MTH410

# Quantitative Business Analysis

## Module 8: One-Way ANOVA and Simple Linear Regression

For Module 8, we will study hypotheses involving the equality of three or more means. This will involve ANOVA, Analysis of Variance. In addition, we will incorporate data to develop equations showing how data sets are related. This process is called regression analysis, and it reflects how a dependent variable is related to an independent variable. In other words, it can be used to analyze how a change in one variable impacts the other variable, such as an increase in marketing budget increasing sales.

In addition, you will complete the final exam. The final exam is comprehensive. Be sure to prepare for the final exam ahead of time.

**Learning Outcomes**

1. Understand and apply one-way analysis of variance, ANOVA.
2. Determine statistical significance of the correlation coefficient.
3. Create a scatter diagram and regression equation using provided data.
4. Describe and explain the components of the simple linear regression model.

## For Your Success & Readings

As we conclude this course, reflect on all that you now know about statistical analysis. Are you surprised about how relevant your knowledge is to what you encounter every day? Can you think of ways that you can use what you have learned to evaluate information quantitatively and to draw inferences that can assist you in your decision making? Hopefully, statistical analysis is no longer something you **must** do, but is now something you *want* to do to enhance your evaluative skills in all aspects of your professional and personal life.

This week’s discussion question is about correlation and regression concepts. Use the internet to find a website that shows an example or application of correlation or regression in an area of interest in your personal or professional life. Discuss how correlation or regression was used, summarize your findings, and share them. Be sure to include the independent and dependent variable. Discuss the impact/relevance of the independent variable. Be sure to support your statements with logic and argument, citing any sources referenced. Post your initial response early and check back often to continue the discussion. Be sure to respond to your peers’ and instructor's posts, as well.

Finally, as you exit this course, you may wish to keep your notes handy so that you can refer to them as you hear or read about research studies and their outcomes. Statistical analysis is prevalent in all aspects of our lives, and we hope that we have equipped you with the knowledge you need to critically analyze and evaluate the next set of statistics that comes your way.

The final exam is comprehensive. Be sure to study for the final exam early in the week.

**Required**

- Sections 12.2–12.4 and 13.1–13.4, 13.6 in *Introductory Business Statistics*

**Recommended**

- Da Silva, G. (2018). **Correlation analysis of exports of manufactured products and basic products in the store of Sao Paulo.** (https://search-proquest-com.csuglobal.idm.oclc.org/docview/2126485956?rfr_id=info%3Axri%2Fsid%3Aprimo) *Independent Journal of Management & Production, 9*(5), 640–652.
- Taylor, C. (2018a, June 29). Example of an ANOVA calculation. *ThoughtCo*. Retrieved from **https://www.thoughtco.com/example-of-an-anova-calculation-3126404** (https://www.thoughtco.com/example-of-an-anova-calculation-3126404)
- Taylor, C. (2018b, September 24). What is ANOVA? *ThoughtCo*. Retrieved from **https://www.thoughtco.com/what-is-anova-3126418** (https://www.thoughtco.com/what-is-anova-3126418)

## 1. One-Way Analysis of Variance

The data from studies conducted to generate data can be used in the statistical procedure called **analysis of variance** or **ANOVA**. ANOVA can be used to test for equality of subgroup means or three or more populations in much the same way that we have already done with two populations.

To use ANOVA analysis, you must assume the following criteria:

Criterion 1

Each population has an approximately normal distribution.

Criterion 2

The variance for each population is approximately the same for all populations.

Criterion 3

The observations are independent.

ANOVA is used when the data are divided into groups according to only one factor, such as the mean of post-graduate pay. The analysis is usually focused on answering the following questions:

a. Is there a significant mean (average) difference between the groups?
b. If there is a difference, which groups are significantly different with respect to the mean or average?

In other words, to determine whether the population means are equal, we must see whether or not there is variability among the sample means and variability of the data within each sample. Statistical tests are then provided to compare group means, group medians, and group standard deviations.

For example, suppose there are three methods to complete a certain task at a factory. Twenty-one experienced employees (employed for more than one year) were randomly chosen. Seven employees used Method A, seven used Method B, and seven used Method C. Suppose we want to know if there was a difference in times between the mean task times. We would use ANOVA. For this example, we would let:

$\mu_1$ = the population mean time (in seconds) to do the task using Method A
$\mu_2$ = the population mean time (in seconds) to do the task using Method B
$\mu_3$ = the population mean time (in seconds) to do the task using Method C

The hypotheses would then be written as follows:
$H_0$: $\mu_1 = \mu_2 = \mu_3$ (The population means are equal.)
$H_a$: Not all the population means are equal.
Suppose these are the times taken (in seconds) to complete the task using each of the three different methods:

| | Method A (Factor 1) | Method B (Factor 2) | Method C (Factor 3) |
|---|---|---|---|
| | 30 | 28 | 29 |
| | 35 | 33 | 31 |
| | 40 | 38 | 35 |
| | 38 | 35 | 39 |
| | 29 | 29 | 28 |
| | 33 | 33 | 33 |
| | 35 | 37 | 32 |
| Sample Mean, | = 34.2857 | = 33.2857 | = 32.4286 |
| Sample Variance, | = 15.9048 | = 14.2381 | = 13.9524 |

The sample sizes are each 7:

$k$ = the number of groups<br>
$n$ = the total number of all the samples combined (total sample size)

In the above example, $k$ = 3 groups and $n$ = 7 + 7 + 7 = 21 total number of samples.

The groups are often called "**treatments**" or "**factors**." For example, the "treatments" above are the types of methods. The groups are called "treatments" because one of the original applications of ANOVA was in applying agricultural treatments to plants.

One uses the F distribution to compare the variation between the groups, *SS*(Factor), relative to the variation within each group, *SS*(Within). In the example above, the variation between groups would be the variation obtained by comparing the three types of methods.

To do ANOVA, one summarizes the results in an **ANOVA Table**:

| **Source of Variation** | **Sum of Squares (*SS*)** | **Degrees of Freedom (*df*)** | **Mean Sum of Squares Square (*MS*)** | **Test Statistic, *F*** | ***p*-value** |
|---|---|---|---|---|---|
| Factor (Between) | | | | | |
| Error (Within) | | | | | |
| Total | SS(Total) | | | | |

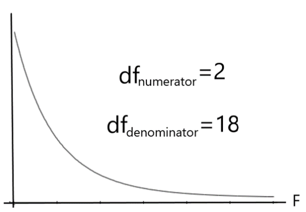
mean of the kth group of factor

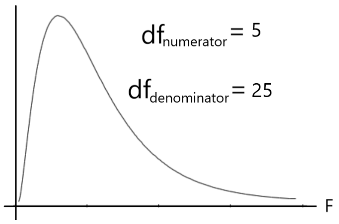
s = SS(Factor) = = a sum of weighted variances

Grand Mean = mean of all the group of factor means = =

**The hypothesis test is always right-tailed.** If the mean sum of variances between samples is large enough, relative to the mean sum of variances within samples, then reject $H_0$.

The **F-Distribution** depends on two degrees of freedom:





Computing the above ANOVA calculations is time consuming. It is a good idea to understand the idea behind ANOVA. Keep in mind that if the $p$-value is smaller than the level of significance, then one rejects $H_0$. Because of the time-consuming calculations, a recommendation is to use technology. For example, we will use Excel to solve the following problem.

EXAMPLE

Recall the example above on the three methods to complete a task:

| | Method A (Factor 1) | Method B (Factor 2) | Method C (Factor 3) |
|---|---|---|---|
| | 30 | 28 | 29 |
| | 35 | 33 | 31 |
| | 40 | 38 | 35 |
| | 38 | 35 | 39 |
| | 29 | 29 | 28 |
| | 33 | 33 | 33 |
| | 35 | 37 | 32 |
| Sample Mean, | = 34.2857 | = 33.2857 | = 32.4286 |
| Sample Std. Dev., *s* | 3.9881 | 3.7733 | 3.7353 |
| Sample Variance, | = 15.9048 | = 14.2381 | = 13.9524 |

Test the hypotheses at the .05 level of significance:

$H_0$: $\mu_1 = \mu_2 = \mu_3$ (The population mean times are equal.)

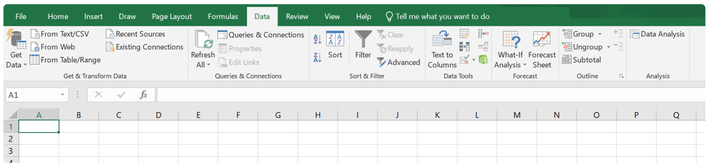
$H_a$: Not all the population means are equal.

