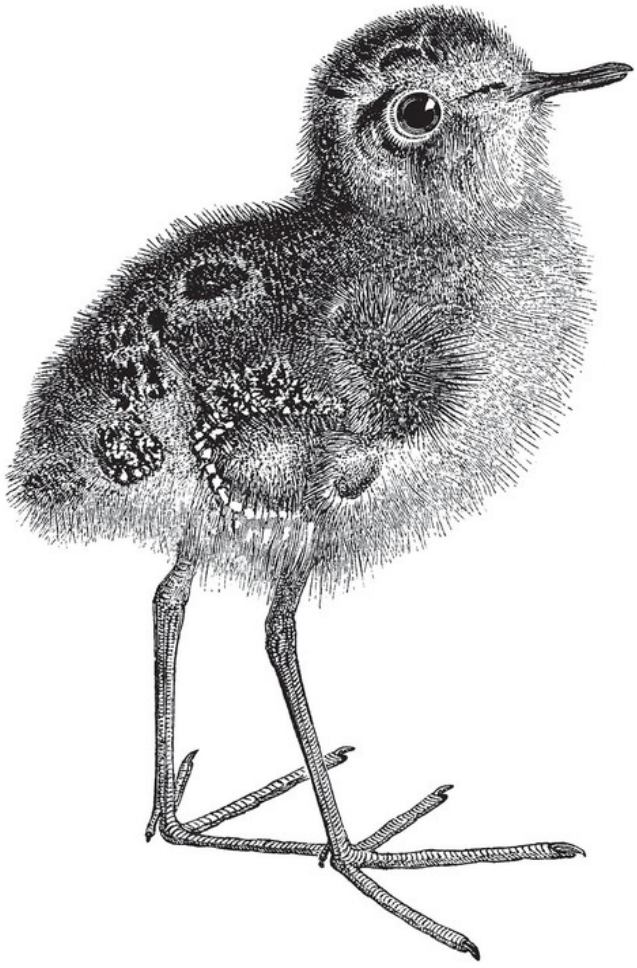
O'REILLY®

Excel Cookbook

Recipes for Mastering Microsoft Excel



Early Release

RAW & UNEDITED

Dawn Griffiths

Excel Cookbook

Recipes for Mastering Microsoft Excel

Dawn Griffiths

Excel Cookbook

by Dawn Griffiths

Printed in the United States of America.

Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O’Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (http://oreilly.com). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: David Michelson

Development Editor: Corbin Collins

Production Editor: Christopher Faucher

Interior Designer: David Futato

Cover Designer: Karen Montgomery

April 2024: First Edition

Revision History for the Early Release

- 2023-02-13: First release

See http://oreilly.com/catalog/errata.csp?isbn=9781098143329 for release details.

978-1-098-14326-8

Chapter 1. The Analysis ToolPak

A NOTE FOR EARLY RELEASE READERS

With Early Release ebooks, you get books in their earliest form—the author's raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.

This will be the 9th chapter of the final book. Please note that the GitHub repo will be made active later on.

If you have comments about how we might improve the content and/or examples in this book, or if you notice missing material within this chapter, please reach out to the editor at ccollins@oreilly.com.

As you've seen in [Link to Come], Excel offers a wealth of functions you can use for statistical analysis. Using these, however, can sometimes be complex and time-consuming.

An alternative approach is to use Excel's Analysis ToolPak. This add-in provides data analysis tools that address statistical and engineering tasks such as generating statistics, performing hypothesis tests and even running Fourier analyses.

This chapter guides you through installing the Analysis ToolPak and using each tool it provides.

9.1 Installing the Analysis ToolPak

Problem

You have some data you want to analyze, and you want to install a suite of tools to help you.

Solution

Load the Analysis ToolPak add-in to Excel. Once loaded, you can access the ToolPak from the Data menu by going to the Analyze group and choosing Data Analysis.

To load the Analysis ToolPak in Excel for Windows:

1. Choose File ⇒ Options.
2. Select Add-ins.
3. In the Manage box at the bottom of the screen, choose the Excel Add-ins option and click Go.
4. In the Add-ins dialog box, place a check in the Analysis ToolPak checkbox and click OK (see Figure 1-1).

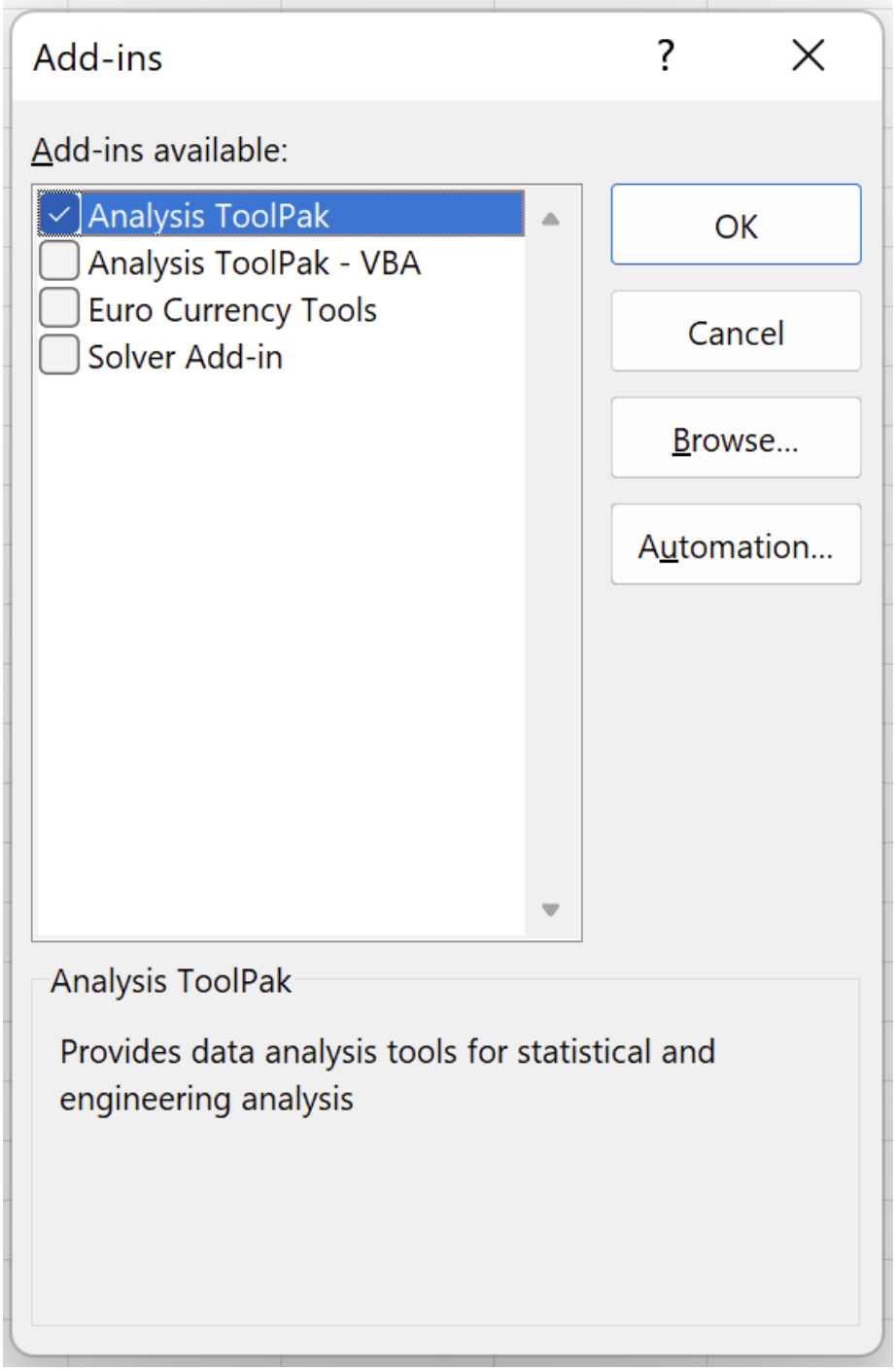


Figure 1-1. Loading the Analysis ToolPak in Excel for Windows.

To load the Analysis ToolPak in Excel for Mac:

1. Choose Tools ⇒ Excel add-ins.

2. In the Add-ins dialog box, place a check in the Analysis ToolPak checkbox and click OK (see Figure 1-2).

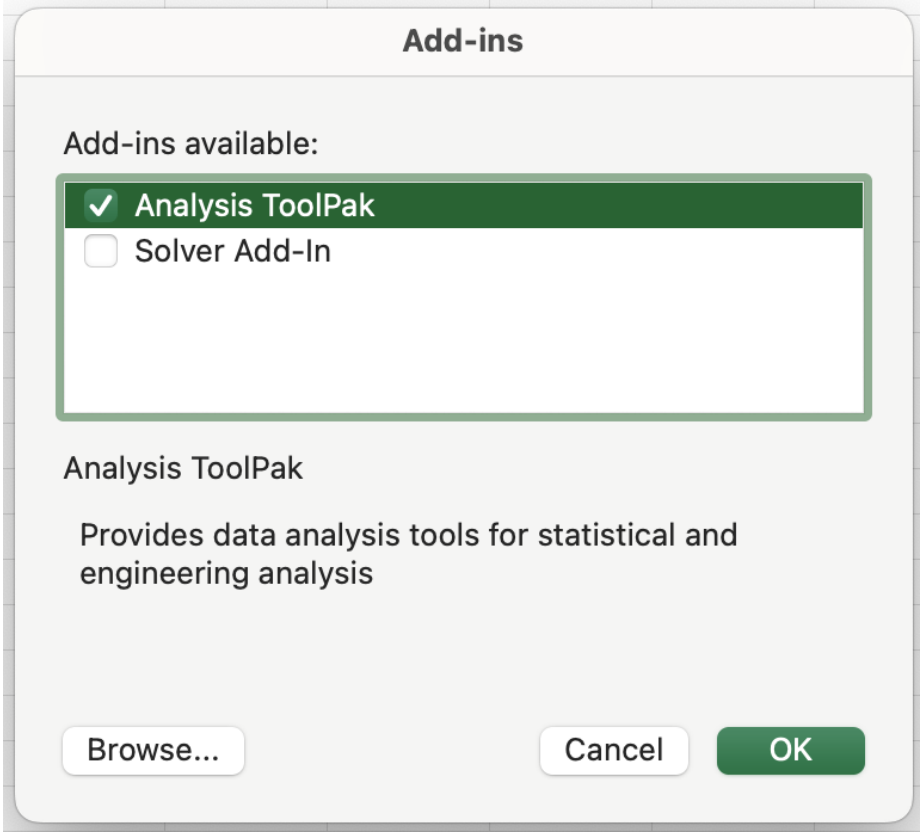


Figure 1-2. Loading the Analysis ToolPak in Excel for Mac.

Discussion

The Analysis ToolPak brings extra features to Excel that help make statistical and engineering analyses less time-consuming. It includes tools to generate statistics, generate random numbers and samples, perform hypothesis tests, and more. I'll discuss these tools in more detail in different recipes.

9.2 Generating Descriptive Statistics

Problem

You have sample data and want to generate standard statistics from it.

Solution

Use the Analysis ToolPak's Descriptive Statistics tool.

Suppose you want to generate standard descriptive statistics for the scores of 15 students. The range A1:A16 lists the scores, including a label in row 1 (see Figure 1-3).

	A	B	C	D	E	F	G	H
1	Scores							
2	59							
3	63							
4	18							
5	28							
6	38							
7	35							
8	22							
9	47							
10	37							
11	55							
12	59							
13	47							
14	44							
15	42							
16	47							

Figure 1-3. Scores data.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Descriptive Statistics from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-4).
3. Choose **A1:A16** in the Input Range box for the range of the scores, including the label in A1.
4. Choose the Columns option in the Grouped By section because the scores are listed in a column.

5. Place a check in the Labels in First Row checkbox because the first row of the input range contains the data's label.
6. Choose one of the Output Options to specify where you want the Descriptive Statistics tool to output the results (for example, the Output Range **C1**).
7. Place a check in the Summary Statistics checkbox.
8. Place checks in the Kth Largest and Kth Smallest checkboxes if you want to include these statistics in the output (optional), and enter a value for *K*.

