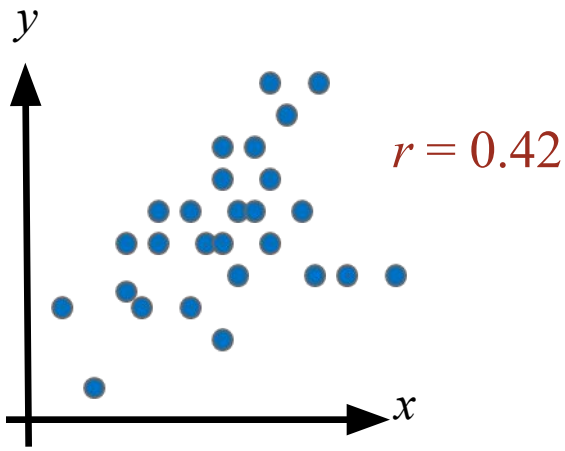
Linear Correlation



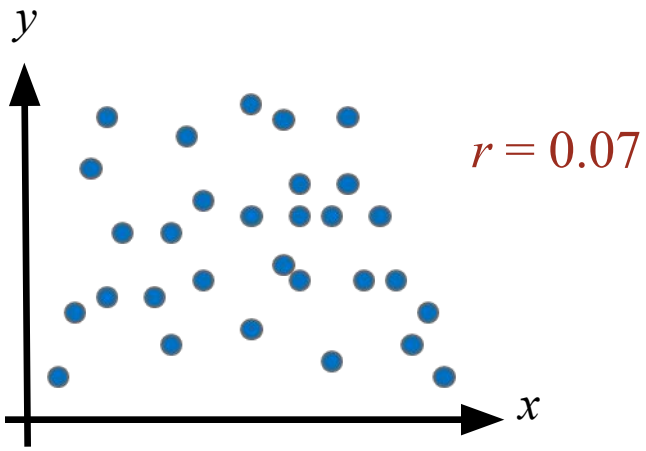
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

Calculating a Correlation Coefficient

In Words	*In Symbols*
1. Find the sum of the x-values.	$\sum x$
2. Find the sum of the y-values.	$\sum y$
3. Multiply each x-value by its corresponding y-value and find the sum.	$\sum xy$

Calculating a Correlation Coefficient

In Words	*In Symbols*
4. Square each x-value and find the sum.	$\sum x^2$
5. Square each y-value and find the sum.	$\sum y^2$
6. Use these five sums to calculate the correlation coefficient.	$r = \dfrac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$

Example: Finding the Correlation Coefficient

Calculate the correlation coefficient for the advertising expenditures and company sales data. What can you conclude?

Advertising expenses, ($1000), x	Company sales ($1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

Solution: Finding the Correlation Coefficient

x	y	xy	x^2	y^2
2.4	225	540	5.76	50,625
1.6	184	294.4	2.56	33,856
2.0	220	440	4	48,400
2.6	240	624	6.76	57,600
1.4	180	252	1.96	32,400
1.6	184	294.4	2.56	33,856
2.0	186	372	4	34,596
2.2	215	473	4.84	46,225
$\Sigma x = 15.8$	$\Sigma y = 1634$	$\Sigma xy = 3289.8$	$\Sigma x^2 = 32.44$	$\Sigma y^2 = 337{,}558$

Solution: Finding the Correlation Coefficient

$\Sigma x = 15.8 \quad \Sigma y = 1634 \quad \Sigma xy = 3289.8 \quad \Sigma x^2 = 32.44 \quad \Sigma y^2 = 337{,}558$

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

$$= \frac{8(3289.8) - (15.8)(1634)}{\sqrt{8(32.44) - 15.8^2}\sqrt{8(337{,}558) - 1634^2}}$$

$$= \frac{501.2}{\sqrt{9.88}\sqrt{30{,}508}} \approx 0.9129$$

$r \approx 0.913$ suggests a strong positive linear correlation. As the amount spent on advertising increases, the company sales also increase.

Example: Using Technology to Find a Correlation Coefficient

Use a technology tool to calculate the correlation coefficient for the Old Faithful data. What can you conclude?