*Click “Solution” to check your thinking.*

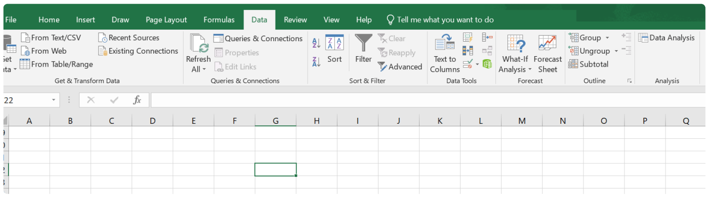
Solution

In Excel, you can click on “**File**”, then on “**Options**.” Then scroll down to select “**Add-ins**” and to the right (under “**Inactive Application Add-ins”** if you do not have the add-in) select "**Analysis ToolPak**" and click on "**Go.**" If the add-in is not visible, you might have to click “**Browse**” to find it. This will then open a new window. Check the box next to “**Analysis ToolPak**" and press “**OK.**”

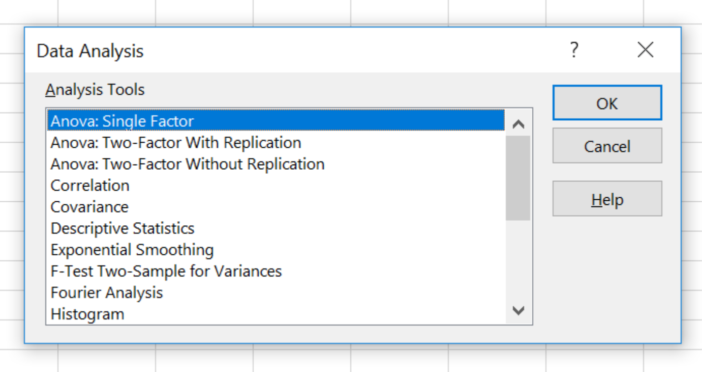
Then press the “**Data**” tab at the top of Excel:



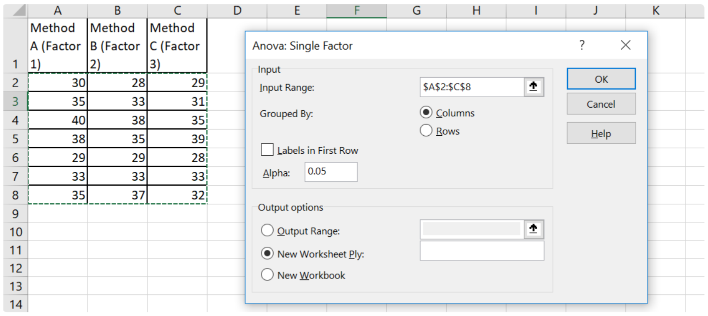
Then to the right, click on **Data Analysis**:



Select “**ANOVA: Single-Factor”** and press **OK**:



Then select the data that you will use in ANOVA. This will place the cell labels in Input Range.



| | A | B | C |
|---|---|---|---|
| 1 | Method A (Factor 1) | Method B (Factor 2) | Method C (Factor 3) |
| 2 | 30 | 28 | 29 |
| 3 | 35 | 33 | 31 |
| 4 | 40 | 38 | 35 |
| 5 | 38 | 35 | 39 |
| 6 | 29 | 29 | 28 |
| 7 | 33 | 33 | 33 |
| 8 | 35 | 37 | 32 |

Anova: Single Factor

Input

Input Range: $A$2:$C$8

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

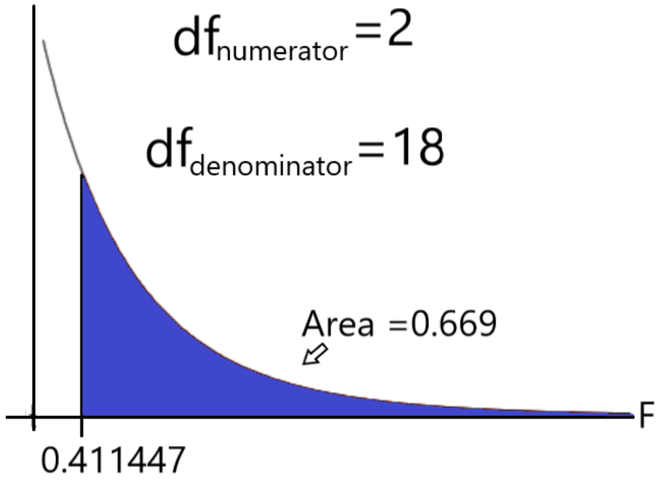
Then select the level of significance, Alpha=0.05. Press “OK” and the ANOVA Excel table will be generated:

| | | | | | | |
|---|---|---|---|---|---|---|
| Anova: Single Factor | | | | | | |
| SUMMARY | | | | | | |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* | | |
| Column 1 | 7 | 240 | 34.28571 | 15.90476 | | |
| Column 2 | 7 | 233 | 33.28571 | 14.2381 | | |
| Column 3 | 7 | 227 | 32.42857 | 13.95238 | | |
| | | | | | | |
| ANOVA | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *p-value* | *F crit* |
| Between Groups | 12.09524 | 2 | 6.047619 | 0.411447 | 0.668767 | 3.554557 |
| Within Groups | 264.5714 | 18 | 14.69841 | | | |
| | | | | | | |
| Total | 276.6667 | 20 | | | | |

Note that the test statistic is the value of the F-distribution:

*p*-**Value Method**:

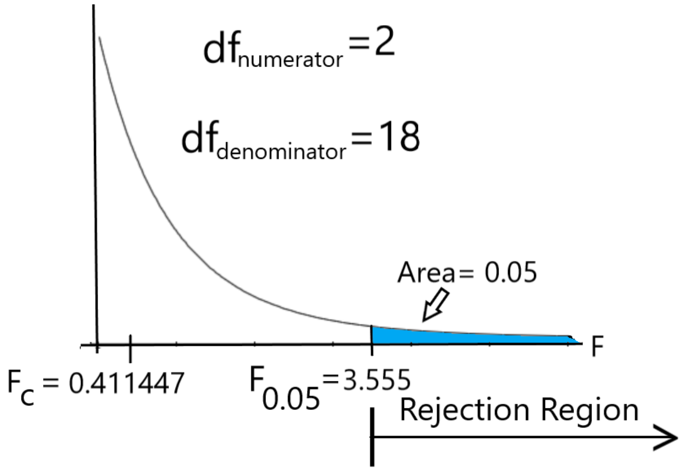
The *p*-value is *P(F > 0.411)= 0.669:*



Since the *p*-value is not less than the level of significance (0.669 > 0.05), then fail to reject $H_0$. We cannot conclude that some of the mean task times differ among the different methods.

**Critical Value/Rejection Region Method**

ANOVA is always a right-tailed test. Thus, since the F-statistic is less than the critical value (0.411 < 3.555), then we would fail to reject $H_0$. We cannot conclude that some of the mean task times differ among the different methods.



The critical value can also be found by Appendix pages 595–606, Tables A1–10. Look up the degrees of freedom:

On page 602, we find that that the critical value is, under the *P* column:

If the Excel Data Analysis ToolPak in Excel is unavailable, you may use an **online calculator** (https://www.danielsoper.com/statcalc/calculator.aspx?id=43).

For example, input the sample sizes of seven and the above sample means and standard deviations.
From this information, we get the online ANOVA Table:

| **Source of Variation** | **Sum of Squares** | **Degrees of Freedom** | **Mean Square** | **F** | ***p*-value** |
|---|---|---|---|---|---|
| Factor<br>(Between) | 12.095 | 2 | 6.048 | 0.411 | 0.669 |
| Error<br>(Within) | 264.571 | 18 | 14.698 | | |
| Total | 276.667 | 20 | | | |

Since the *p*-value is not less than the level of significance (0.669 > 0.05), we would fail to reject $H_0$. We cannot conclude that some of the mean task times differ among the different methods.

## 1.1. More on ANOVA, *p*-value Method

ANOVA is a method that one can use provided that the samples are independent, and normally distributed with equal standard deviations. In the following example, we will assume such is the case.

EXAMPLE

Suppose we have four sets of data regarding recent post-graduate pay for an M.S. in Leadership, an M.S. in Management, an M.S. in Organizational Leadership, and an M.S. in Accounting, and we want to know if there was a difference in pay between them, we would use ANOVA.

$\mu_1$ = the population mean post-graduate pay of graduates with an M.S. in Leadership
$\mu_2$ = the population mean post-graduate pay of graduates with an M.S. in Management
$\mu_3$ = the population mean post-graduate pay of graduates with an M.S. in Organizational Leadership
$\mu_4$ = the population mean post-graduate pay of graduates with an M.S. in Accounting

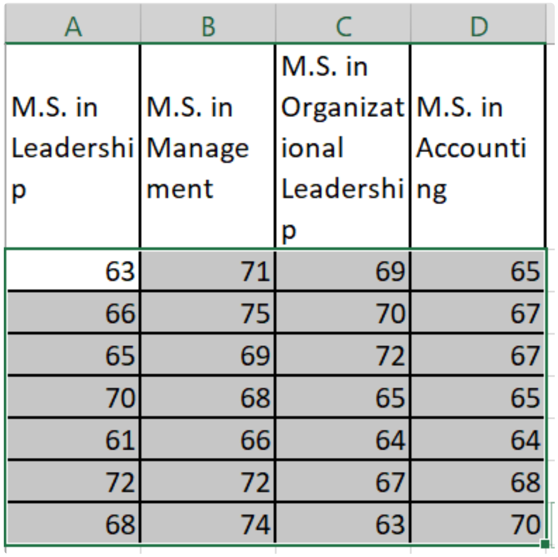
Test the following hypotheses at the 0.1 level of significance:

$H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (The population mean salaries are equal.)
$H_a$: Not all the population means are equal.

Here are the salaries in thousands of dollars per year:

| M.S. in Leadership | M.S. in Management | M.S. in Organizational Leadership | M.S. in Accounting |
|---|---|---|---|
| 63 | 71 | 69 | 65 |
| 66 | 75 | 70 | 67 |
| 65 | 69 | 72 | 67 |
| 70 | 68 | 65 | 65 |
| 61 | 66 | 64 | 64 |
| 72 | 72 | 67 | 68 |
| 68 | 74 | 63 | 70 |

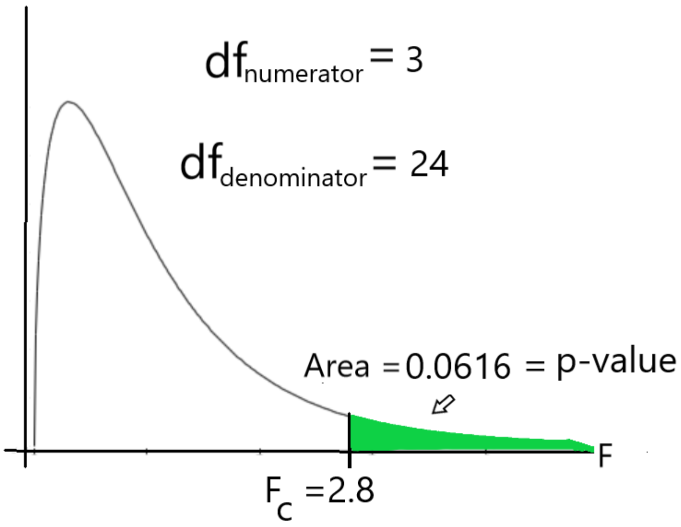
Click on the Data tab in Excel and then on Data Analysis. Then select ANOVA: Single Factor and press OK. Select the data in cells Excel:



| A | B | C | D |
|---|---|---|---|
| M.S. in Leadership | M.S. in Management | M.S. in Organizational Leadership | M.S. in Accounting |
| 63 | 71 | 69 | 65 |
| 66 | 75 | 70 | 67 |
| 65 | 69 | 72 | 67 |
| 70 | 68 | 65 | 65 |
| 61 | 66 | 64 | 64 |
| 72 | 72 | 67 | 68 |
| 68 | 74 | 63 | 70 |