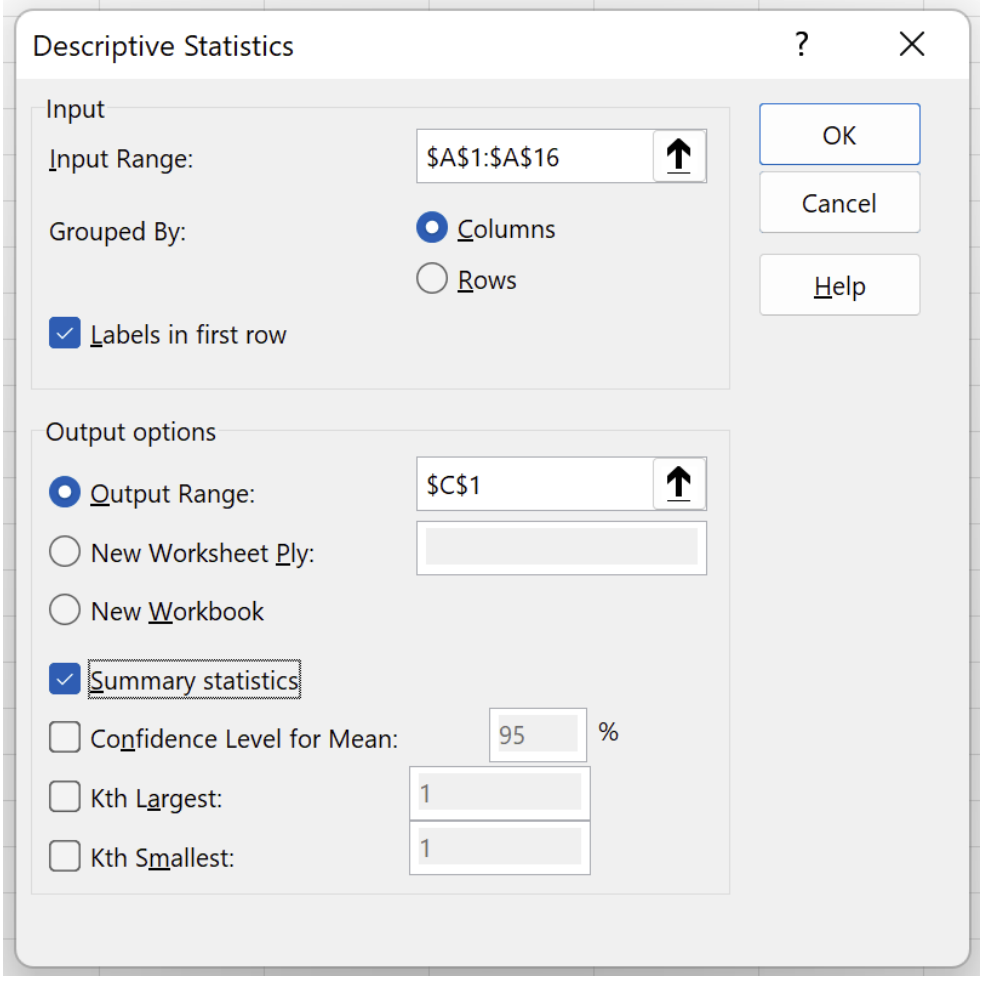


Figure 1-4. The Descriptive Statistics tool's dialog box.

When you click OK, the Descriptive Statistics tool generates descriptive statistics for the data (see Figure 1-5).

Scores		*Scores*	
59			
63		Mean	42.73333
18		Standard Error	3.464743
28		Median	44
38		Mode	47
35		Standard Deviation	13.41889
22		Sample Variance	180.0667
47		Kurtosis	-0.571
37		Skewness	-0.30647
55		Range	45
59		Minimum	18
47		Maximum	63
44		Sum	641
42		Count	15
47			

Figure 1-5. The Descriptive Statistics tool's output.

Discussion

The Descriptive Statistics tool offers a quick way of generating standard statistics for a sample data set. These include measures of central tendency (such as the mean, median, and mode) and measures of variability (such as the sample variance, standard deviation, and skew).

This recipe uses the Descriptive Statistics tool to generate statistics for a single data set. However, you can also generate statistics for multiple data sets at a time by listing each one in a separate column and choosing an Input Range that includes the entire range (see Figure 1-6).

Scores A	Scores B		*Scores A*		*Scores B*	
59	76					
63	93		Mean	42.73333	Mean	63.46667
18	28		Standard Error	3.464743	Standard Error	4.675027
28	62		Median	44	Median	62
38	57		Mode	47	Mode	82
35	78		Standard Deviation	13.41889	Standard Deviation	18.1063
22	77		Sample Variance	180.0667	Sample Variance	327.8381
47	82		Kurtosis	-0.571	Kurtosis	-0.52724
37	82		Skewness	-0.30647	Skewness	-0.30691
55	42		Range	45	Range	65
59	66		Minimum	18	Minimum	28
47	53		Maximum	63	Maximum	93
44	54		Sum	641	Sum	952
42	61		Count	15	Count	15
47	41					

Figure 1-6. The Descriptive Statistics tool's output for multiple data sets.

> **WARNING**
>
> The Descriptive Statistics tool outputs values instead of formulas, so you must rerun the tool if the underlying data changes. Alternatively, consider calculating the statistics using formulas.

9.3 Generating a Confidence Interval for the Population Mean

Problem

You have sample data and want to use it to estimate the population mean plus or minus some margin of error.

Solution

Use the Analysis ToolPak's Descriptive Statistics tool.

Follow the recipe for "9.2 Generating Descriptive Statistics", place an additional check in the Confidence Level for the Mean checkbox, and specify a confidence level, for example, 95% (see Figure 1-7). The

confidence level is the probability that the population mean falls within the confidence interval the tool generates.

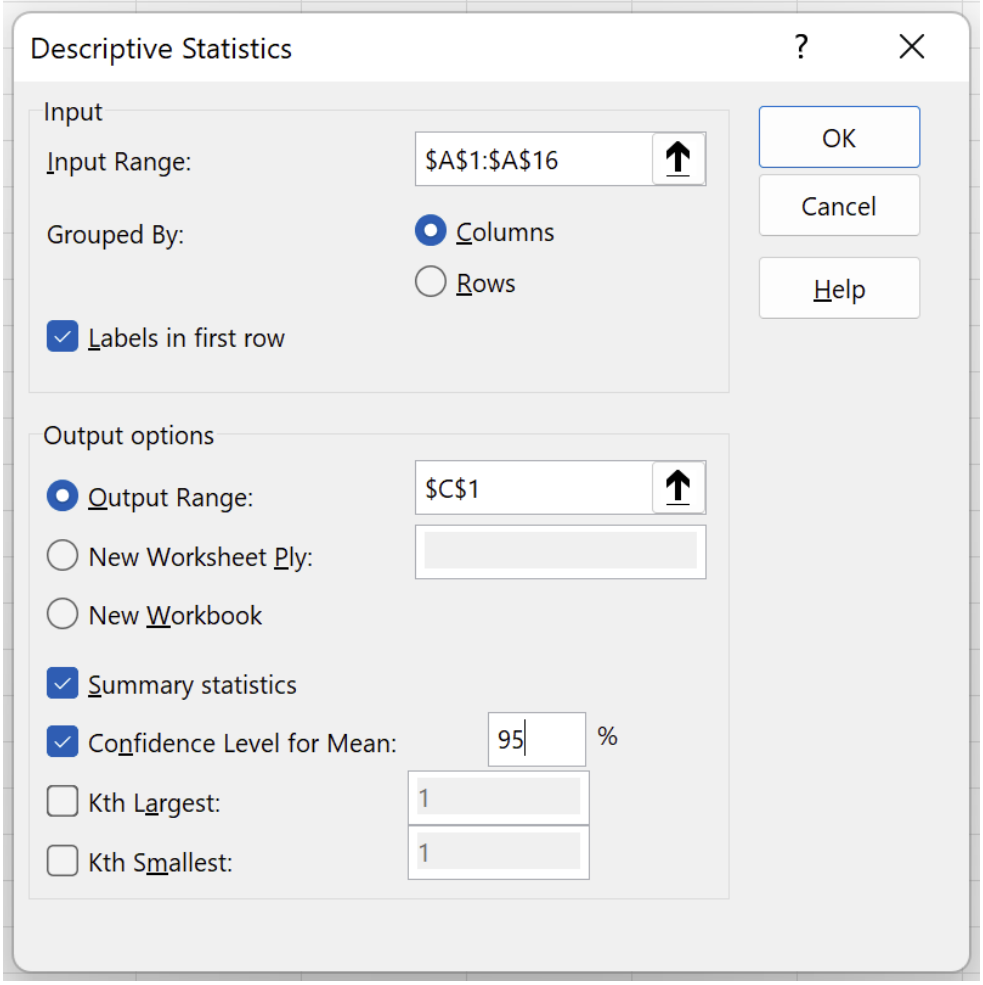


Figure 1-7. The Descriptive Statistics dialog box for creating a confidence interval.

When you click OK, the Descriptive Statistics tool generates a confidence statistic for the population mean (see Figure 1-8). The confidence interval is the sample mean ± the confidence statistic. In this example, there's a 95% probability that the population mean is 42.73333 ± 7.431135).

Scores		Scores	
59			
63		Mean	42.73333
18		Standard Error	3.464743
28		Median	44
38		Mode	47
35		Standard Deviation	13.41889
22		Sample Variance	180.0667
47		Kurtosis	-0.571
37		Skewness	-0.30647
55		Range	45
59		Minimum	18
47		Maximum	63
44		Sum	641
42		Count	15
47		Confidence Level(95.0%)	7.431135

Figure 1-8. The Descriptive Statistics tool's output including confidence.

Discussion

This recipe shows you how to estimate the population mean from sample data, plus or minus some margin of error given by the confidence statistic. The confidence level specifies the probability that the population mean falls within the confidence interval.

Notice that this recipe places a check in both the Summary Statistics and Confidence Level for the Mean checkboxes. Doing so means that the Descriptive Statistics tool outputs the sample mean and confidence statistic, which are both needed to specify the confidence interval.

> **TIP**
>
> The Descriptive Statistics tool calculates the confidence statistic using the t-distribution. If you want to calculate the confidence statistic using the Normal distribution, use the `CONFIDENCE.NORM` function instead of the Descriptive Statistics tool.

9.4 Generating Ordinal and Percentage Rank Statistics

Problem

You have a set of numbers and want to find the rank of each one when sorted in descending order.

Solution

Use the Analysis ToolPak's Rank and Percentile tool.

Suppose you have a set of scores from 15 students, and you want to analyze the relative standing of each one. The range A1:A16 lists the scores, including a label in row 1 (see Figure 1-3).

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Rank and Percentile from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-9).

3. Choose **A1:A16** in the Input Range box for the range of the scores, including the label in A1.
4. Choose the Columns option in the Grouped By section because the scores are listed in a column.
5. Place a check in the Labels in First Row checkbox because the first row of the input range contains the data's label.
6. Choose one of the Output Options to specify where you want the Rank and Percentile tool to output the results (for example, the Output Range **C1**).

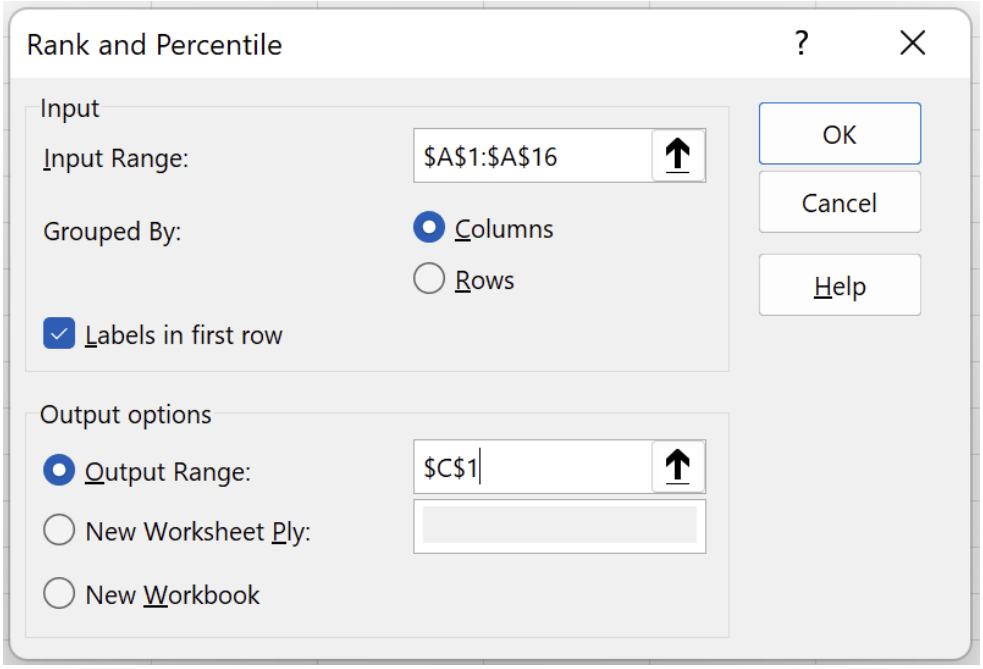


Figure 1-9. The Rank and Percentile tool's dialog box.

When you click OK, the Rank and Percentile tool outputs the scores in descending order, along with each one's rank and percentage rank (see Figure 1-10).