Duration x	Time, y	Duration x	Time, y
1.8	56	3.78	79
1.82	58	3.83	85
1.9	62	3.88	80
1.93	56	4.1	89
1.98	57	4.27	90
2.05	57	4.3	89
2.13	60	4.43	89
2.3	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.6	92
3.27	77	4.63	91
3.65	77		

Solution: Using Technology to Find a Correlation Coefficient

MINITAB

Correlations: C1, C2

Pearson correlation of C1 and C2 = 0.979

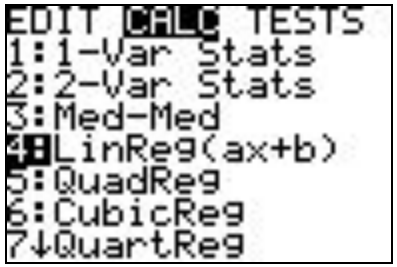
EXCEL

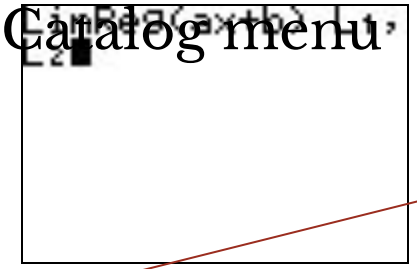
	A	B
26	=CORREL(A1:A25,B1:B25)	0.978659212930679
27		

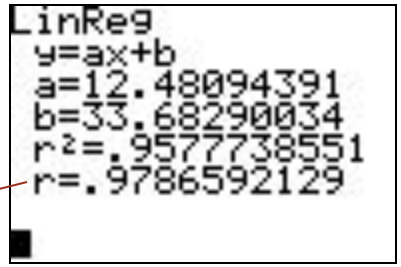
TI-83/84

STAT > Calc

To calculate ***r***, you must first enter the ***DiagnosticOn*** command found in the Catalog menu.







$r \approx 0.979$ suggests a strong positive correlation.

Using a Table to Test a Population Correlation Coefficient ρ

- Once the sample correlation coefficient r has been calculated, we need to determine whether there is enough evidence to decide that the population correlation coefficient ρ is significant at a specified level of significance.
- Use Table 11 in Appendix B.
- If $|r|$ is greater than the critical value, there is enough evidence to decide that the correlation coefficient ρ is significant.

Using a Table to Test a Population Correlation Coefficient ρ

- Determine whether ρ is significant for five pairs of data ($n = 5$) at a level of significance of $\alpha = 0.01$

Reject $H_0: \rho = 0$ if the absolute value of r is greater than the value given in the table.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

Number of pairs of data in sample

level of significance

- If $|r| > 0.959$, the correlation is significant. Otherwise, there is not enough evidence to conclude that the correlation is significant.

Using a Table to Test a Population Correlation Coefficient ρ

In Words	*In Symbols*
1. Determine the number of pairs of data in the sample.	Determine n.
2. Specify the level of significance.	Identify α.
3. Find the critical value.	Use Table 11 in Appendix B.

Using a Table to Test a Population Correlation Coefficient ρ

In Words	*In Symbols*
4. Decide if the correlation is significant.	If $\lvert r \rvert >$ critical value, the correlation is significant. Otherwise, there is not enough evidence to support that the correlation is significant.
5. Interpret the decision in the context of the original claim.	

Example: Using a Table to Test a Population Correlation Coefficient ρ

Using the Old Faithful data, you used 25 pairs of data to find
$r \approx 0.979$. Is the correlation coefficient significant? Use $\alpha = 0.05$.

Duration x	Time, y	Duration x	Time, y
1.8	56	3.78	79
1.82	58	3.83	85
1.9	62	3.88	80
1.93	56	4.1	89
1.98	57	4.27	90
2.05	57	4.3	89
2.13	60	4.43	89
2.3	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.6	92
3.27	77	4.63	91
3.65	77		

Solution: Using a Table to Test a Population Correlation Coefficient ρ

- $n = 25, \quad \alpha = 0.05$
- $|r| \approx 0.979 > 0.396$
- There is enough evidence at the 5% level of significance to conclude that there is a significant linear correlation between the duration of Old Faithful's eruptions and the time between eruptions.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
20	0.444	0.561
21	0.433	0.549
22	0.423	0.537
23	0.413	0.526
24	0.404	0.515
25	0.396	0.505
26	0.388	0.496
27	0.381	0.487
28	0.374	0.479

Hypothesis Testing for a Population Correlation Coefficient ρ

- A hypothesis test can also be used to determine whether the sample correlation coefficient r provides enough evidence to conclude that the population correlation coefficient ρ is significant at a specified level of significance.
- A hypothesis test can be one-tailed or two-tailed.