Then select the level of significance, Alpha=0.1. The following Excel ANOVA table is generated:

| | | | | | | |
|---|---|---|---|---|---|---|
| ANOVA: Single Factor | | | | | | |
| SUMMARY | | | | | | |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* | | |
| Column 1 | 7 | 465 | 66.42857 | 14.95238 | | |
| Column 2 | 7 | 495 | 70.71429 | 10.57143 | | |
| Column 3 | 7 | 470 | 67.14286 | 11.14286 | | |
| Column 4 | 7 | 466 | 66.57143 | 4.285714 | | |
| | | | | | | |
| ANOVA | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *p-value* | *F crit* |
| Between Groups | 86 | 3 | 28.66667 | 2.8 | 0.061643 | 2.32739 |
| Within Groups | 245.7143 | 24 | 10.2381 | | | |
| | | | | | | |
| Total | 331.7143 | 27 | | | | |



The *p*-value is *P(F > 2.8) = 0.0616*. Since the *p*-value is less than the level of significance (0.0616 < 0.10), we would reject the null hypothesis. At the 0.10 level of significance, there is strong enough evidence to conclude that the mean wages are not all equal.

Note that the null hypothesis would not be rejected if the level of significance had been less than 0.0616. For example, if the level of significance had been 0.05, then we would fail to reject the null hypothesis.

## 1.2. More on ANOVA, Critical Value/Rejection Region Method

Recall that ANOVA tests whether three or more means are equal. To apply ANOVA, we will assume that the samples are independent, and normally distributed with equal standard deviations. The following is an example obtained from Secondary Page 1.1. This is the same as the page 1.2 example, except that we are using the critical value/rejection region method.

EXAMPLE

Suppose we have four sets of data regarding recent post-graduate pay for an M.S. in Leadership, an M.S. in Management, an M.S. in Organizational Leadership, and an M.S. in Accounting, and we want to know if there was a difference in pay between them, we would use ANOVA.

$\mu_1$ = the population mean post-graduate pay of graduates with an M.S. in Leadership
$\mu_2$ = the population mean post-graduate pay of graduates with an M.S. in Management
$\mu_3$ = the population mean post-graduate pay of graduates with an M.S. in Organizational Leadership
$\mu_4$ = the population mean post-graduate pay of graduates with an M.S. in Accounting

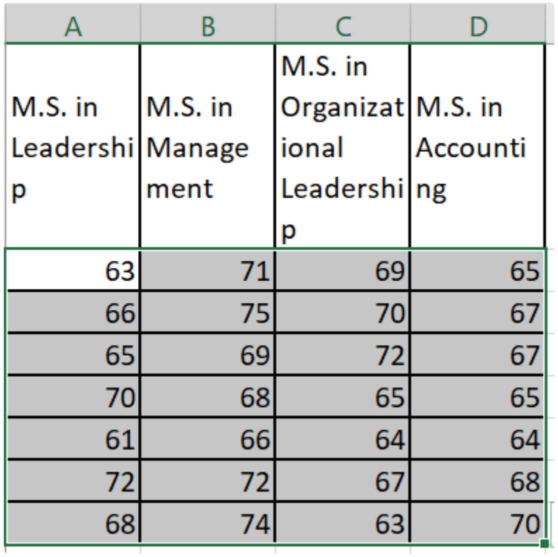
Test the following hypotheses at the 0.1 level of significance:

$H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (The population mean salaries are equal.)
$H_a$: Not all the population means are equal.

Here are the salaries in thousands of dollars per year:

| M.S. in Leadership | M.S. in Management | M.S. in Organizational Leadership | M.S. in Accounting |
|---|---|---|---|
| 63 | 71 | 69 | 65 |
| 66 | 75 | 70 | 67 |
| 65 | 69 | 72 | 67 |
| 70 | 68 | 65 | 65 |
| 61 | 66 | 64 | 64 |
| 72 | 72 | 67 | 68 |
| 68 | 74 | 63 | 70 |

Click on the Data tab in Excel and then on Data Analysis. Then select ANOVA: Single Factor and press OK. Select the data in cells Excel:



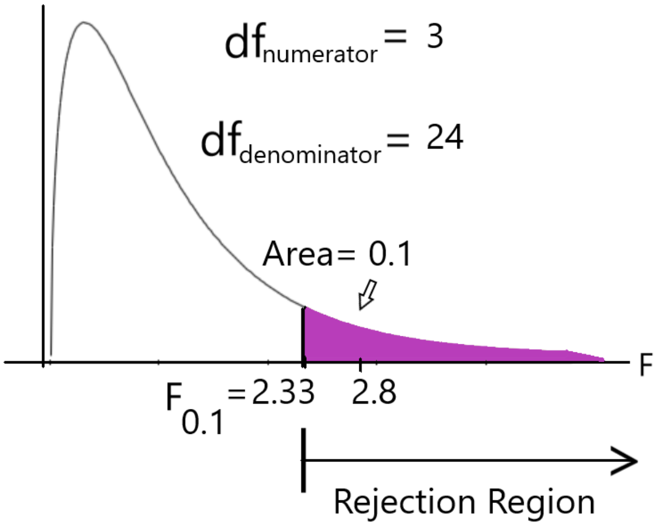
| A | B | C | D |
|---|---|---|---|
| M.S. in Leadership | M.S. in Management | M.S. in Organizational Leadership | M.S. in Accounting |
| 63 | 71 | 69 | 65 |
| 66 | 75 | 70 | 67 |
| 65 | 69 | 72 | 67 |
| 70 | 68 | 65 | 65 |
| 61 | 66 | 64 | 64 |
| 72 | 72 | 67 | 68 |
| 68 | 74 | 63 | 70 |

Then select the level of significance, Alpha=0.1. The following Excel ANOVA table is generated:

| | | | | | | |
|---|---|---|---|---|---|---|
| ANOVA: Single Factor | | | | | | |
| SUMMARY | | | | | | |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* | | |
| Column 1 | 7 | 465 | 66.42857 | 14.95238 | | |
| Column 2 | 7 | 495 | 70.71429 | 10.57143 | | |
| Column 3 | 7 | 470 | 67.14286 | 11.14286 | | |
| Column 4 | 7 | 466 | 66.57143 | 4.285714 | | |
| | | | | | | |
| ANOVA | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *p-value* | *F crit* |
| Between Groups | 86 | 3 | 28.66667 | 2.8 | 0.061643 | 2.32739 |
| Within Groups | 245.7143 | 24 | 10.2381 | | | |
| | | | | | | |
| Total | 331.7143 | 27 | | | | |

The critical value can also be found by Appendix pages 595-660, Tables A1-A10. Look up the degrees of freedom:

On page 602, we find that the critical value is, under the *P* column:



ANOVA tests are always right-tailed. Thus, since the test statistic is greater than the critical value (2.8 > 2.33), we would reject the null hypothesis. At the 0.10 level of significance, there is strong enough evidence to conclude that the mean wages are not all equal.

Note that the null hypothesis would not be rejected if the level of significance had been less than 0.0616. For example, if the level of significance had been 0.05, then we would fail to reject the null hypothesis.

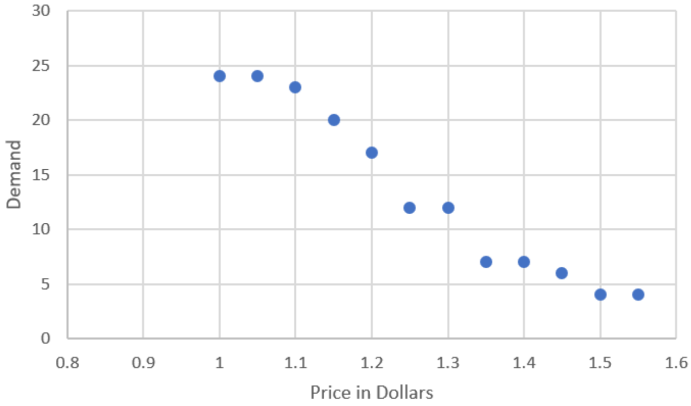
## 2. Correlation Coefficient and Test for Significance of the Correlation Coefficient

A **scatterplot** visually describes a relationship between two numerical variables. For example, suppose a manufacturer is interested in the relationship between demand in thousands of items and the price, in dollars. The following data were collected from 10 stores.

| Price in Dollars | Demand in Thousands of Items |
|---|---|
| 1.45 | 6 |
| 1.55 | 4 |
| 1.1 | 23 |
| 1.5 | 4 |
| 1.2 | 17 |
| 1.25 | 12 |
| 1.3 | 12 |
| 1.35 | 7 |
| 1.4 | 7 |
| 1 | 24 |
| 1.15 | 20 |
| 1.05 | 24 |

The **scatterplot** is found with Excel by selecting the 2 columns of data, and then clicking on the **Insert** tab. Next, we selected **Insert Scatter**. The scatter diagram is then selected.