Scores		Point	Scores	Rank	Percent
59		2	63	1	100.00%
63		1	59	2	85.70%
18		11	59	2	85.70%
28		10	55	4	78.50%
38		8	47	5	57.10%
35		12	47	5	57.10%
22		15	47	5	57.10%
47		13	44	8	50.00%
37		14	42	9	42.80%
55		5	38	10	35.70%
59		9	37	11	28.50%
47		6	35	12	21.40%
44		4	28	13	14.20%
42		7	22	14	7.10%
47		3	18	15	0.00%

Figure 1-10. The Rank and Percentile tool's output.

Discussion

The Rank and Percentile tool uses the `RANK.EQ` and `PERCENTRANK.INC` functions to return the ordinal and percentage rank of each value in the input range.

The `RANK.EQ(`*`value, range`*`)` function returns the rank of a value in the range. Where there are tied values, it returns their top rank. If you need to know the average rank of the tied values instead, use the `RANK.AVG` function.

The `PERCENTRANK.INC(`*`range, value`*`)` returns the rank of a value in a range as a percentage of the data set, including the range's first and last values. To exclude the first and last values, use the `PERCENTRANK.EXC` function instead.

9.5 Generating a Frequency Distribution

Problem

You have a set of numeric data and want to determine how values are distributed by dividing them into intervals and counting how

many are in each interval. You optionally want to show a running total for the percentage of values for each bin, and display the results on a chart.

Solution

Use the Analysis ToolPak's Histogram tool.

Suppose you have a set of scores for 15 students. You want to group the data into bins and count how many values occur in each bin (the frequency). The range A2:A16 lists the scores, C2:C11 lists the upper limits for each bin, and the first row contains labels (see Figure 1-11).

	A	B	C	D	E	F	G
1	Scores		Bin upper limits				
2	59		10				
3	63		20				
4	18		30				
5	28		40				
6	38		50				
7	35		60				
8	22		70				
9	47		80				
10	37		90				
11	55		100				
12	59						
13	47						
14	44						
15	42						
16	47						

Figure 1-11. Scores and bin upper limits.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Histogram from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-12).
3. Choose **A1:A16** in the Input Range box for the range of the scores, including the label in A1.

4. Choose **C1:C11** in the Bin Range box for the range of the bin upper limits, including the label in C1.
5. Place a check in the Labels checkbox because the first row of the input and bin ranges contain labels.
6. Choose one of the Output Options to specify where you want the Histogram tool to output the results (for example, the Output Range **E1**).
7. Place a check in the Pareto (sorted histogram) checkbox (optional) to show extra columns with the bins sorted by frequency.
8. Place a check in the Cumulative Percentage checkbox (optional) to output a running total for the percentage of values held in each bin.
9. Place a check in the Chart Output checkbox (optional) to output a histogram or Pareto chart (depending on the other selected options).

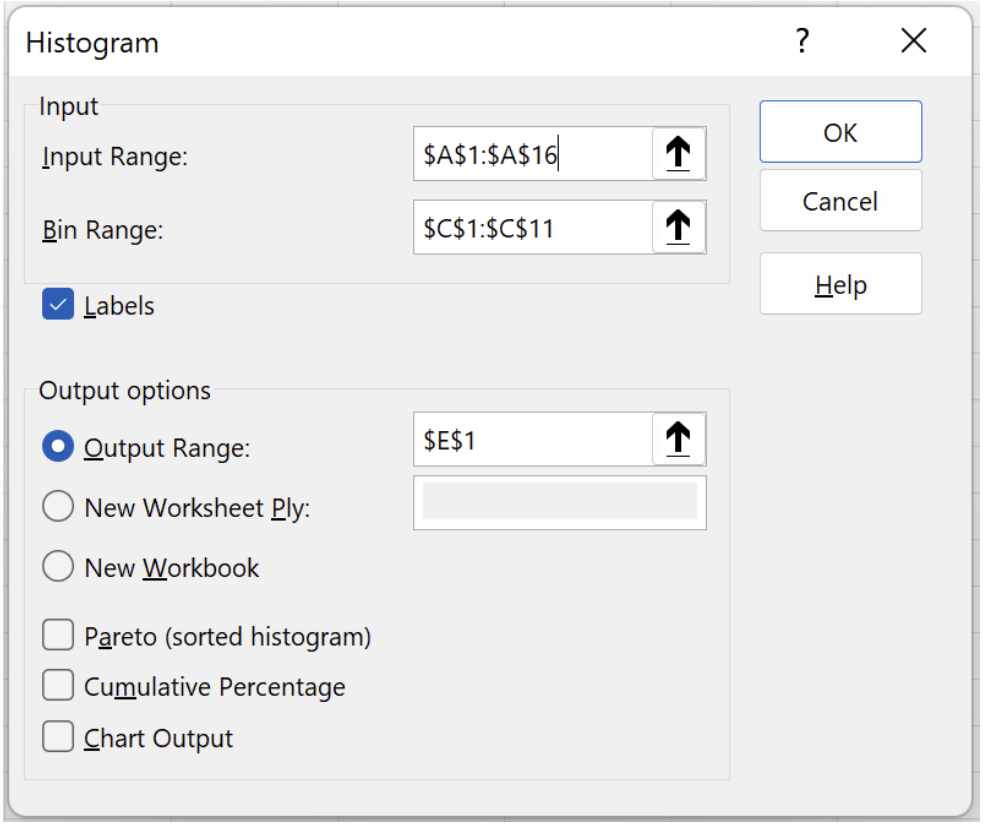


Figure 1-12. The Histogram tool's dialog box.

When you click OK, the Histogram tool outputs a table showing the bin upper limits and the frequency of each one. Depending on your selected options, it may also show you the cumulative percentage and sort the results by bin frequency (see Figure 1-13).

Bin upper limits	*Frequency*	*Cumulative %*	*Bin upper limits*	*Frequency*	*Cumulative %*
10	0	0.00%	50	5	33.33%
20	1	6.67%	40	3	53.33%
30	2	20.00%	60	3	73.33%
40	3	40.00%	30	2	86.67%
50	5	73.33%	20	1	93.33%
60	3	93.33%	70	1	100.00%
70	1	100.00%	10	0	100.00%
80	0	100.00%	80	0	100.00%
90	0	100.00%	90	0	100.00%
100	0	100.00%	100	0	100.00%
More	0	100.00%	More	0	100.00%

Figure 1-13. The Histogram tool's output.

If you select the Chart Output option, the Histogram tool generates an additional histogram or Pareto chart, depending on your other options (see Figure 1-14).

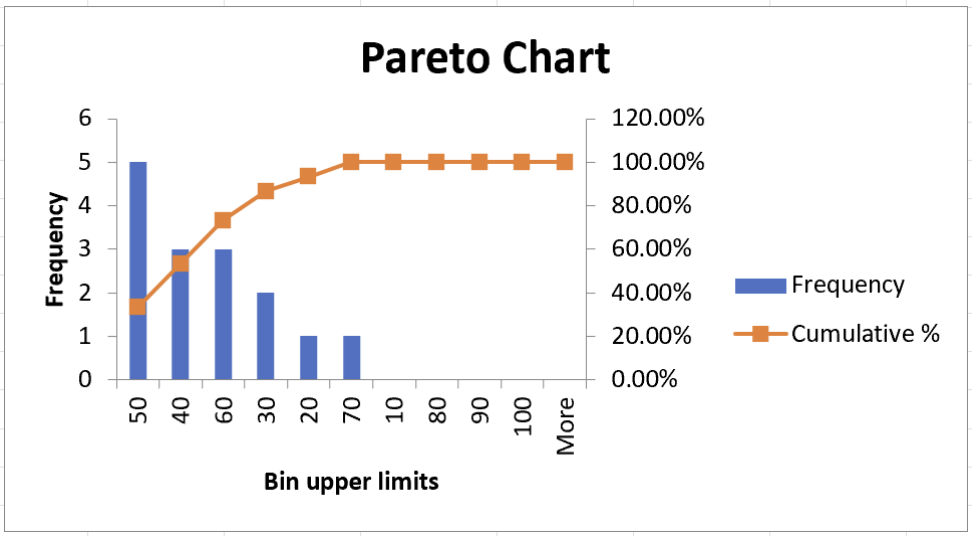


Figure 1-14. The Histogram tool's chart output.

Discussion

This recipe is a convenient way of generating various statistics relating to frequencies. The primary output shows the frequency of each bin you specify. If you don't provide the tool with a range of bin limits, it will choose them for you.

The cumulative percentage option shows the total percentages for each bin and its predecessors. For example, the output shown on the left of Figure 1-13 shows us that 20% of the score values are in the first three bins with upper limits of 10, 20, and 30. If you select the Pareto (sorted histogram) option, the results also show the cumulative percentage sorted by bin frequency. For example, the output shown on the right of Figure 1-13 shows us that 53.33% of the score values are in the bins with upper limits of 50 and 40.

Notice that the Histogram tool outputs results as values instead of formulas, so you'll need to rerun the tool if the underlying data changes.

> **TIP**
>
> Strictly speaking, there should be no gaps between the histogram and Pareto chart columns. To correct this, format the columns so that the gap width between each column is 0%.

See also

See recipes [Link to Come] and [Link to Come] for other techniques you can use to generate frequencies.

> **WARNING**
>
> If you use the Pareto (Sorted Histogram) option, you must ensure that the bin range contains values instead of formulas. If it uses formulas, you may encounter errors.

9.6 Generating Moving Averages

Problem

You have a set of data with seasonal variability and want to analyze the general trend.

Solution

Use the Analysis ToolPak's Moving Average tool.

Suppose you have sales data where revenue tends to be lower for the third quarter of each year, and you want to analyze the general trend. The range A2:A20 lists the year and quarter, B2:B20 lists the upper limits for each bin, and the first row contains labels (see Figure 1-15).

	A	B	C	D	E	F	G	H
1	Year/Quarter	Sales						
2	2018 Q1	78000						
3	2018 Q2	60000						
4	2018 Q3	45000						
5	2018 Q4	80000						
6	2019 Q1	75000						
7	2019 Q2	70000						
8	2019 Q3	50000						
9	2019 Q4	90000						
10	2020 Q1	80000						
11	2020 Q2	52000						
12	2020 Q3	60000						
13	2020 Q4	115000						
14	2021 Q1	105000						
15	2021 Q2	60000						
16	2021 Q3	80000						
17	2021 Q4	120000						
18	2022 Q1	110000						
19	2022 Q2	63000						
20	2022 Q3	80000						

Figure 1-15. Sales data per quarter.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Moving Average from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-16).
3. Choose **B1:B20** in the Input Range box for the range of the sales values, including the label in B1.
4. Place a check in the Labels in First Row checkbox because the first row of the input range contains the data's label.
5. Type **4** in the Interval box because the data lists sales for each quarter, and the seasonal pattern repeats every four quarters.
6. Choose an Output Range of **C2** so that the moving average value for each row is output next to the original sales data.
7. Place a check in the Chart Output checkbox if you want to generate a line chart with a moving average trendline (optional).

8. Place a check in the Standard Errors checkbox if you want to generate standard error alongside the moving averages (optional). This option shows the degree of variability between the actual values and the moving averages.

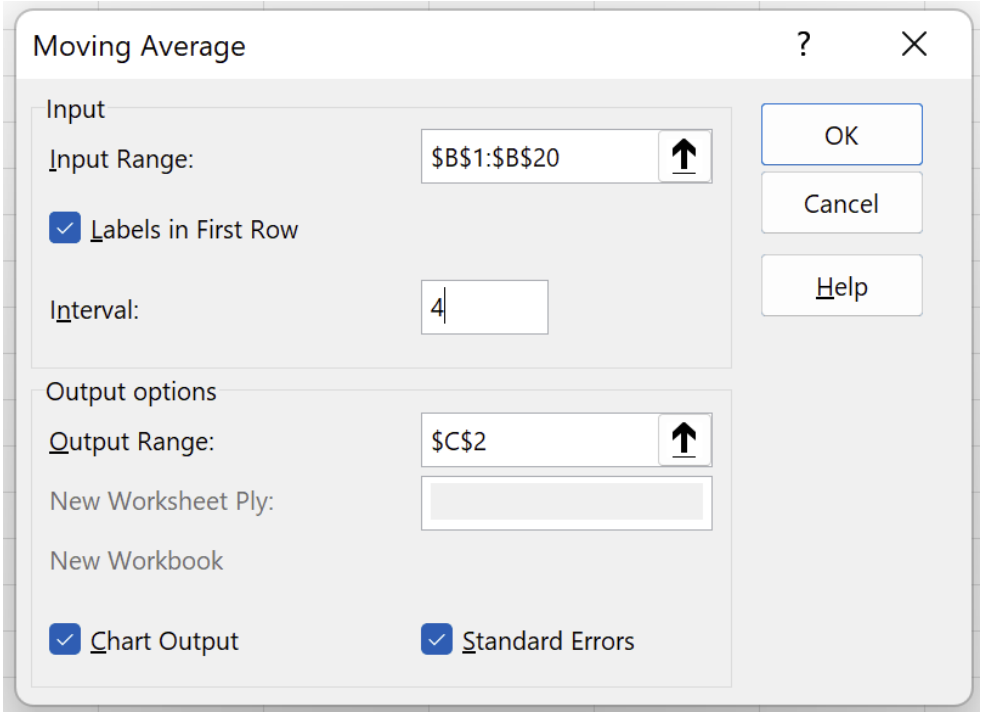


Figure 1-16. The Moving Average tool's dialog box.