Hypothesis Testing for a Population Correlation Coefficient ρ

- Left-tailed test

  $H_0: \rho \geq 0$ (no significant negative correlation)
  $H_a: \rho < 0$ (significant negative correlation)

- Right-tailed test

  $H_0: \rho \leq 0$ (no significant positive correlation)
  $H_a: \rho > 0$ (significant positive correlation)

- Two-tailed test

  $H_0: \rho = 0$ (no significant correlation)
  $H_a: \rho \neq 0$ (significant correlation)

The *t*-Test for the Correlation Coefficient

- Can be used to test whether the correlation between two variables is significant.
- The **test statistic** is r
- The **standardized test statistic**

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

follows a *t*-distribution with d.f. = $n - 2$.

- In this text, only two-tailed hypothesis tests for ρ are considered.

Using the *t*-Test for ρ

In Words	*In Symbols*
1. State the null and alternative hypothesis.	State H_0 and H_a.
2. Specify the level of significance.	Identify α.
3. Identify the degrees of freedom.	d.f. = $n - 2$.
4. Determine the critical value(s) and rejection region(s).	Use Table 5 in Appendix B.

Using the *t*-Test for ρ

In Words	*In Symbols*
5. Find the standardized test statistic.	$t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$
6. Make a decision to reject or fail to reject the null hypothesis.	If t is in the rejection region, reject H_0. Otherwise fail to reject H_0.
7. Interpret the decision in the context of the original claim.	

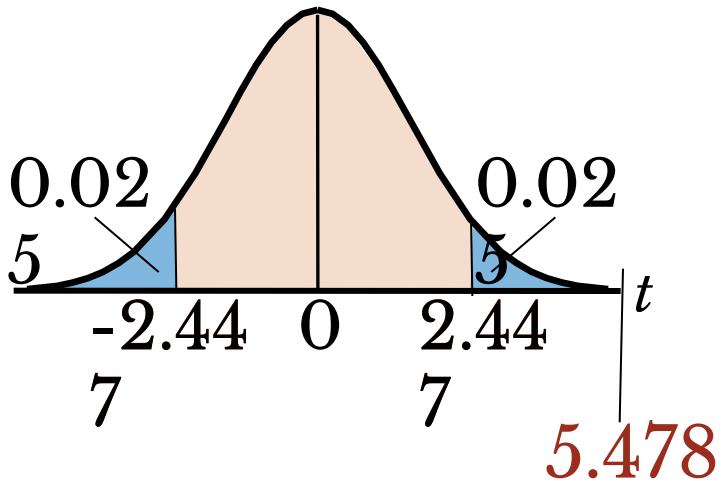
Example: *t*-Test for a Correlation Coefficient

Previously you calculated $r \approx 0.9129$. Test the significance of this correlation coefficient. Use $\alpha = 0.05$.

Advertising expenses, ($1000), x	Company sales ($1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

Solution: *t*-Test for a Correlation Coefficient

- H_0: $\rho = 0$
- H_a: $\rho \neq 0$
- $\alpha =$
- d.f. = 8 – 2 =
- Rejection Region:



- **Test Statistic:**

$$t = \frac{0.9129}{\sqrt{\frac{1-(0.9129)^2}{8-2}}} \approx 5.478$$

- **Decision** Reject H_0

At the 5% level of significance, there is enough evidence to conclude that there is a significant linear correlation between advertising expenses and

Correlation and Causation

- The fact that two variables are strongly correlated does not in itself imply a cause-and-effect relationship between the variables.
- If there is a significant correlation between two variables, you should consider the following possibilities.
  1. Is there a direct cause-and-effect relationship between the variables?
     - Does x cause y?

Correlation and Causation

2. Is there a reverse cause-and-effect relationship between the variables?
   - Does y cause x?
3. Is it possible that the relationship between the variables can be caused by a third variable or by a combination of several other variables?
4. Is it possible that the relationship between two variables may be a coincidence?

Section 9.1 Summary

- Introduced linear correlation, independent and dependent variables and the types of correlation
- Found a correlation coefficient
- Tested a population correlation coefficient ρ using a table
- Performed a hypothesis test for a population correlation coefficient ρ
- Distinguished between correlation and causation