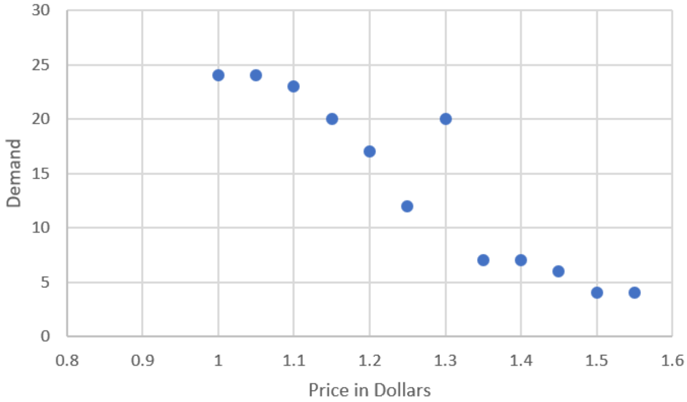
The scatterplot derived from the above data is as follows:



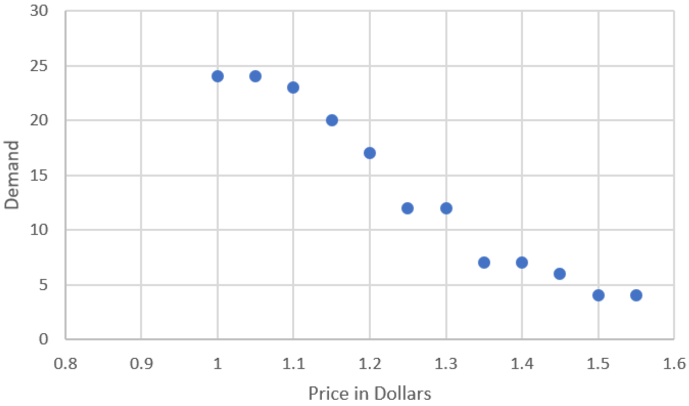
From the scatterplot, we may see trends. For example, it seems a higher price generally corresponds to less demand. The scatterplot pattern seems to be somewhat linear.

Also, a scatterplot may help identify **outliers**. Outliers might be visible from unusual $x$-values, or unusual $y$-values, or both, that do not fit the overall scatterplot pattern.

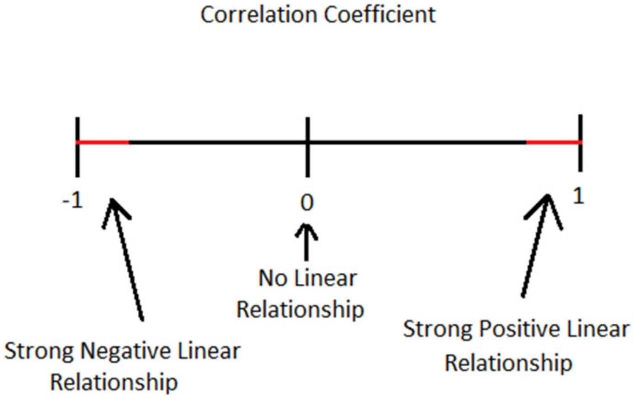
For example, the following scatterplot seems to identify the point (1.3, 20) as a potential outlier. The $y$-value appears to be unusually large.



In the following example, the leftmost point appears to be an outlier:



Recall, from Module 2, the **correlation coefficient** ($r$, where $-1 \le r \le 1$) is a number that indicates both the direction and the strength of a linear relationship between the dependent variable ($y$) and the independent variable ($x$). If $r$ is positive, then $x$ and $y$ are directly related, and if $r$ is negative, then they are inversely related. The closer the absolute value of $r$ is to 1, the stronger the linear relationship between $x$ and $y$. If $r = -1$ or $r = +1$, then the best-fit linear equation graph passes through all the data values.

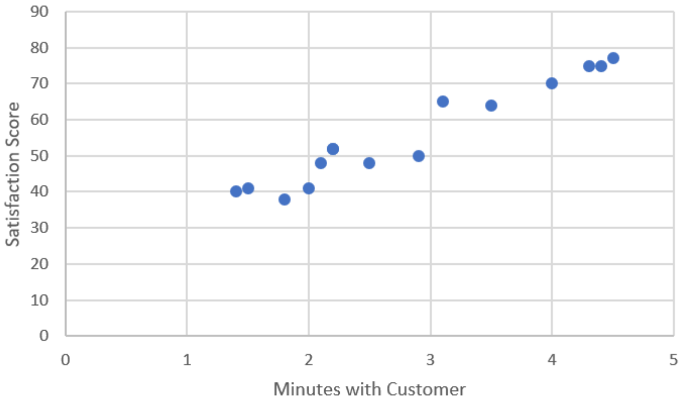

Correlation Coefficient
-1
0
1
No Linear Relationship
Strong Negative Linear Relationship
Strong Positive Linear Relationship

See Figure 13.2 on page 554 for different scatterplots and the corresponding value of $r$.

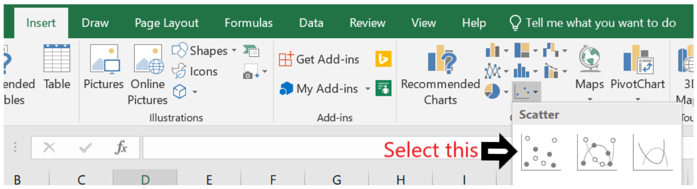
Suppose a manager recorded the number of minutes its phone operators spent on the phone with customers. Afterward, customers answered questions on how satisfied they were with service. A score of 100 represents the highest level of customer satisfaction. Fifteen calls with times and corresponding scores were recorded and shown here:

| **Minutes with a Customer, *x*** | **Satisfaction Score, *y*** |
|---|---|
| 2.9 | 50 |
| 2 | 41 |
| 1.5 | 41 |
| 4 | 70 |
| 2.1 | 48 |
| 2.2 | 52 |
| 4.3 | 75 |
| 3.1 | 65 |
| 4.4 | 75 |
| 3.5 | 64 |
| 1.8 | 38 |
| 4.5 | 77 |
| 1.4 | 40 |
| 2.2 | 52 |
| 2.5 2.827 | 48 55.73 |
| | |
| | |

By using Excel, the corresponding scatterplot is shown:



To draw the scatterplot, select the two rows of data. Then press the Insert Tab, and then press Scatter. Select the chart with the dots that are not joined by any line:



By looking at the scatterplot, there appears to be a positive linear correlation. We need to compute the sample correlation coefficient, *r*.

r =

The sample correlation coefficient, *r*, can be calculated with the following formula:

Here, and are the sample standard deviations for *x* and *y*, respectively, and *n* is the sample size.

EXAMPLE

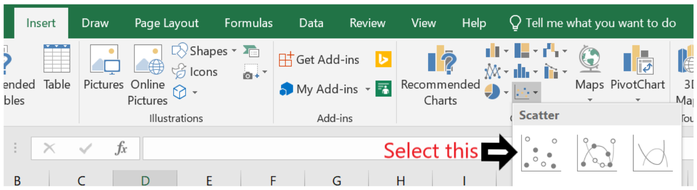
Suppose the 15 times spent with a customer on the phone and corresponding customer satisfaction score are shown in the first two columns:

| **Minutes with a Customer,** *x* | **Satisfaction Score,** *y* | | | |
|---|---|---|---|---|
| 2.9 | 50 | 0.073 | -5.73 | -0.41829 |
| 2 | 41 | -0.827 | -14.73 | 12.18171 |
| 1.5 | 41 | -1.327 | -14.73 | 19.54671 |
| 4 | 70 | 1.173 | 14.27 | 16.73871 |
| 2.1 | 48 | -0.727 | -7.73 | 5.61971 |
| 2.2 | 52 | -0.627 | -3.73 | 2.33871 |
| 4.3 | 75 | 1.473 | 19.27 | 28.38471 |
| 3.1 | 65 | 0.273 | 9.27 | 2.53071 |
| 4.4 | 75 | 1.573 | 19.27 | 30.31171 |
| 3.5 | 64 | 0.673 | 8.27 | 5.56571 |
| 1.8 | 38 | -1.027 | -17.73 | 18.20871 |
| 4.5 | 77 | 1.673 | 21.27 | 35.58471 |
| 1.4 | 40 | -1.427 | -15.73 | 22.44671 |
| 2.2 | 52 | -0.627 | -3.73 | 2.33871 |
| 2.5 2.827 | 48 55.73 | -0.327 | -7.73 | 2.52771 |
| | | | | *Sum* = 203.9067 |
| | | | | |

$r = \quad = \quad = 0.967$

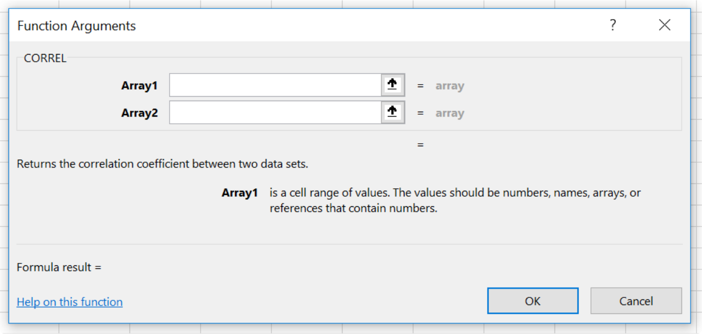
From the table above, we may compute the sample correlation coefficient.

To save time and errors, one may instead use **Excel** to find the sample correlation coefficient, *r*. Select the **Formulas** tab at the top. Then select **More Functions** and click on **Statistical**. Next, scroll down to **CORREL**.



Select this

For **Array1** and **Array2**, select the first and second data columns, respectively.



Returns the correlation coefficient between two data sets.

**Array1** is a cell range of values. The values should be numbers, names, arrays, or references that contain numbers.

### Interpreting the Correlation Coefficient and Test for Significance

Denote the population correlation coefficient by the Greek letter, $\rho$. Recall that a correlation coefficient sufficiently close to 0 implies no linear relationship between the two sets of data.

If the population correlation coefficient is different from 0, we will say that the correlation coefficient is "**significant**." Thus, we would like to test these hypotheses regarding the population correlation coefficient:

$H_0$: $\rho = 0$ (There is no significant linear relationship between the two variables.)
$H_a$: $\rho \neq 0$ (There is a significant linear relationship between the two variables.)

The above is a two-tailed hypothesis test.

There are two methods of determining whether $H_0$ should be rejected.

**Method 1: Use the t-test for a correlation coefficient:**

Use the test statistic and the t-distribution with *n-2* degrees of freedom:

*n* is the number of pairs and *r* is the sample correlation coefficient.

Since the hypothesis test is two-tailed, reject $H_0$ if the test statistic is less than the negative critical value or greater than the positive critical value.