When you click OK, the Moving Average tool outputs the moving average values for the data, along with a chart and the standard errors if you selected these options (see Figure 1-17).

Year/Quarter	Sales	Moving Average	Standard Error
2018 Q1	78000	#N/A	#N/A
2018 Q2	60000	#N/A	#N/A
2018 Q3	45000	#N/A	#N/A
2018 Q4	80000	65750	#N/A
2019 Q1	75000	65000	#N/A
2019 Q2	70000	67500	#N/A
2019 Q3	50000	68750	12853.74459
2019 Q4	90000	71250	14224.75835
2020 Q1	80000	72500	13834.96476
2020 Q2	52000	68000	15932.4747
2020 Q3	60000	70500	13910.98936
2020 Q4	115000	76750	21711.53207
2021 Q1	105000	83000	24048.45369
2021 Q2	60000	85000	25895.52326
2021 Q3	80000	90000	25845.99824
2021 Q4	120000	91250	22558.60423
2022 Q1	110000	92500	21551.17456
2022 Q2	63000	93250	23172.58617
2022 Q3	80000	93250	23576.67014

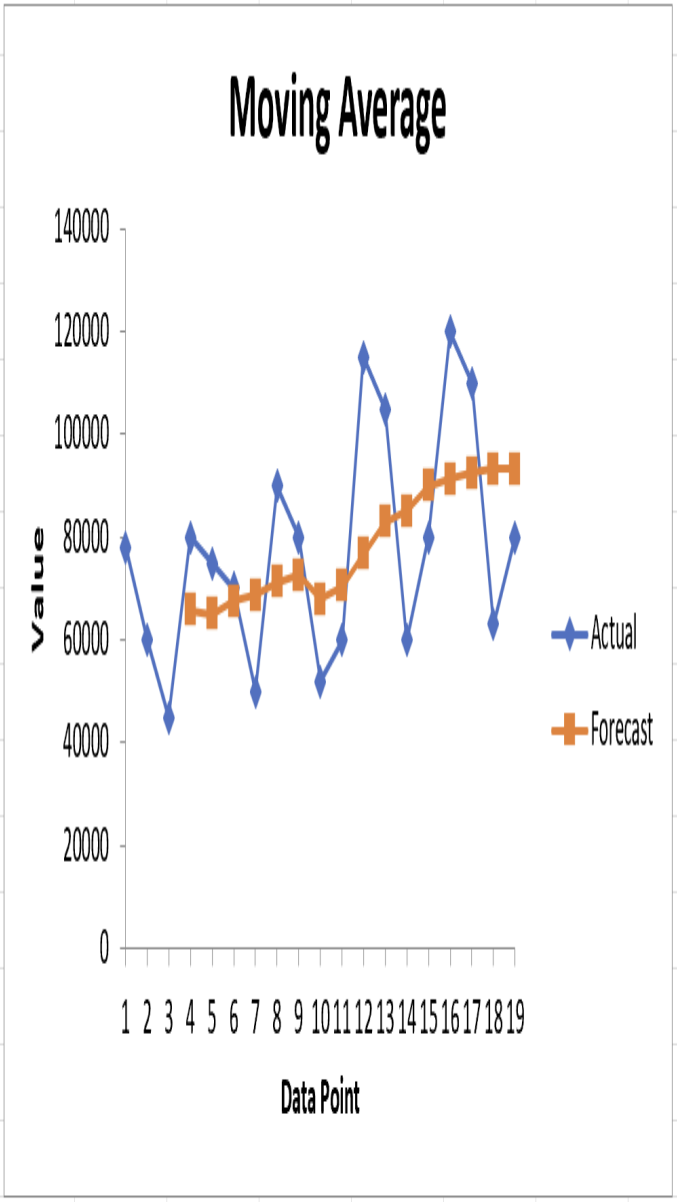


Figure 1-17. The Moving Average tool's output[1].

Discussion

A moving average is a series of averages that “moves” with the data. The interval defines how many data points it uses for these averages. In this recipe, the example uses an interval of 4, which generates the average of four data points: the current data point and the three preceding ones (see Figure 1-18). Using a four-quarter moving average smooths out seasonal variability between quarters because each average contains one data point from each quarter.

	A	B	C
1	Year/Quarter	Sales	Moving Average
2	2018 Q1	78000	#N/A
3	2018 Q2	60000	#N/A
4	2018 Q3	45000	#N/A
5	2018 Q4	80000	=AVERAGE(B2:B5)
6	2019 Q1	75000	=AVERAGE(B3:B6)
19	2022 Q2	63000	=AVERAGE(B16:B19)
20	2022 Q3	80000	=AVERAGE(B17:B20)

Figure 1-18. Formulas generated by the Moving Average tool.

9.7 Using Exponential Smoothing

Problem

You have a time series with irregular peaks and troughs and want to smooth out these irregularities to see the general trend.

Solution

Use the Analysis ToolPak’s Exponential Smoothing tool.

Suppose you have a set of sales data for 12 periods that have irregular peaks and troughs, and you want to analyze the general trend. The range A2:A13 lists the periods, B2:B13 lists the sales, and the first row contains labels (see Figure 1-19).

	A	B	C	D	E	F	G
1	Period	Sales					
2	1	110					
3	2	140					
4	3	250					
5	4	500					
6	5	200					
7	6	350					
8	7	110					
9	8	890					
10	9	400					
11	10	1200					
12	11	450					
13	12	900					

Figure 1-19. Sales data per period.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Exponential Smoothing from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-20).
3. Choose **B1:B13** in the Input Range box. for the range of the sales values, including the label in B1.
4. Type a number between 0 and 1 in the Damping Factor box, for example, **0.7** (see the "Discussion").
5. Place a check in the Labels checkbox because the first row of the input range contains the data's label.
6. Choose an Output Range of **C2** so that the value generated for each row is output next to the original sales data.
7. Place a check in the Chart Output checkbox if you want to generate a chart showing lines for the actual data and the exponential smoothing output (optional).
8. Place a check in the Standard Errors checkbox if you want to generate standard error values (optional). These show the degree of variability between the actual and predicted values.

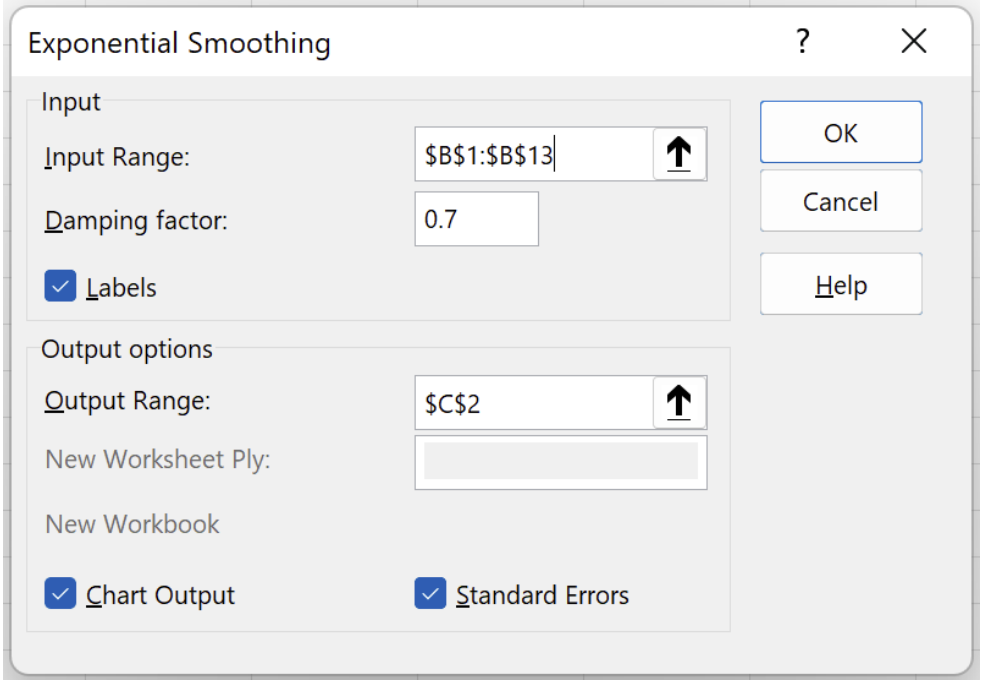


Figure 1-20. The Exponential Smoothing tool's dialog box.

When you click OK, the Exponential Smoothing tool outputs exponential smoothing values for the data, along with a chart and the standard errors if you selected these options (see Figure 1-21).

Period	Sales	Damping Factor 0.7	Standard Error
1	110	#N/A	#N/A
2	140	110	#N/A
3	250	119	#N/A
4	500	158.3	#N/A
5	200	260.81	211.9904793
6	350	242.567	214.1788552
7	110	274.7969	209.7606538
8	890	225.35783	118.8805056
9	400	424.750481	400.1870737
10	1200	417.3253367	395.6091584
11	450	652.1277357	592.9984342
12	900	591.489415	466.9217766

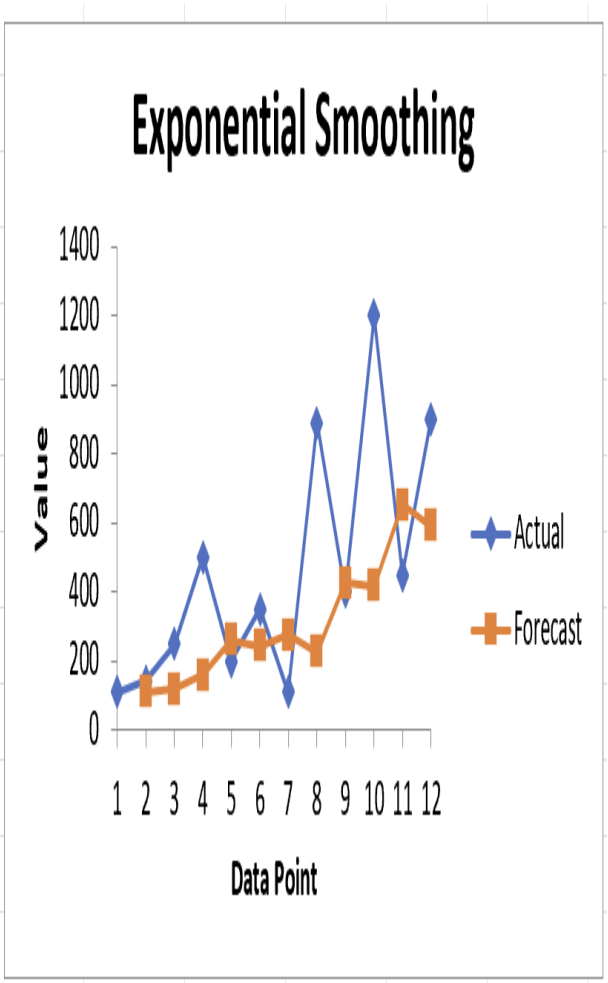


Figure 1-21. The Exponential Smoothing tool's output[2].

Discussion

Exponential smoothing predicts values based on the previous data point and its predicted value.

A damping factor is a number between 0 and 1 that adjusts the weighting between the two values. Using a damping factor of 0.7, for example, multiplies the previous data point by 0.3, multiples its previously predicted value by 0.7, and then returns the sum of these results (see Figure 1-22). Values of 0.7 and 0.8 are reasonable damping factors because they adjust the predicted values by 20% to 30% for error.

	A	B	C
1	Period	Sales	Damping Factor 0.7
2	1	110	#N/A
3	2	140	=B2
4	3	250	=0.3*B3+0.7*C3
5	4	500	=0.3*B4+0.7*C4
12	11	450	=0.3*B11+0.7*C11
13	12	900	=0.3*B12+0.7*C12

Figure 1-22. Formulas generated by the Exponential Smoothing tool.

9.8 Generating a Random Sample

Problem

You have a data set and want to generate a random sample from it.

Solution

Use the Analysis ToolPak's Sampling tool.

Suppose you have a set of customers, and you want to generate a random sample from this data. A1:E201 lists the data: the first column includes a unique number for each customer, and the first row contains labels (see Figure 1-23).