In the above example, suppose we want to test the hypotheses at the 0.05 level of significance.

The critical values with *n-2=15-2=13* degrees of freedom are obtained from Appendix page 608, Table A12:

Since the test statistic is greater than the critical value (13.685 > 2.160), we would reject the null hypothesis. At the 0.05 level of significance, there is a significant linear relationship between the hours studying and the score. Since the correlation coefficient is positive, we can conclude there is a significant positive linear relationship between the minutes with a customer on the phone and the customer satisfaction score.

**Method 2:** Use a **table of critical values for the correlation coefficient**:

The tables use the above formula and the level of significance. The critical values listed on those tables are not the same as the critical values for the t-distribution. Rather, they are specifically calculated values for the sample correlation coefficient, *r*. Many such tables only list values for a 0.05 level of significance. For example, see this **Table of Critical Values for the Correlation Coefficient** (https://www.statisticssolutions.com/table-of-critical-values-pearson-correlation/).

If the absolute value of the sample correlation coefficient, $|r|$, is less than the correlation critical value, then fail to reject the null hypothesis. Otherwise, reject the null hypothesis.

In our example above, suppose we want to test the hypotheses at the 0.05 level of significance. The critical value with *n-2=15-2=13* degrees of freedom is obtained by scrolling down to 13 degrees of freedom and to the 0.05 level of significance. The correlation critical value (CV) is

Since $|r|= 0.997$ is greater than 0.514, then reject the null hypothesis. At the 0.05 level of significance, there is a significant linear relationship between the minutes with a customer on the phone and the customer satisfaction score. Since the correlation coefficient is positive, we would conclude there is a significant positive linear relationship between the minutes with a customer on the phone and the customer satisfaction score.

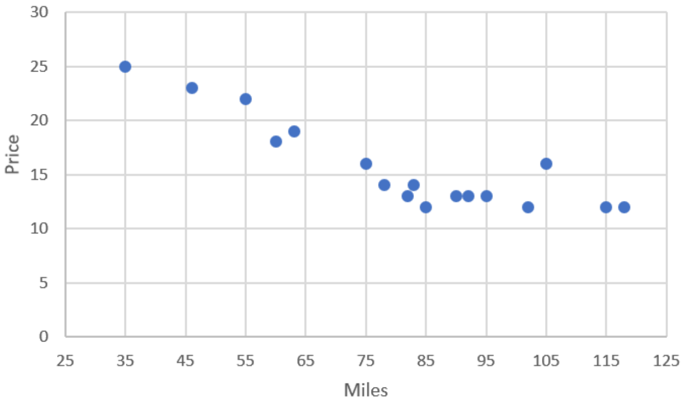
## 2.1. Constructing a Scatterplot and Interpreting the Correlation Coefficient

The following is an example of constructing and interpreting a scatterplot. We will also compute and determine the significance of the correlation coefficient.

A car dealership is interested in the cost of a 3-year old Model A vehicle. The dealership is interested in the relationship between Model A car mileage (in thousands) and cost. The following data were obtained on 17 randomly chosen Model A cars:

| Model A Miles<br>(in thousands) | Model A Price<br>(in thousands) |
|---|---|
| 82 | 13 |
| 90 | 13 |
| 92 | 13 |
| 115 | 12 |
| 63 | 19 |
| 118 | 12 |
| 78 | 14 |
| 83 | 14 |
| 35 | 25 |
| 102 | 12 |
| 105 | 16 |
| 55 | 22 |
| 95 | 13 |
| 85 | 12 |
| 46 | 23 |
| 60 | 18 |
| 75 | 16 |

The Excel generated scatterplot is:



The relationship here appears to indicate that cars with more miles have a lower price. The car with 105 thousand miles and price of $16 thousand appears to be an outlier. Using Excel, the correlation coefficient is computed to be $r$ = -0.882.

We would like to determine at the 0.05 level of significance if there is a significant linear relationship between Model A miles and price.

$H_0$: $\rho = 0$ (There is no significant linear relationship between Model A miles and price.)
$H_a$: $\rho \neq 0$ (There is a significant linear relationship between Model A miles and price.)

We would like to use a **table of critical values for the correlation coefficient** (https://www.statisticssolutions.com/table-of-critical-values-pearson-correlation/).

We want to test the hypotheses at the 0.05 level of significance. The critical value with *n-2=17-2=15* degrees of freedom is obtained by scrolling down to 15 degrees of freedom and to the 0.05 level of significance. The correlation critical value is

Since |*r*|=|*-0.882*|=*0.882* is greater than 0.482, we would reject the null hypothesis. At the 0.05 level of significance, there is a significant linear relationship between Model A miles and price. Since the correlation coefficient is negative, we can conclude there is a significant negative linear relationship between Model A miles and price.

## 2.2. More on Interpreting the Correlation Coefficient and Test for Significance

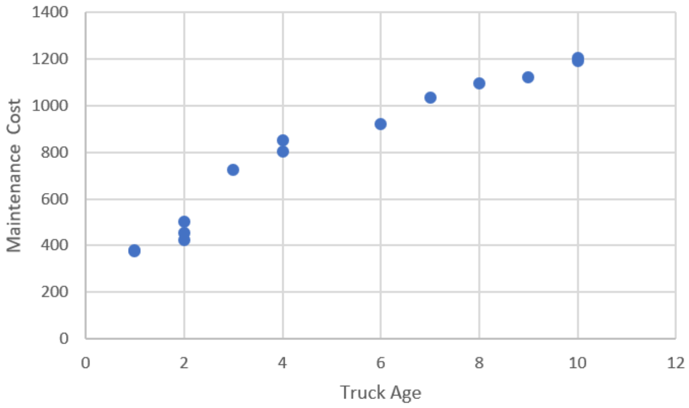
We would like to determine if there is a significant linear relationship between two variables.

EXAMPLE

A trucking company wants to determine if there is a linear relationship between the age of its trucks and the yearly maintenance costs. Suppose the ages and maintenance costs for last year of 14 of its trucks are obtained.

| Truck Age in Years | Maintenance Cost |
|---|---|
| 3 | 723 |
| 1 | 374 |
| 7 | 1032 |
| 10 | 1205 |
| 1 | 380 |
| 4 | 805 |
| 2 | 453 |
| 9 | 1123 |
| 10 | 1189 |
| 8 | 1095 |
| 6 | 920 |
| 2 | 422 |
| 2 | 500 |
| 4 | 850 |

With Excel, the scatterplot is:



Using the CORREL feature of Excel, the sample correlation coefficient is $r = 0.969$.

We would like to determine at the 0.05 level of significance if there is a significant linear relationship between truck age and maintenance cost.

$H_0$: $\rho = 0$ (There is no significant linear relationship between truck age and maintenance cost.)
$H_a$: $\rho \neq 0$ (There is a significant linear relationship between truck age and maintenance cost.)

There are two methods of determining whether $H_0$ should be rejected.

**Method 1: Use the t-test for a correlation coefficient:**

The test statistic is

The critical value with *n-2=14-2=12* degrees of freedom is obtained from Appendix page 608, Table A12:

Since the test statistic is greater than the critical value (13.587 > 2.179), we would reject the null hypothesis. At the 0.05 level of significance, there is a significant linear relationship between truck age and maintenance cost. Since the correlation coefficient is positive, we can conclude there is a significant positive linear relationship between truck age and maintenance cost.

**Method 2:** Use a **table of critical values for the correlation coefficient** (https://www.statisticssolutions.com/table-of-critical-values-pearson-correlation/).

We want to test the hypotheses at the 0.05 level of significance. The critical value with *n-2=14-2=12* degrees of freedom is obtained by scrolling down to 12 degrees of freedom and to the 0.05 level of significance. The correlation critical value is

Since $|r|$ = 0.969 is greater than 0.532, we would reject the null hypothesis. At the 0.05 level of significance, there is a significant linear relationship between truck age and maintenance cost. Since the correlation coefficient is positive, we can conclude there is a significant positive linear relationship between truck age and maintenance cost.

## 3. Linear Regression Equation

**Linear Equations**

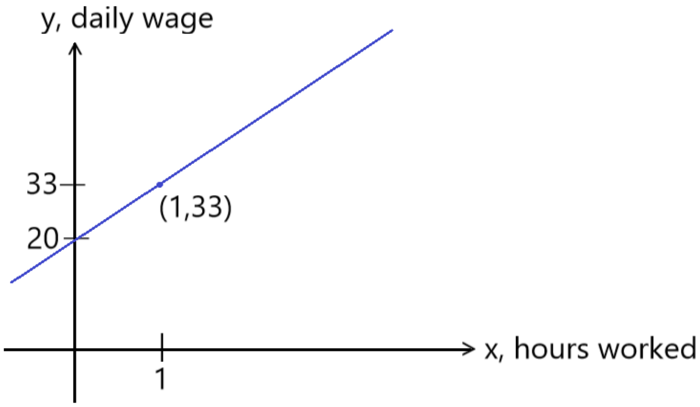
Since we will study linear regression, here is a short review of linear equations. A linear equation is of the form

The term *b* is the slope, and a represents the *y*-intercept. The graph of the above equation is a straight line. The term *x* is called the **independent variable,** and *y* is the **dependent variable**. The slope, *b*, measures the steepness of the line.

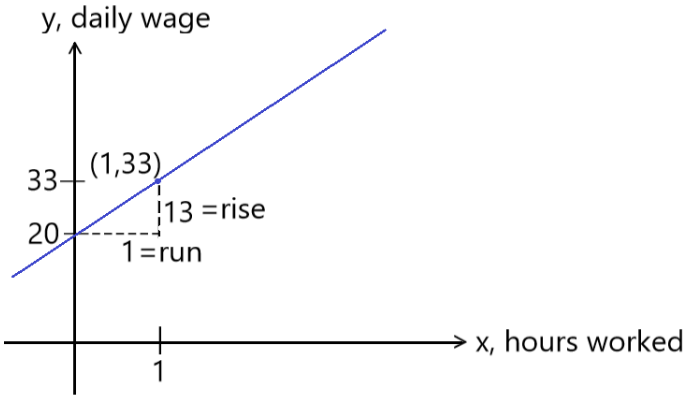
For example, suppose a worker get paid $13 per hour and is given $20 per day for meals. Let *x* represent the number of hours worked. The linear equation representing the worker's daily wages, *y*, is

Since the *y*-intercept is *a* = 20, the point (0, 20) lies on the graph. The slope is *b* = 13. To graph the equation, we need to find one additional point. For example, if *x* = 1, then .

The graph is shown here:



The slope of *b* = 13 = 13/1 means that for each hour worked, the wage increases by $13:



Since hours cannot be negative, the above equation applies only to non-negative *x*-values. A positive slope means that the line is increasing, as shown above. This means that as *x* increases, *y* increases, too. A negative slope means that the line is decreasing. A zero slope means that the line is horizontal. For graphical explanations regarding the slope, see Figure 13.5 on page 557.

**Linear Regression Equation**

If given a scatterplot, the "line of best fit" is given by the regression equation

The slope, $b_1$ of the line is

$s_y$ and $s_x$ are the sample standard deviations of *y* and *x*, respectively. *r* is the sample correlation coefficient.
The *y*-intercept, of the line is

and are the sample means of *x* and *y*, respectively.

EXAMPLE

Suppose a manager recorded the number of minutes their phone operators spent on the phone with customers. Afterward, customers answered questions on how satisfied they were with the service they received. A score of 100 represents the highest level of customer satisfaction. Fifteen calls with times and corresponding scores were recorded and are shown here:

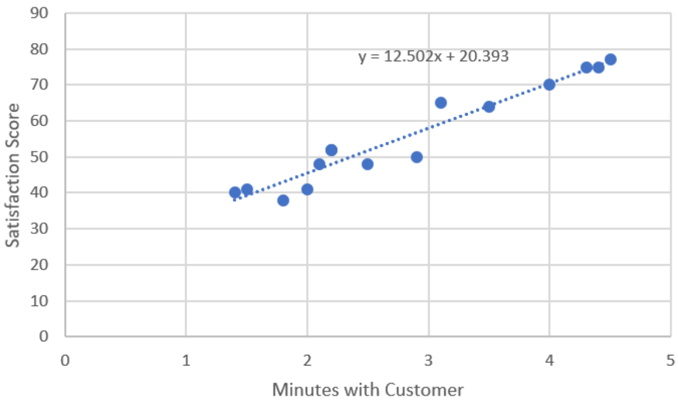
| **Minutes with a Customer,** $x$ | **Satisfaction Score,** $y$ |
|---|---|
| 2.9 | 50 |
| 2 | 41 |
| 1.5 | 41 |
| 4 | 70 |
| 2.1 | 48 |
| 2.2 | 52 |
| 4.3 | 75 |
| 3.1 | 65 |
| 4.4 | 75 |
| 3.5 | 64 |
| 1.8 | 38 |
| 4.5 | 77 |
| 1.4 | 40 |
| 2.2 | 52 |
| 2.5 | 48 |
| | |
| | |

Recall that the sample correlation coefficient is $r$ = 0.967.

Thus, the regression equation is

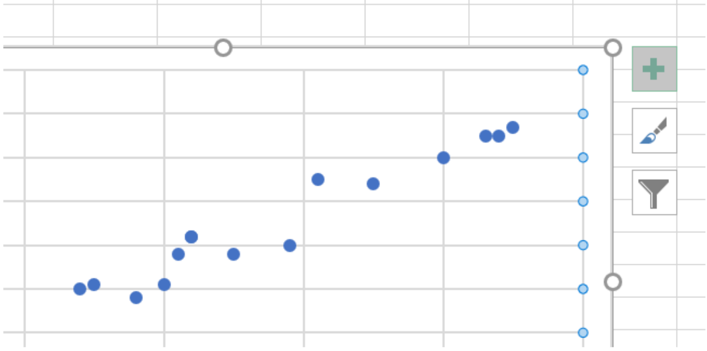
Since the slope, 12.51, is positive, we expect the graph of the line to be increasing. This means that as x increases, then y increases, too.

Alternatively, we can use Excel to draw the graph of the regression line and determine the regression equation:

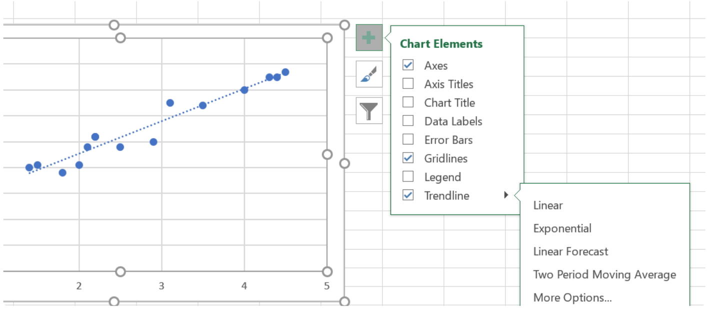


Due to roundoff error, the regression equation might differ from what we got without Excel. See the above example, where the Excel generated equation for the regression line differs from the one we had originally obtained by using the formula. The difference is due to roundoff error. Because of that potential difference, a recommendation is to use software, like Excel, to determine the linear equation. The line  is called the "best fit line" because it is the one that can best "represent" all the points of the scatterplot.

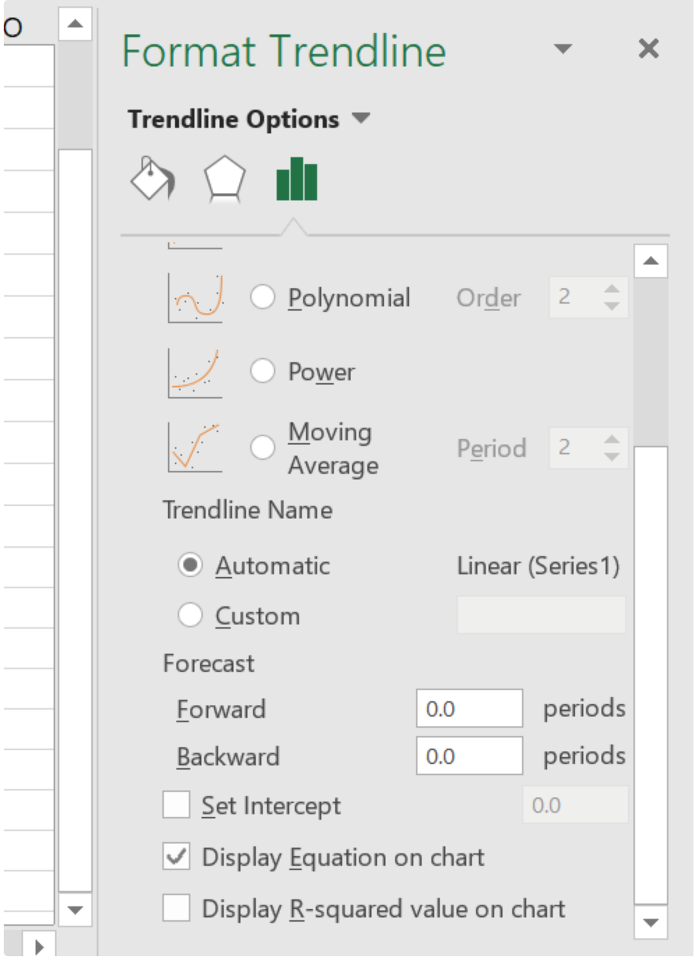
To draw the regression line with Excel, once the scatterplot is drawn, press on the graph area and press the plus sign, "+."



Once you select “Trendline,” the regression line will appear. Then select “More Options.”



Next, scroll down and check “Display Equation on chart.” The linear regression will now appear on the chart.



Format Trendline

Trendline Options

○ Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Linear (Series1)

○ Custom

Forecast

Forward 0.0 periods

Backward 0.0 periods

☐ Set Intercept 0.0

☑ Display Equation on chart

☐ Display R-squared value on chart

From the regression line the slope of the line is

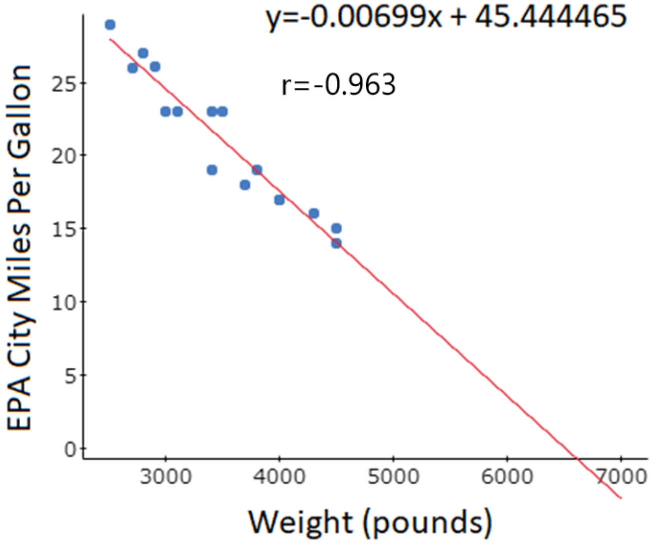
This means that for every increase of 1 minute with the customer, the score increases by approximately 12.51.

We may use the regression equation to make **predictions** about *y* if given the value of *x*.

For example, if a worker spends 3 minutes on the phone, the predicted score is

A problem that could arise with prediction occurs when one estimates y based on *x* values that are outside of the range of the data that were initially used to obtain the regression line. This is referred to as **extrapolation**.

A great example illustrating the potential pitfall is shown in the following figure that shows a car that weighs 6,501 pounds gets zero miles per gallon, which is not possible. The cars that were used to obtain this regression line ranged from approximately 2,500 to 4,500 pounds, and the car in question is beyond the last data point by about 2,000 pounds.



The assumed simple linear regression model is as follows:

From the estimated regression equation, , is an estimate of the population slope, , and is an estimate of . The error term, , is a random variable whose expected value (or mean) is zero, .