	A	B	C	D	E	F	G
1	Customer	Gender	State	Sales	VIP		
2	1	Female	CA	124887.5	No		
3	2	Female	OR	200.7	No		
4	3	Female	CA	166915.8	No		
5	4	Male	OR	40985.41	No		
6	5	Male	NY	127730.7	No		
7	6	Female	CA	17892.42	No		
8	7	Male	MA	251997.9	No		
9	8	Male	OR	47831.14	No		
10	9	Male	TX	143358.9	No		
11	10	Female	TX	29027.61	No		
200	199	Female	CA	183976.7	No		
201	200	Female	OR	58869.67	No		

Figure 1-23. Customer data.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Sampling from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-24).
3. Choose **A1:A201** in the Input Range box for the range of the unique customer numbers, including the label in A1.
4. Place a check in the Labels checkbox because the first row of the input range contains the data's label.
5. Choose a Sampling Method of Random.
6. Type the sample size you want to generate in the Number of Samples box (for example, **10**).
7. Choose one of the Output Options to specify where you want the Sampling tool to output the results (for example, the Output Range **G2**).

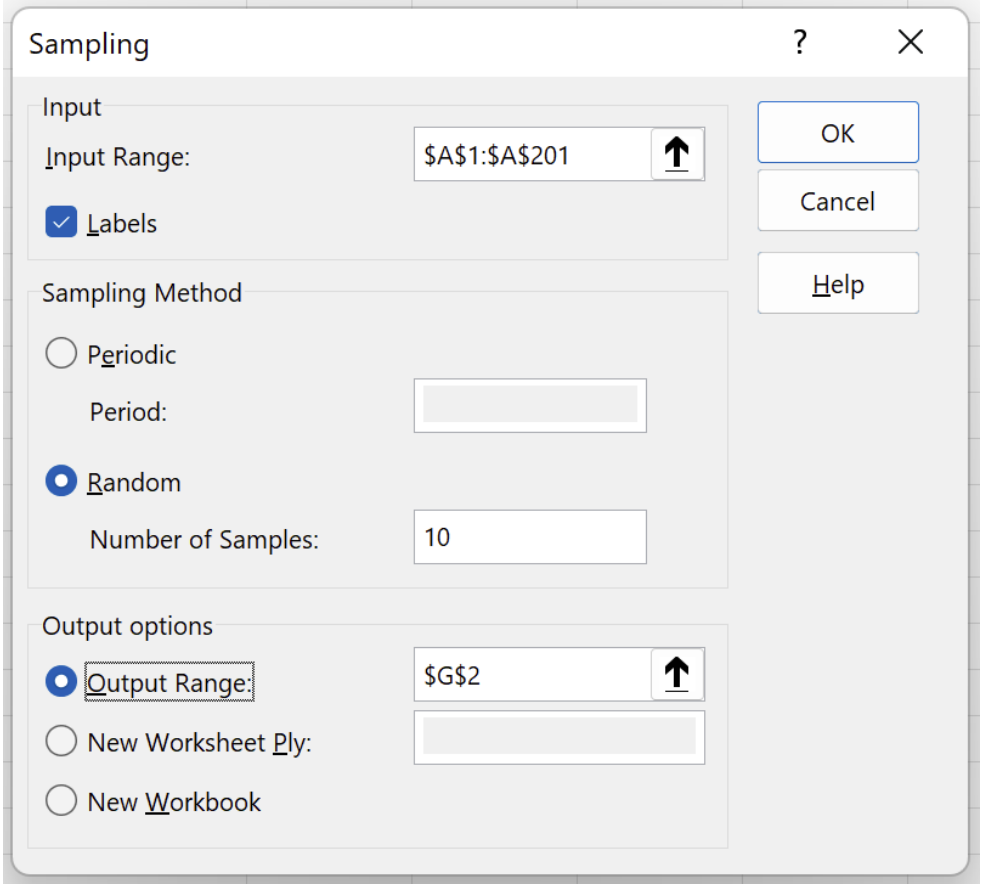


Figure 1-24. The Random Sampling tool's dialog box.

When you click OK, the Sampling tool outputs a random sample (see Figure 1-25).

Customer	Gender	State	Sales	VIP		Random sample
1	Female	CA	124887.5	No		138
2	Female	OR	200.7	No		73
3	Female	CA	166915.8	No		172
4	Male	OR	40985.41	No		6
5	Male	NY	127730.7	No		196
6	Female	CA	17892.42	No		20
7	Male	MA	251997.9	No		75
8	Male	OR	47831.14	No		4
9	Male	TX	143358.9	No		8
10	Female	TX	29027.61	No		176

Figure 1-25. The Random Sampling tool's output[3].

Discussion

This recipe gives you a simple way of generating a simple random sample from a range of values where every value has an equal probability of being selected.

> **WARNING**
>
> Generating a random sample using this may return duplicate values where a participant is selected more than once.

9.9 Generating a Periodic Sample

Problem

You have a data set and want to generate a sample from it by choosing every *n*th item.

Solution

Use the Analysis ToolPak's Sampling tool.

Follow the recipe for "9.8 Generating a Random Sample", but this time choose a Sampling Method of Random and type a value for *n* (for example, **20**) in the Period box (see Figure 1-26). This option tells the Sampling tool to return every *n*th item.

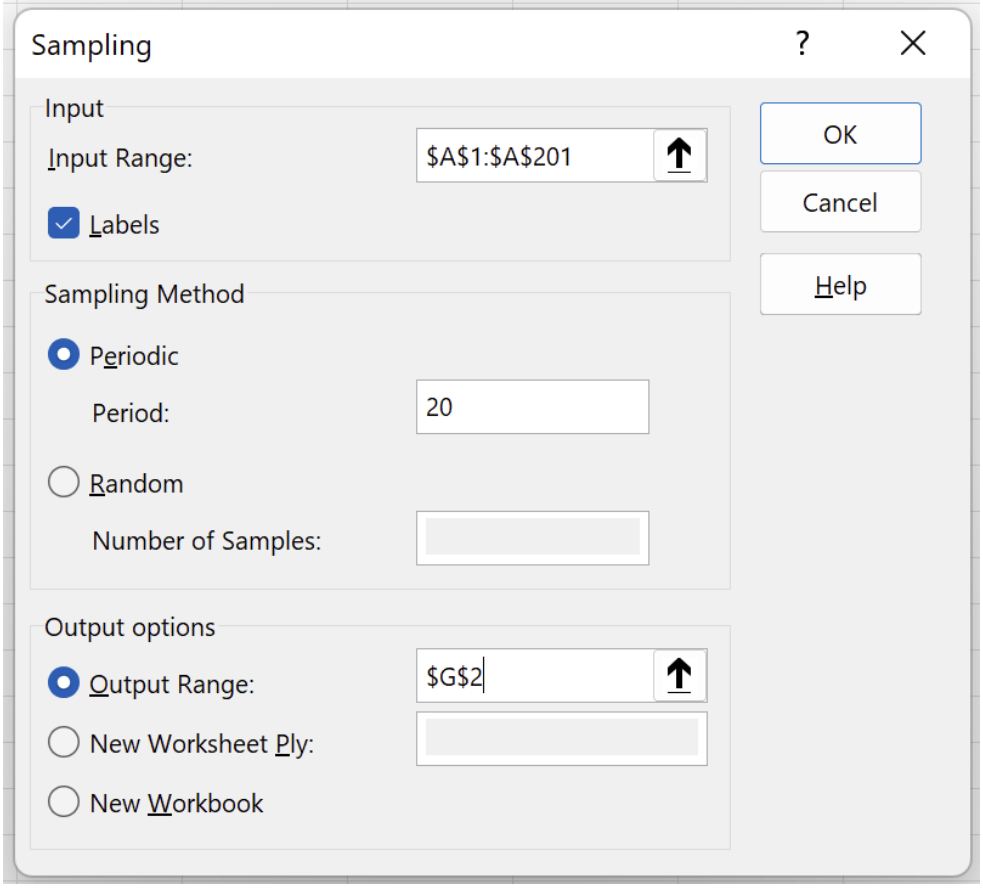


Figure 1-26. The Periodic Sampling tool's dialog box.

When you click OK, the Sampling tool outputs the sample (see Figure 1-27).

Customer	Gender	State	Sales	VIP		Periodic sample
1	Female	CA	124887.5	No		20
2	Female	OR	200.7	No		40
3	Female	CA	166915.8	No		60
4	Male	OR	40985.41	No		80
5	Male	NY	127730.7	No		100
6	Female	CA	17892.42	No		120
7	Male	MA	251997.9	No		140
8	Male	OR	47831.14	No		160
9	Male	TX	143358.9	No		180
10	Female	TX	29027.61	No		200

Figure 1-27. The Periodic Sampling tool's output[4].

Discussion

If a data set is periodic, you can use this recipe to return values from a particular part of the cycle. For quarterly sales figures, for example, you can use a period of 4 to output values from the same quarter each year. Likewise, if you have monthly figures, a period of 12 will return values from the same month.

9.10 Generating Random Numbers Drawn from a Distribution

Problem

You want to create a set of random numbers that follow a specific distribution, such as normal, binomial, or Poisson.

Solution

Use the Analysis ToolPak's Random Number Generation tool.

Suppose you want to generate a set of independent random numbers drawn from a specific distribution, for example, the normal distribution.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Random Number Generation from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-28).
3. Type a number in the Number of Variables box (for example, **1**) to specify how many columns of random numbers you want to return.
4. Type a number in the Number of Random Numbers box (for example, **100**) to specify how many random numbers you want to generate in each column.

5. Choose the type of distribution you want to draw numbers from (for example, Normal).
6. Enter Parameters for the distribution. This section varies depending on the type of distribution you've selected. The normal distribution, for example, has Mean and Standard Deviation parameters, which I've set to **50** and **20** (see Figure 1-28).
7. Choose one of the Output Options to specify where you want the tool to output the random numbers (for example, the Output Range **A1**).

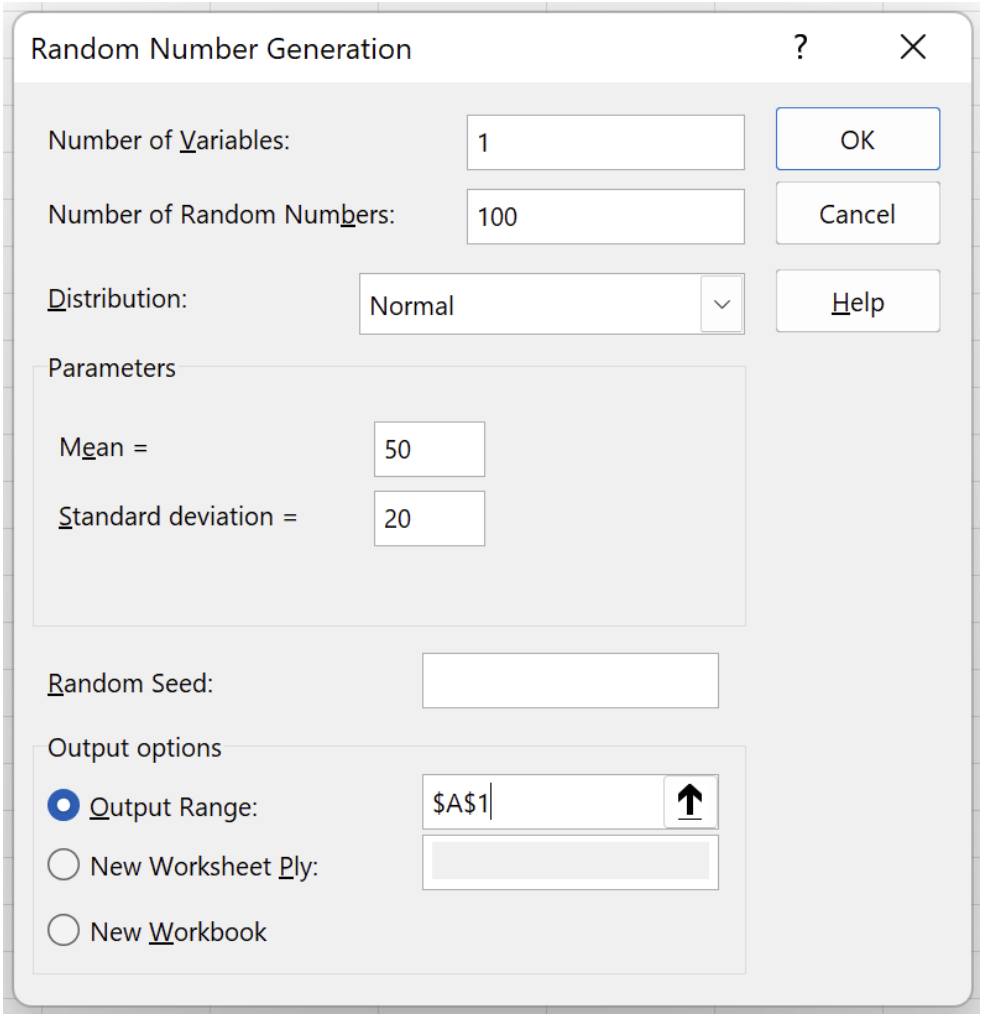


Figure 1-28. The Random Number Generation tool's dialog box.

When you click OK, the Random Number Generation tool outputs independent random numbers drawn from the specified distribution (see Figure 1-29).

	A	B	C	D	E	F
1	48.92775					
2	13.12063					
3	38.37079					
4	48.64876					
5	54.5532					
6	28.78604					
98	81.27097					
99	46.69171					
100	42.75226					

Figure 1-29. The Random Number Generation tool's output.

Discussion

This recipe offers a flexible way of generating numbers from several distributions. The following options are available:

Uniform

Generates random numbers between two limits (that you specify) where each number has an equal chance of being chosen.

Normal

Draws random numbers from a normal distribution with the specified mean and standard deviation.

Bernoulli

Generates 0 or 1 at random using the specified probability of success.

Binomial

Draws random numbers from a binomial distribution using the specified probability of success and number of trials.

Poisson

Draws random numbers from a Poisson distribution using the specified Lambda value (the expected number of occurrences in an interval)

Patterned

Repeats a series of numbers in steps between two limits.

Discrete

Returns values where each value has a specific probability of being chosen. This option requires a two-column input range where the first column holds the values, and the second column holds the corresponding probability. The sum of the probabilities must equal 100%.

> **TIP**
>
> To generate the same random number sequence multiple times, type an integer between 1 and 32,767 in the Random Seed box. This option specifies the starting value for the Random Number Generation tool's algorithm.

9.11 Generating a Correlation Matrix

Problem

You have pairs of measurements and want to determine if there's a linear relationship, its strength, and its direction.

Solution

Use the Analysis ToolPak's Correlation tool.

Suppose you have a set of sales data over 12 months for three products—paper, ink cartridges, and computers—and you want to analyze the relationships between the products. The range A1:D13 lists the data, with the months in column 1 and labels in the first row (see Figure 1-30).

	A	B	C	D
1	Month	Paper	Ink cartridges	Computers
2	Jan	1729	4380	36185
3	Feb	1373	2260	35146
4	Mar	1880	4210	35873
5	Apr	1217	2130	30077
6	May	1640	3768	38508
7	Jun	1990	4779	36751
8	Jul	1402	1995	31200
9	Aug	1751	2665	33457
10	Sep	1784	4465	20453
11	Oct	1271	1650	32926
12	Nov	1229	2335	34058
13	Dec	1184	2640	27999

Figure 1-30. Paper, ink cartridge, and computer sales data.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Correlation from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-31).
3. Choose **B1:B13** in the Input Range box for the range of the sales values, including the labels.
4. Place a check in the Labels checkbox because the first row of each input range contains data labels.
5. Choose one of the Output Options to specify where you want the Descriptive Statistics tool to output the results (for example, the Output Range **E1**).

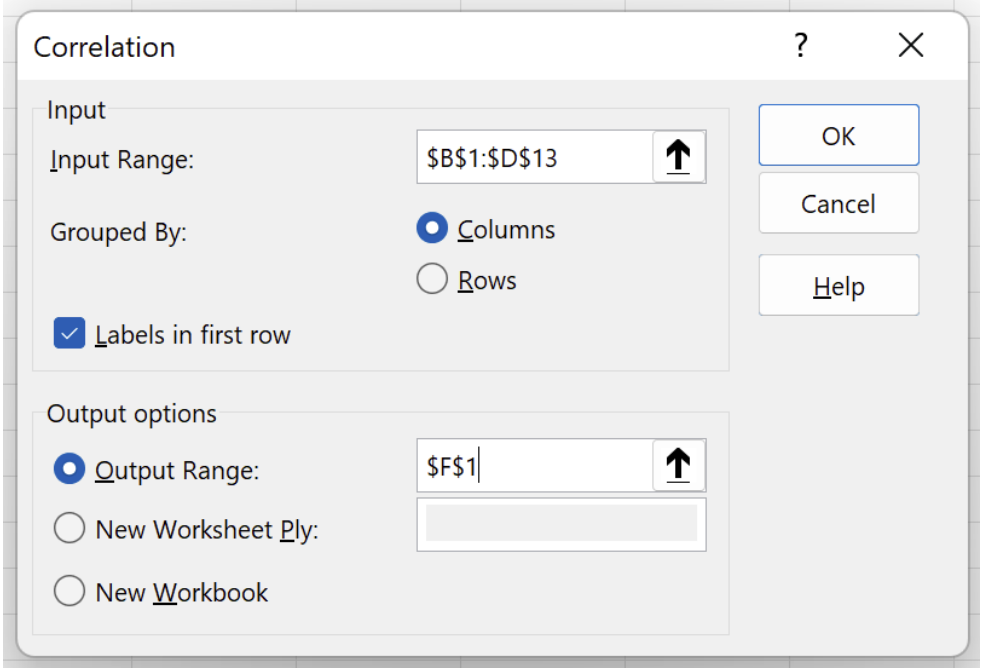


Figure 1-31. The Correlation tool's dialog box.

When you click OK, the Correlation tool outputs a correlation matrix showing the correlation coefficient value for each possible pair of measurements (see Figure 1-32).

	Paper	*Ink cartridges*	*Computers*
Paper	1		
Ink cartridges	0.862268599	1	
Computers	0.203414941	0.075400268	1

Figure 1-32. The Correlation tool's output.

Discussion

The correlation coefficient measures how two sets of values vary together. The closer it is to plus or minus 1, the stronger the relationship, while a value of 0 indicates no linear relationship.

The Correlation tool is handy when you have more than two sets of values and want to find the correlation between each possible pair. The correlation applies the `CORREL` function for each combination.

> **WARNING**
>
> The Correlation tool outputs values instead of formulas, so you'll need to rerun the tool if the underlying data changes. Alternatively, consider using the `CORREL` function.

See also

Correlation and covariance are closely related. See "9.12 Generating a Covariance Matrix".

9.12 Generating a Covariance Matrix

Problem

You have pairs of measurements and want to determine the degree to which they vary together.

Solution

Use the Analysis ToolPak's Covariance tool.

Follow the recipe for "9.11 Generating a Correlation Matrix", but this time choose the Covariance tool (see Figure 1-33). All of the other steps remain the same.

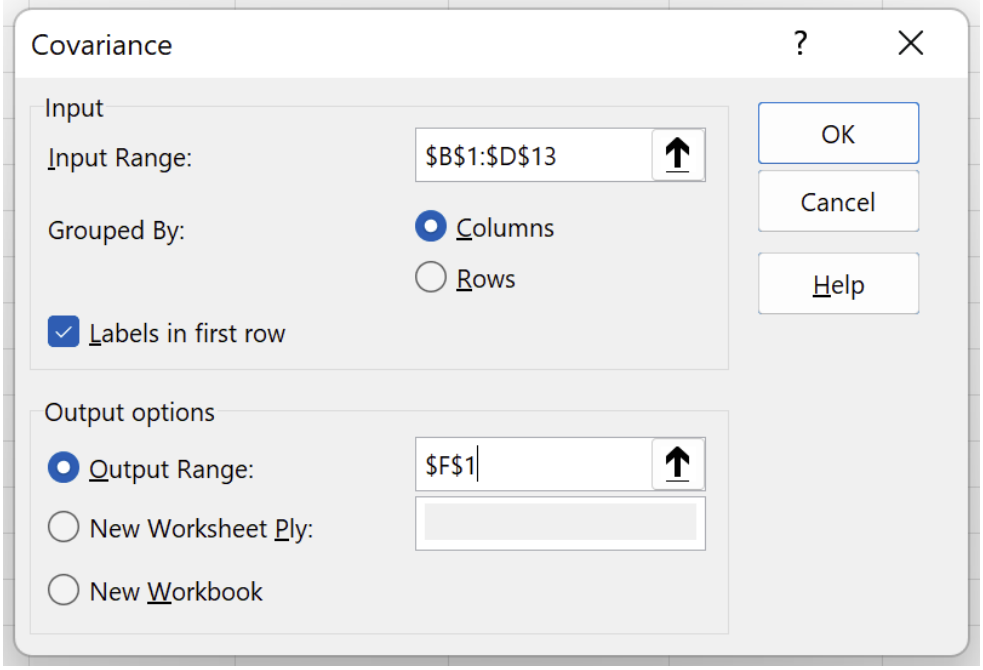


Figure 1-33. The Covariance tool's dialog box.

When you click OK, the Covariance tool outputs a covariance matrix showing the covariance value for each possible pair of measurements (see Figure 1-34).

	Paper	*Ink cartridges*	*Computers*
Paper	76240.25		
Ink cartridges	256745.2083	1162880.91	
Computers	262453.9583	379943.1597	21835153.24

Figure 1-34. The Covariance tool's output.

Discussion

Both Correlation and covariance measure how two sets of values vary together. The main difference is correlation coefficients are scaled to lie between plus or minus 1, while covariances are unscaled.

Suppose you have measurements that include the weight in pounds. Then, if you convert the weight to kilograms, the correlation coefficient will stay the same, while the covariance will change.

> **WARNING**
>
> The Covariance tool outputs values instead of formulas, so you'll need to rerun the tool if the underlying data changes. Alternatively, consider using the `COVARIANCE.P` function.

9.13 Performing a Linear Regression Analysis

Problem

You have a range of data and want to analyze how one or more other ranges affect it. You also want to find a linear equation for the relationship you can use to make predictions.

Solution

Use the Analysis ToolPak's Regression tool.

Suppose you have data showing the number of items sold, product price, and the amount spent on advertising for different periods. You want to use this to predict how many items you'll sell, depending on the price and advertising costs. The range A2:A8 lists the number of items sold, B2:B8 lists the product price, C2:C8 C1:C11 lists the advertising cost, and the first row contains labels (see Figure 1-35).

	A	B	C
1	Sold	Price	Advertising
2	10000	4	3500
3	5000	10	500
4	7000	6	1000
5	9000	4	1500
6	8000	10	5000
7	8500	6	4000
8	5000	8	2000

Figure 1-35. The number of items sold, product price, and advertising data.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Regression from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-36).
3. Choose **A1:A8** in the Input Y Range box for the range listing the number of items sold.
4. Choose **B1:C8** in the Input X Range box for the range listing the price and advertising amount.
5. Choose the Columns option in the Grouped By section because the values are listed in columns.
6. Place a check in the Labels in First Row checkbox because the first row of the input ranges contains data labels.
7. Choose one of the Output Options to specify where you want the tool to output the results (for example, the Output Range **F1**).

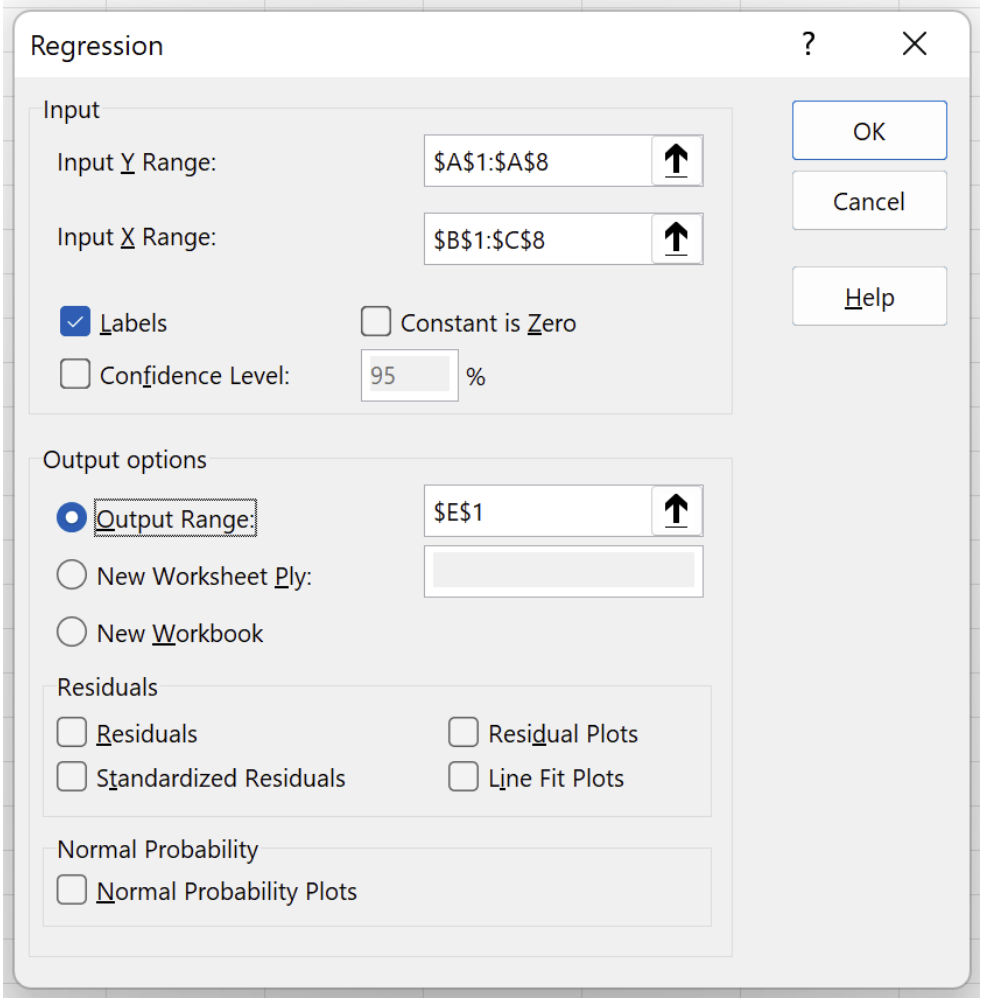


Figure 1-36. The Regression tool's dialog box.