Similarly, the assumed **multiple linear regression model** is as follows:

For example, suppose a manager wants to see if there is a linear relationship between yearly pay, years of service, and number of professional development activities. The multiple regression equation would be:

$y$ = yearly pay
$x_1$ = years of service
$x_2$ = number of professional development activities

Here is another example. Suppose a car dealership owner wants to determine if there is a linear relationship between Model A price, miles, and age. The multiple regression equation would be:

$y$ = Model A price
$x_1$ = the number of miles that the Model A car has
$x_2$ = the Model A car age

The “best fit” regression line would be

### 3.1. More on Linear Equations

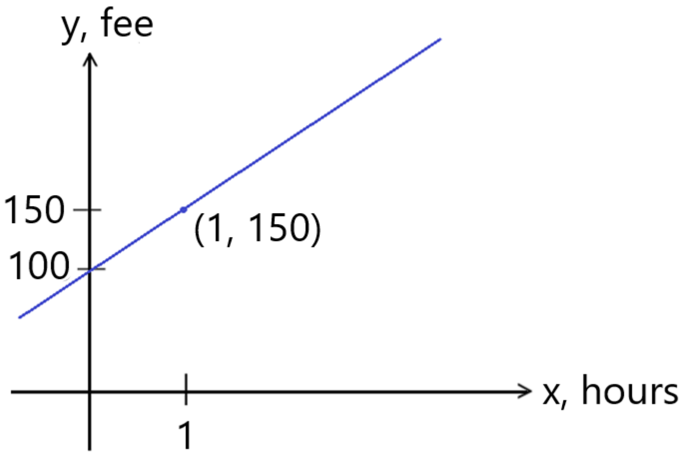
For example, suppose an accountant charges \$50 per hour plus an initial fee of \$100. If we let $y$ represent the total cost (fee) to employ the accountant for $x$ hours, then the cost equation is the linear expression

The $x$-term is the independent variable. The y-term is the dependent variable. The value of $y$ depends on the $x$-value. The $y$-intercept is the point $(x, y) = (0, 100)$. To graph the equation, we need one more point. If we choose $x = 1$, then the corresponding $y$-values is

Hence, $(x, y) = (0,150)$ is an additional point on the line. The graph is thus:

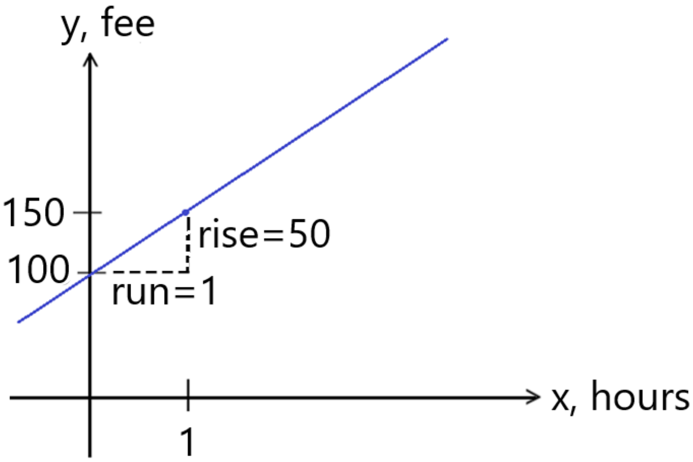
The $x$-term is the independent variable. The y-term is the dependent variable. The value of $y$ depends on the $x$-value. The $y$-intercept is the point $(x, y) = (0, 100)$. To graph the equation, we need one more point. If we choose $x = 1$, then the corresponding $y$-values is

Hence, $(x, y) = (0,150)$ is an additional point on the line. The graph is thus:



Keep in mind that the linear equation is only defined for non-negative $x$-values. This is because hours worked cannot be negative. The $y$-intercept of $y = 100$ represents the initial fee right before the accountant begins consultation. The slope:

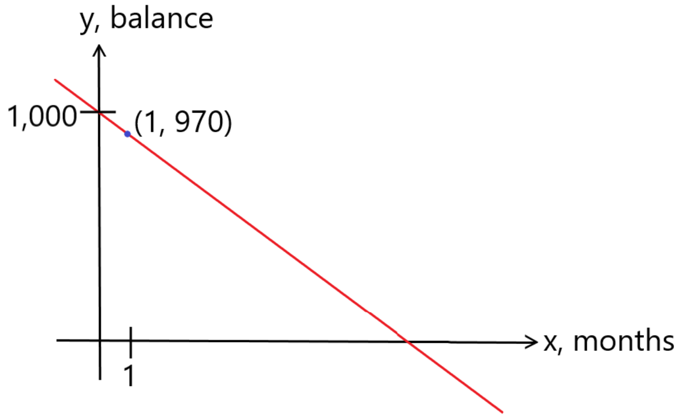
means that for every hour the accountant is employed, the fee increases by \$50:



Suppose an account initially has \$1000, and that \$30 are withdrawn each month for toll road charges. If we let $y$ denote the balance after $x$ months then the linear equation is:

The slope is -30, and the $y$-intercept is $(0, 1000)$. To graph the line, we can pick $x = 1$; for example, to find the $y$-value:

Hence, the point $(x, y) = (1, 970)$ is on the line:



Note that the slope is negative, -30. This means that the line is decreasing. In other words, *y* decreases as *x* increases. Also, the slope of -30:

means that an increase of 1 month corresponds to a balance that decreases by $30.

## 3.2. More on the Linear Regression Equation

Recall that the regression equation provides the “best fit line” that describes all of the data. The following is an additional example.

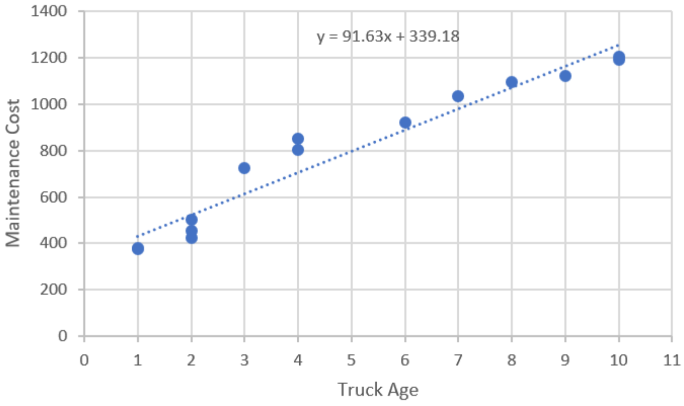
EXAMPLE

A trucking company wants to determine if there is a linear relationship between the age of its trucks and the yearly maintenance costs. Suppose the ages and maintenance costs for last year of 14 of its trucks are obtained.

| Truck Age in Years | Maintenance Cost |
|---|---|
| 3 | 723 |
| 1 | 374 |
| 7 | 1032 |
| 10 | 1205 |
| 1 | 380 |
| 4 | 805 |
| 2 | 453 |
| 9 | 1123 |
| 10 | 1189 |
| 8 | 1095 |
| 6 | 920 |
| 2 | 422 |
| 2 | 500 |
| 4 | 850 |

Using the CORREL feature of Excel, the sample correlation coefficient is $r$ = 0.969. Using the methods of the previous page, we can conclude there is a significant positive linear relationship between truck age and maintenance cost.

The Excel scatterplot with the regression line included is:



Thus, the regression equation is

The slope is positive. This indicates that as truck age increases, the yearly maintenance cost also increases. The slope means that for each 1-year increase in truck age, the yearly maintenance cost increases by approximately $91.63.

We can also predict the maintenance cost for a 5-year old truck:

We can predict that a 5-year old truck will have a yearly maintenance cost of $797.33.

## 4. Summary

For this last module, we first used ANOVA to determine if more than two population means are equal. We then studied whether there is a significant linear relation between two variables. We also covered how to obtain and understand the linear regression equation.

In this course we have examined the basics of statistical analysis. We have learned to examine the characteristics of data and to determine whether a relationship exists between variables and data. We have also explored how to represent data and their outcomes graphically.

The use of statistics can facilitate effective and efficient operations in all areas of organizations. By being able to analyze data, you can create projections and scenarios that can then be evaluated not only for accuracy but also for real-world application. As we have learned in this course, while statistical formulas may look intimidating, the math calculations are quite simple. We have also learned how established data tables can further reduce statistical calculation workload when completing such calculations manually.

Fortunately, in today’s technological world, we have the benefit of numerous software programs that allow us to identify data parameters simply, and then compute whatever statistical calculations we are seeking. The job of the decision maker, then, is to be able to interpret the outcomes correctly and make appropriate decisions from the information. Sound easy? Since you have now been through the course and had to calculate statistics manually, it probably does sound simple—and it should. You are now well-equipped to understand the variety of formulas available to you in order to analyze and evaluate data fully.

As you leave this course, continue to observe statistics in your everyday life and consider how the numbers might have been derived. Better yet, ask. You might be surprised by just how much more you know than those who conducted the studies, and—at a minimum—you will be able to evaluate the methods used and, therefore, the validity of the corresponding claims.

We hope that you have enjoyed the course and that you will incorporate statistics in your future decisions for optimum solutions and decisions.
Here is the list of the objectives that we have covered and are part of the Mastery Exercises in Knewton Alta:

- Determine appropriate situations for a one-way ANOVA test, and identify the null and alternative hypotheses
- Determine the degrees of freedom for the numerator and denominator for one-way ANOVA test
- Determine the critical value and rejection region for one-way ANOVA test
- Calculate the test statistic for a one-way ANOVA test
- Make a decision for the hypothesis test using critical value/rejection region method and interpret results—Excel
- Make a decision for the hypothesis test using the *p*-value method and interpret results—Excel
- Understand properties of linear equations in business applications
- Understand the relationship between scatter plots and table and determine patterns in business applications
- Find the linear regression equation given a list of data points with business applications
- Find and interpret the correlation coefficient in business contexts
- Make predictions about business scenarios using a line of best fit
- Find outliers in a business-related data set
- Identify applications where multiple regression can be performed
- Calculate the correlation coefficient using Technology—Excel
- Determine the best fit linear regression equation using Technology—Excel

### Check Your Understanding

Check your understanding of the concepts presented here in Module 8. If you struggle to answer these questions, review the primary and secondary pages in the module and try these activities again.