When you click OK, the Regression tool outputs three tables.

The first table shows regression statistics (see Figure 1-37). It includes an R Square statistic which indicates how well the analysis fits the data—the closer this value is to 1, the better the fit. In this example, R Square is 0.838, which means it's a good fit: the price and advertising costs explain 83.8% of the variation in the items sold.

Regression Statistics	
Multiple R	0.915523108
R Square	0.838182562
Adjusted R Square	0.757273843
Standard Error	954.0561244
Observations	7

Figure 1-37. Statistics generated by the Regression tool, including R Square.

The ANOVA table includes a Significance F statistic (see Figure 1-38). This statistic shows you the statistical significance of the results (see "Discussion").

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	18859107.65	9429553.823	10.35960738	0.026184883
Residual	4	3640892.354	910223.0886		
Total	6	22500000			

Figure 1-38. The Regression tool's ANOVA output including Significance F.

The third table shows the regression analysis results (Figure 1-39). Here, the coefficients describe the straight line that best fits the data, which you can use to make predictions. In this example, the equation for the line is y (items sold) = 9660.3855 - 557.559 * price + 0.6651505 * advertising.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	9660.3855	1233.280747	7.8331	0.0014	6236.2492	13084.52
Price	-557.559	153.1677974	-3.64	0.022	-982.821	-132.297
Advertising	0.6651505	0.231567943	2.8724	0.0454	0.0222148	1.308086

Figure 1-39. The Regression Analysis output including coefficients.

Discussion

The Regression tool performs linear regression analysis using the least squares method. Behind the scenes, it uses the `LINEST` function to fit a straight line through the points in the input ranges you provide. The Input Y Range box specifies the dependent variable

you want to be able to predict, and the Input X Range box specifies the independent variables you want to use in this prediction.

You generally want the Significance F statistic in the ANOVA table to be less than 0.05. If it's greater than 0.05, the results may be unreliable; try rerunning the tool, this time leaving out the variable with the highest P-value in the results table (see Figure 1-39).

The Regression dialog box includes the following additional options:

Constant is Zero

Forces the regression line to pass through the origin so that when the X values are 0, the Y value is also 0.

Confidence Level

The confidence level for the regression analysis.

Residuals

Generates residuals: the differences between the actual and predicted data points.

Normal Probability

Generates an extra table showing probabilities.

> **WARNING**
>
> The Regression tool outputs values instead of formulas, so you'll need to rerun the tool if the underlying data changes.

9.14 Performing a Two-Sample t-Test

Problem

You have two samples and want to compare the means of the populations they're drawn from when you don't know the population variances.

Solution

Use the t-Test: Two-Sample Assuming Equal Variances tool if you believe the populations have the same variances or the t-Test: Two-Sample Assuming Unequal Variances tool if you think the variances are different.

Suppose you want to test whether two different product brands have the same mean weight, and you have a sample of weights for each brand. The range A2:A16 lists the weights for Brand A, B2:B16 lists the weights for Brand B, and the first row contains labels (see Figure 1-40).

	A	B
1	Brand A	Brand B
2	492.06	494.789
3	516.981	509.706
4	499.622	495.298
5	501.352	497.619
6	499.065	501.196
7	503.103	497.548
8	506.986	490.797
9	491.529	489.223
10	507.399	501.161
11	501.367	490.019
12	503.963	494.748
13	503.778	501.11
14	496.074	505.233
15	502.516	492.631
16	502.203	497.099

Figure 1-40. Weight samples for the two brands.

To solve this problem, you test whether the mean difference between the two populations (the weights for Brand A and Brand B) is 0.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. From the list of analysis tools, select t-Test: Two-Sample Assuming Equal Variances or t-Test: Two-Sample Assuming Unequal Variances (this example uses the Equal Variances option). Then click OK to open the tool's dialog box (see Figure 1-41).
3. Choose **A1:A16** in the Variable 1 Range box for the range for Brand A.

4. Choose **B1:B16** in the Variable 1 Range box for the range for Brand B.
5. Type a value for the Hypothesized Mean Difference. This example uses **0** to test whether the two population means are equal.
6. Place a check in the Labels checkbox because the first row of the variable ranges contains data labels.
7. Type a value for Alpha, the significance level, for example, **0.05**.
8. Choose one of the Output Options to specify where you want the tool to output the results (for example, the Output Range **D1**).

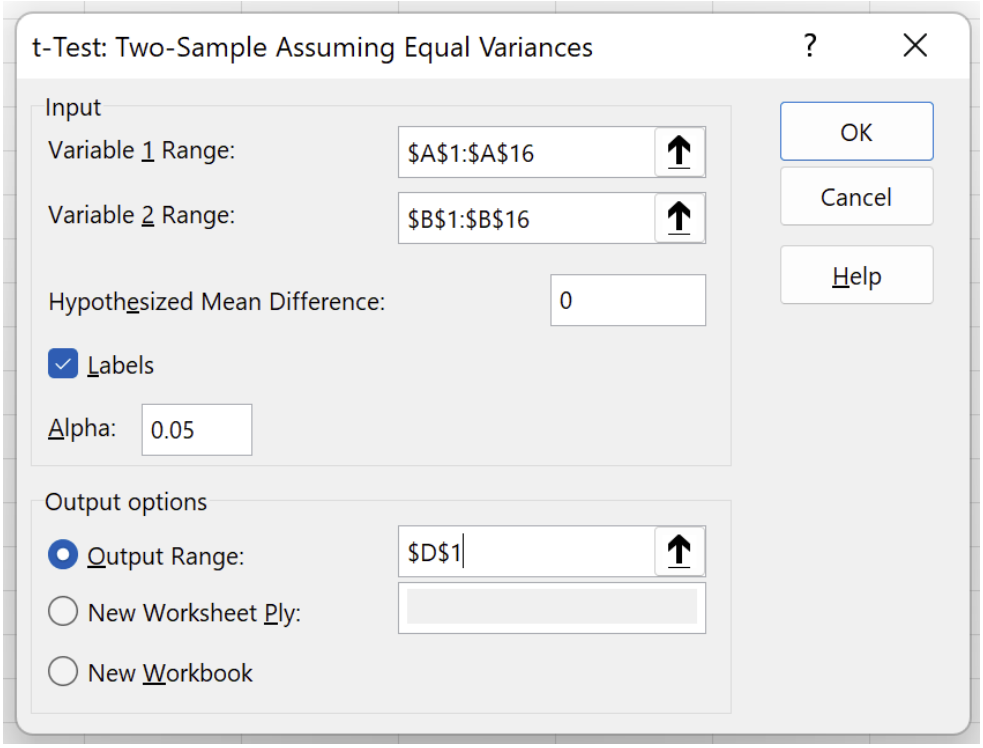


Figure 1-41. The t-Test: Two-Sample Assuming Equal Variances tool's dialog box.

When you click OK, the tool outputs the results (see Figure 1-42).

t-Test: Two-Sample Assuming Equal Variances		
	Brand A	*Brand B*
Mean	501.8665333	497.2118
Variance	38.95456455	32.90684489
Observations	15	15
Pooled Variance	35.93070472	
Hypothesized Mean Difference	0	
df	28	
t Stat	2.126633102	
P(T<=t) one-tail	0.021199796	
t Critical one-tail	1.701130934	
P(T<=t) two-tail	0.042399592	
t Critical two-tail	2.048407142	

Figure 1-42. The t-Test: Two-Sample Assuming Equal Variances tool's output.

Discussion

This recipe is helpful when you want to find the likelihood that two samples came from populations with equal means. It tests the null hypothesis that the means are the same (or have the hypothesized difference entered in step 5) against the alternate hypothesis that they're not.

The most important statistics generated by the tool are the p-values for the one-tailed and two-tailed tests. P(T<=t) one-tail gives the p-value for a one-tailed test: a test that detects differences in the population means in one direction (for example, the mean for Brand A is greater than the mean for Brand B). On the other hand, P(T<=t) two-tail gives the p-value for a two-tailed test used to detect differences in either direction (for example, the mean for Brand A is greater than or less than the mean for Brand B).

In general, decide whether you need a one-tailed or two-tailed test, and then see if its p-value is less than the significance level (Alpha) you chose in step 7. If the p-value is smaller than the significance

level, there's enough evidence to reject the null hypothesis. In the example used in this recipe, both p-values are less than the significance level (0.05), so there's enough evidence to conclude that the mean weight of Brand A is different from the mean weight of Brand B.

> **WARNING**
>
> The t-Test tools output values instead of formulas, so you'll need to rerun the tool if the underlying data changes.

See also

This recipe is for samples where you don't know the value of the population variances. If you know the variances, use "9.15 Performing a Two-Sample z-Test" instead.

If you have more than two samples, use "9.18 Performing a One-Way ANOVA Test".

9.15 Performing a Two-Sample z-Test

Problem

You have two samples and want to compare the means of the populations they're drawn from when you know the population variances.

Solution

Use the Analysis ToolPak's z-Test: Two Sample for Means option.

Suppose you want to test whether two different product brands have the same mean weight, and you have a sample of weights for each brand. The range A2:A16 lists the weights for Brand A, B2:B16 lists

the weights for Brand B, and the first row contains labels (see Figure 1-40). Furthermore, Brand A has a variance of 40, and Brand B has a variance of 31.

Follow the recipe for "9.14 Performing a Two-Sample t-Test", but this time choose the z-Test: Two Sample for Means tool and include these extra steps:

1. Type **40** in the Variable 1 Variance (known) box: this is the variance for Brand A (see Figure 1-43).
2. Type **31** in the Variable 2 Variance (known) box: this is the variance for Brand B.

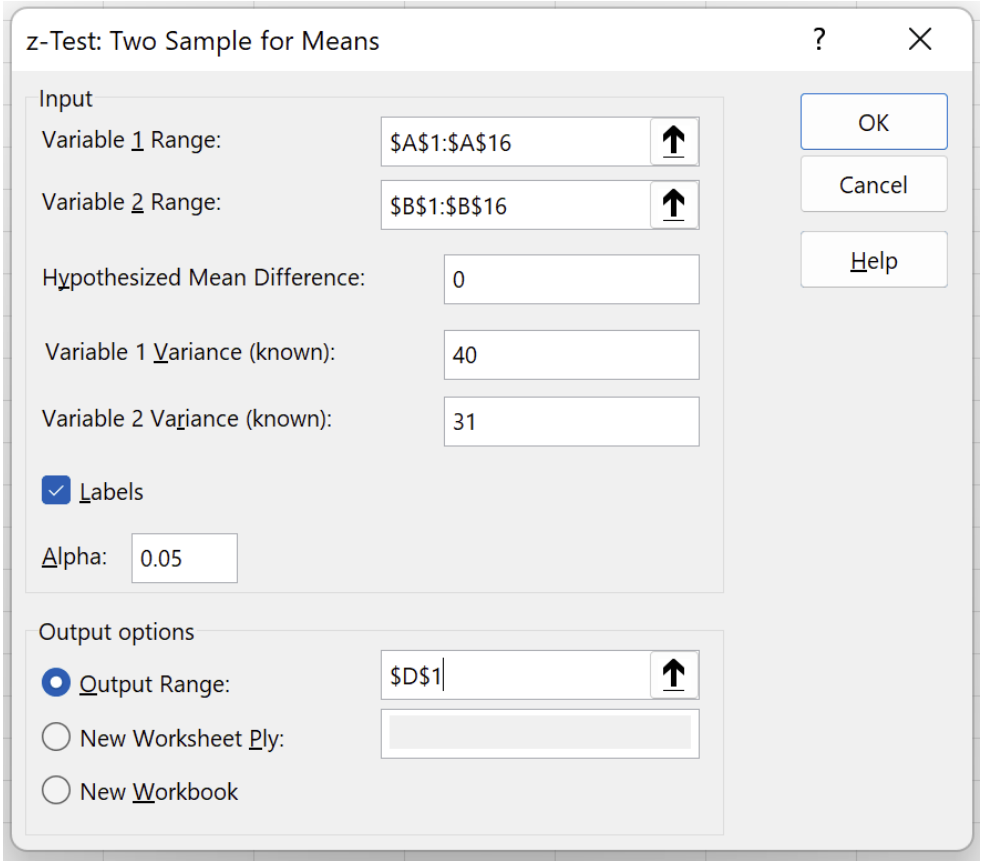


Figure 1-43. The z-Test: Two Sample for Means tool's dialog box.