To complete these activities, read the question and the given answers. Click on the answer you think is right to check your thinking.

**Question #1:**

A manufacturer makes three types of car batteries, small, medium and large. The number of months that randomly selected batteries of each type lasted is shown below. Assume that the samples are independent and that the lasting times are approximately normally distributed with equal standard deviations.

| Small | Medium | Large |
|---|---|---|
| 54 | 48 | 70 |
| 53 | 51 | 53 |
| 58 | 55 | 52 |
| 52 | 50 | 67 |
| 55 | 60 | 48 |
| 49 | 58 | 52 |
| 52 | 52 | 51 |

A researcher wants to test these hypotheses at the 0.1 level of significance regarding the mean lasting times:

$H_0$: $\mu_1 = \mu_2 = \mu_3$ (The three population mean lasting times are equal.)
$H_a$: Not all the population means are equal.

*Find the ANOVA test statistic.*

a. 0.594

Not quite! Re-read the question and examine the table, and try again!

b. 18.143

Not quite! Re-read the question and examine the table, and try again!

c. 2.624

Not quite! Re-read the question and examine the table, and try again!

d. 0.535

Correct! See the ANOVA table:

| ANOVA: Single Factor | | | | | | |
|---|---|---|---|---|---|---|
| SUMMARY | | | | | | |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* | | |
| Column 1 | 7 | 373 | 53.28571 | 7.904762 | | |
| Column 2 | 7 | 374 | 53.42857 | 19.28571 | | |
| Column 3 | 7 | 393 | 56.14286 | 74.47619 | | |
| | | | | | | |
| ANOVA | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *p-value* | *F crit* |
| Between Groups | 36.28571 | 2 | 18.14286 | 0.535363 | 0.594494 | 2.623947 |
| Within Groups | 610 | 18 | 33.88889 | | | |
| | | | | | | |
| Total | 646.2857 | 20 | | | | |

**Question #2:**

A manufacturer makes three types of car batteries, small, medium and large. The number of months that randomly selected batteries of each type lasted is shown below. Assume that the samples are independent and that the lasting times are approximately normally distributed with equal standard deviations.

| Small | Medium | Large |
|---|---|---|
| 54 | 48 | 70 |
| 53 | 51 | 53 |
| 58 | 55 | 52 |
| 52 | 50 | 67 |
| 55 | 60 | 48 |
| 49 | 58 | 52 |
| 52 | 52 | 51 |

A researcher wants to test these hypotheses at the 0.1 level of significance regarding the mean lasting times:

$H_0$: $\mu_1 = \mu_2 = \mu_3$ (The three population mean lasting times are equal.)
$H_a$: Not all the population means are equal.

*Find the p-value.*

a. 0.594

Correct! See the ANOVA table:

| | | | | | | |
|---|---|---|---|---|---|---|
| ANOVA: Single Factor | | | | | | |
| SUMMARY | | | | | | |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* | | |
| Column 1 | 7 | 373 | 53.28571 | 7.904762 | | |
| Column 2 | 7 | 374 | 53.42857 | 19.28571 | | |
| Column 3 | 7 | 393 | 56.14286 | 74.47619 | | |
| | | | | | | |
| ANOVA | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *p-value* | *F crit* |
| Between Groups | 36.28571 | 2 | 18.14286 | 0.535363 | 0.594494 | 2.623947 |
| Within Groups | 610 | 18 | 33.88889 | | | |
| | | | | | | |
| Total | 646.2857 | 20 | | | | |

b. 18.143

Not quite! Re-read the question and examine the table, and try again!

c. 2.624

Not quite! Re-read the question and examine the table, and try again!

d. 0.535

Not quite! Re-read the question and examine the table, and try again!

**Question #3**:

A manufacturer makes three types of car batteries, small, medium and large. The number of months that each type of randomly selected batteries lasted is shown below. Assume that the samples are independent and that the lasting times are approximately normally distributed with equal standard deviations.

| Small | Medium | Large |
|---|---|---|
| 54 | 48 | 70 |
| 53 | 51 | 53 |
| 58 | 55 | 52 |
| 52 | 50 | 67 |
| 55 | 60 | 48 |
| 49 | 58 | 52 |
| 52 | 52 | 51 |

A researcher wants to test these hypotheses at the 0.1 level of significance regarding the mean lasting times:

$H_0$: $\mu_1 = \mu_2 = \mu_3$ (The three population mean lasting times are equal.)
$H_a$: Not all the population means are equal.

Find the critical value.

a. 0.594

Not quite! Re-read the question and examine the table, and try again!

b. 18.143

Not quite! Re-read the question and examine the table, and try again!

c. 2.624

Correct! See the ANOVA table:

| ANOVA: Single Factor | | | | | | |
|---|---|---|---|---|---|---|
| SUMMARY | | | | | | |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* | | |
| Column 1 | 7 | 373 | 53.28571 | 7.904762 | | |
| Column 2 | 7 | 374 | 53.42857 | 19.28571 | | |
| Column 3 | 7 | 393 | 56.14286 | 74.47619 | | |
| | | | | | | |
| ANOVA | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *p-value* | *F crit* |
| Between Groups | 36.28571 | 2 | 18.14286 | 0.535363 | 0.594494 | 2.623947 |
| Within Groups | 610 | 18 | 33.88889 | | | |
| | | | | | | |
| Total | 646.2857 | 20 | | | | |

d. 0.535

Not quite! Re-read the question and examine the table, and try again!

**Question #4**

A manufacturer makes three types of car batteries, small, medium and large. The number of months that each type of randomly selected batteries lasted is shown below. Assume that the samples are independent and that the lasting times are approximately normally distributed with equal standard deviations.

| Small | Medium | Large |
|---|---|---|
| 54 | 48 | 70 |
| 53 | 51 | 53 |
| 58 | 55 | 52 |
| 52 | 50 | 67 |
| 55 | 60 | 48 |
| 49 | 58 | 52 |
| 52 | 52 | 51 |

A researcher wants to test these hypotheses at the 0.1 level of significance regarding the mean lasting times:

$H_0$: $\mu_1 = \mu_2 = \mu_3$ (The three population mean lasting times are equal.)
$H_a$: Not all the population means are equal.

*Determine if the null hypothesis should be rejected.*

a. Fail to reject the null hypothesis.

Correct! Since the *p*-value is greater than the level of significance (0.594 > 0.1), we would fail to reject the null hypothesis. See the ANOVA table:

| ANOVA: Single Factor | | | | | | |
|---|---|---|---|---|---|---|
| SUMMARY | | | | | | |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* | | |
| Column 1 | 7 | 373 | 53.28571 | 7.904762 | | |
| Column 2 | 7 | 374 | 53.42857 | 19.28571 | | |
| Column 3 | 7 | 393 | 56.14286 | 74.47619 | | |
| | | | | | | |
| ANOVA | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *p-value* | *F crit* |
| Between Groups | 36.28571 | 2 | 18.14286 | 0.535363 | 0.594494 | 2.623947 |
| Within Groups | 610 | 18 | 33.88889 | | | |
| | | | | | | |
| Total | 646.2857 | 20 | | | | |

b. Reject the null hypothesis.

Not quite! Re-read the question and examine the table, and try again!

c. Accept the alternative hypothesis.

Not quite! Re-read the question and examine the table, and try again!

d. This cannot be determined.

Not quite! Re-read the question and examine the table, and try again!

**Question #5**

The following are the population (in thousands) and cost of living indices for 10 small cities.

| Population | Cost of living index |
|---|---|
| 16.5 | 92 |
| 21.5 | 105 |
| 18.5 | 92 |
| 23.5 | 105 |
| 17 | 78 |
| 18.5 | 77 |
| 21 | 104 |
| 20.5 | 98 |
| 16 | 68 |
| 22 | 100 |

*What is the correlation coefficient?*

a. 0.698

Not quite! Use the CORREL function of Excel and try again!

b. 0.781

Not quite! Use the CORREL function of Excel and try again!

c. 0.839

Using the CORREL function of Excel, the correlation coefficient is r=0.839

d. 0.855

Not quite! Use the CORREL function of Excel and try again!

**Question #6**

The following are the population (in thousands) and cost of living indices for 10 small cities.

| Population | Cost of living index |
|---|---|
| 16.5 | 92 |
| 21.5 | 105 |
| 18.5 | 92 |
| 23.5 | 105 |
| 17 | 78 |
| 18.5 | 77 |
| 21 | 104 |
| 20.5 | 98 |
| 16 | 68 |
| 22 | 100 |

*Determine if there is a significant linear relationship between the population of a small city and the cost of living index.*

a. There is a significant linear relationship between the population of a small city and the cost of living index.

The answer is correct! Using the CORREL function of Excel, the correlation coefficient is r=0.839. The test statistic is

The critical values with *n-2=10-2=8* degrees of freedom are obtained from Appendix page 608, Table A12:

Since the test statistic is larger than the positive critical value (4.36 > 2.306), we would conclude that at the 0.05 level of significance, there is a significant linear relationship between the two variables.

b. There is not a significant linear relationship between the population of a small city and the cost of living index.

Not quite! Use the CORREL function of Excel and try again!

c. It cannot be determined if there is a significant linear relationship between the population of a small city and the cost of living index.

Not quite! Use the CORREL function of Excel and try again!

d. There is a possible linear relationship between the population of a small city and the cost of living index.

Not quite! Use the CORREL function of Excel and try again!

**Question #7**

The following are the population (in thousands) and cost of living indices for 10 small cities.

| Population | Cost of living index |
|---|---|
| 16.5 | 92 |
| 21.5 | 105 |
| 18.5 | 92 |
| 23.5 | 105 |
| 17 | 78 |
| 18.5 | 77 |
| 21 | 104 |
| 20.5 | 98 |
| 16 | 68 |
| 22 | 100 |

*What is the regression equation? Choose the best response.*

a.

Not quite! Review the question and table, and try again.

b.

The answer is Correct! With Excel, the regression equation is

c.

Not quite! Review the question and table, and try again.

d.

Not quite! Review the question and table, and try again.

## References

None