When you click OK, the tool outputs the results (see Figure 1-44).

z-Test: Two Sample for Means		
	Brand A	*Brand B*
Mean	501.8665333	497.2118
Known Variance	40	31
Observations	15	15
Hypothesized Mean Difference	0	
z	2.139494925	
P(Z<=z) one-tail	0.016197803	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	0.032395606	
z Critical two-tail	1.959963985	

Figure 1-44. The z-Test: Two Sample for Means tool's output.

Discussion

This recipe is handy when you want to find the likelihood that two samples came from populations with equal means and you know the value of the population variances. It tests the null hypothesis that the means are the same—or there's a specified difference between them—against the alternate hypothesis that they're not.

The tool outputs p-values—P(Z<=z)--for one-tailed and two-tailed tests. The discussion for "9.14 Performing a Two-Sample t-Test" explains these tests.

> **WARNING**
>
> The z-Test: Two Sample for Means tool outputs values instead of formulas, so you'll need to rerun the tool if the underlying data changes.

9.16 Performing a Paired Two-Sample t-Test

Problem

You have a sample composed of pairs of measurements and want to test the mean difference between the measurements.

Solution

Use the Analysis ToolPak's t-Test: Paired Two Sample for Means option.

Suppose you want to test whether a boot camp for athletes makes a difference in how fast they can run. You have a sample of athletes, and you've recorded how long it takes each one to sprint down a track before and after the boot camp. The range A2:A16 lists the ID of each athlete, B2:B16 lists the times recorded before the boot camp, C2:C16 lists the times recorded after the boot camp, and the first row contains labels (see Figure 1-45).

	A	B	C
1	Athlete	Before	After
2	1	99.6	96.25
3	2	94.67	82.44
4	3	83.8	82.96
5	4	106.54	105.98
6	5	104.41	102.61
7	6	107.53	95.02
8	7	104.18	92.57
9	8	95.92	102.75
10	9	116.58	96.94
11	10	103.71	85.04
12	11	86.61	86.7
13	12	97.22	94.4
14	13	98.41	89.59
15	14	97.11	86.54
16	15	121.48	100.69

Figure 1-45. Pairs of measurements for the athletes.

To solve this problem, you test whether the mean difference between the two times (recorded before and after the boot camp) is 0.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select t-Test: Paired Two Sample for Means from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-46).
3. Choose **B1:B16** in the Variable 1 Range box for the range for the times recorded before the boot camp.
4. Choose **C1:C16** in the Variable 1 Range box for the range for the times recorded after the boot camp.
5. Type a value for the Hypothesized Mean Difference. This example uses **0** to test whether the two means are equal.
6. Place a check in the Labels checkbox because the first row of the variable ranges contains data labels.
7. Type a value for Alpha, the significance level, for example, **0.05**.
8. Choose one of the Output Options to specify where you want the tool to output the results (for example, the Output Range **E1**).

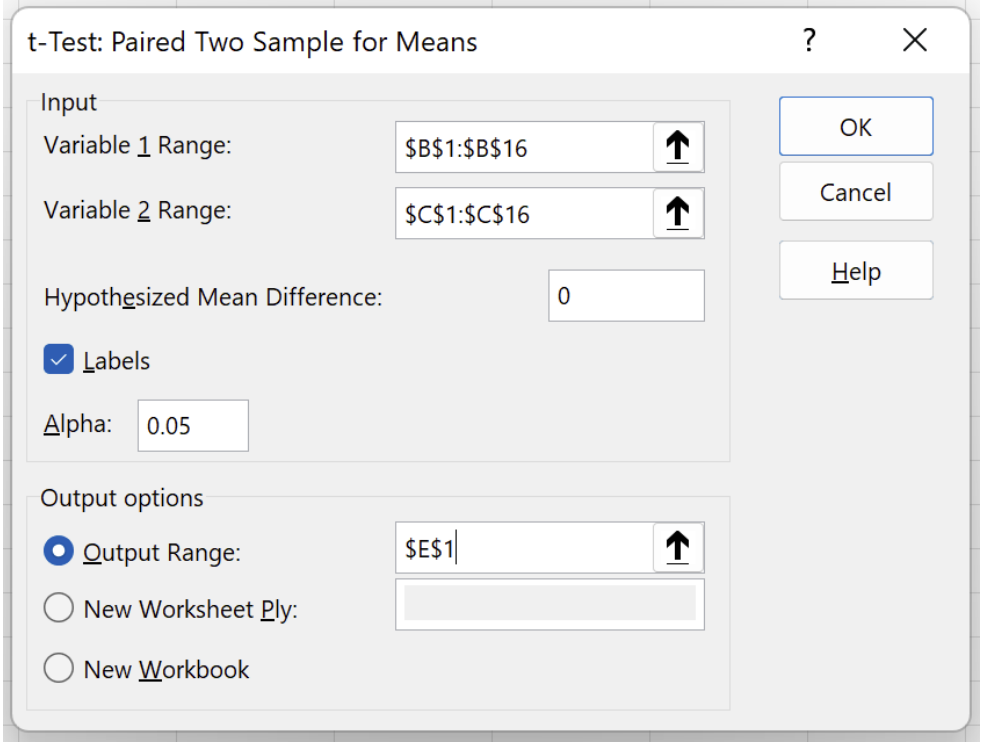


Figure 1-46. The t-Test: Paired Two Sample for Means tool's dialog box.

When you click OK, the tool outputs the results (see Figure 1-47).

t-Test: Paired Two Sample for Means

	Before	*After*
Mean	101.1846667	93.36533333
Variance	97.92488381	58.14764095
Observations	15	15
Pearson Correlation	0.584586371	
Hypothesized Mean Difference	0	
df	14	
t Stat	3.676614457	
P(T<=t) one-tail	0.001245007	
t Critical one-tail	1.761310136	
P(T<=t) two-tail	0.002490013	
t Critical two-tail	2.144786688	

Figure 1-47. The t-Test: Paired Two Sample for Means tool's output.

Discussion

This recipe is a helpful way of testing whether there's a significant difference between two measurements taken from the same item or individual. Examples of when you might want to use this test include:

- Comparing measurements before and after a treatment
- Comparing the speeds of two car models where each individual gets to drive each car
- Comparing the effects of two skin lotions where one is applied to each individual's left arm and the other to their right.

The most important statistics generated by the tool are the p-values for the one-tailed and two-tailed tests. See "Discussion" for an explanation of these statistics.

9.17 Performing a Two-Sample F-Test for Variances

Problem

You have two samples and want to compare the variances of the populations they're drawn from.

Solution

Use the F-Test Two-Sample for Variances tool.

Suppose you want to test whether the scores for two teams have the same variability, and you have a sample of scores for each team. The range A2:A16 lists the scores for Team A, B2:B16 lists the scores for Team B, and the first row contains labels (see Figure 1-48).

	A	B
1	Team A	Team B
2	54.85	57.42
3	62.2	66.91
4	64.48	71.78
5	57.06	52.09
6	62.67	49.44
7	68.74	67.31
8	55.04	64.46
9	59.38	49.53
10	73.79	71.51
11	57.77	68.24
12	56.88	66.49
13	61.83	57.56
14	59.86	51.12
15	65.36	53.47
16	49.31	40.48

Figure 1-48. Score samples for the two teams.

To solve this problem, you test whether each population (the scores for Team A and Team B) has the same variance.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select F-Test Two-Sample for Variances from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-49).
3. Choose **A1:A16** in the Variable 1 Range box for the range for Team A.
4. Choose **B1:B16** in the Variable 1 Range box for the range for Team B.

5. Place a check in the Labels checkbox because the first row of the variable ranges contains data labels.
6. Type a value for Alpha, the significance level, for example, **0.05**.
7. Choose one of the Output Options to specify where you want the tool to output the results (for example, the Output Range **D1**).

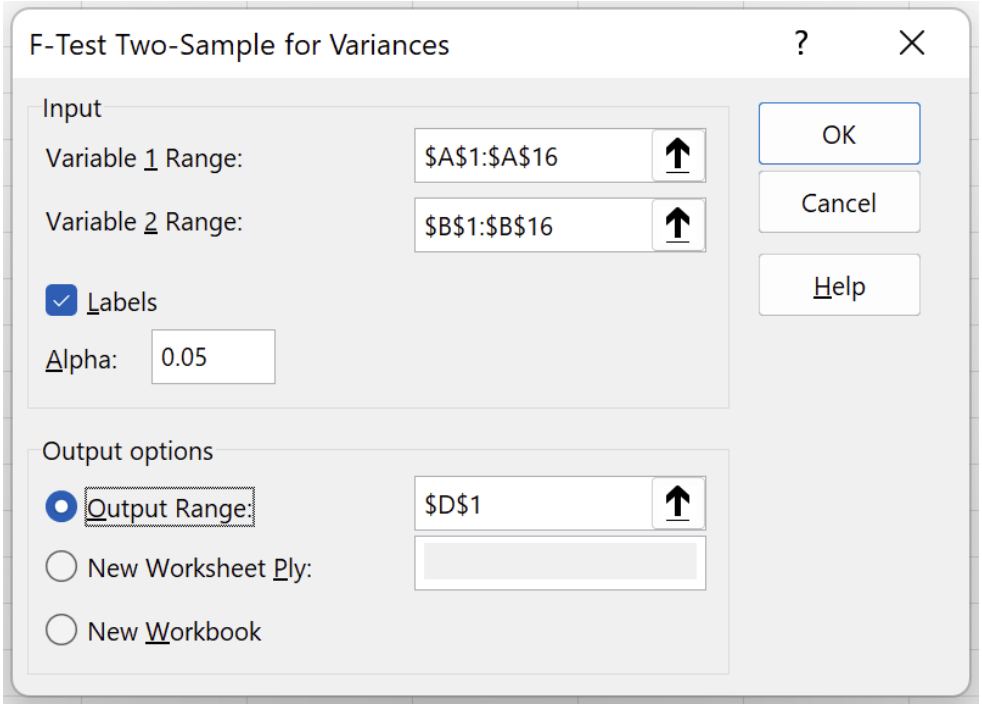


Figure 1-49. The F-Test Two-Sample for Variances tool's dialog box.

When you click OK, the tool outputs the results (see Figure 1-50).

F-Test Two-Sample for Variances		
	Team A	*Team B*
Mean	60.61466667	59.18733333
Variance	36.72552667	92.23384952
Observations	15	15
df	14	14
F	0.3981784	
P(F<=f) one-tail	0.048029033	
F Critical one-tail	0.402620943	

Figure 1-50. The F-Test Two-Sample for Variances tool's output.

Discussion

This recipe helps you find the likelihood that two samples came from populations with equal variances. It tests the null hypothesis that the variances are the same against the alternate hypothesis that they're not.

The most important statistic generated by the tool is P(F<=f) one-tail, which is the p-value for a one-tailed test. If the p-value is less than the significance level (Alpha) you chose in step 6, there's sufficient evidence to reject the null hypothesis.

> **WARNING**
>
> The F-Test Two-Sample for Variances tool outputs values instead of formulas, so you'll need to rerun the tool if the underlying data changes.

9.18 Performing a One-Way ANOVA Test

Problem

You have two or more samples that depend on a single factor and want to test whether they're drawn from populations with the same means.

Solution

Use the ANOVA: Single Factor tool.

Suppose you want to test whether three different product brands have the same mean weight, and you have a sample of weights for each brand. The range A2:A11 lists the weights for Brand A, B2:B11 lists the weights for Brand B, C2:C11 lists the weights for Brand C, and the first row contains labels (see Figure 1-51).

	A	B	C
1	Brand A	Brand B	Brand C
2	489.57	497.38	501.15
3	507.8	499	477.48
4	489.29	505.65	483.16
5	519.48	509.09	469.45
6	496.48	505.4	499.93
7	511.76	498.62	474.87
8	485	497.96	480.93
9	514.68	477.68	496.52
10	496.5	496.78	476.21
11	503.23	513.32	504.78

Figure 1-51. Weight samples for the three brands.

Since the weights depend on a single factor—the product brand—you can use a one-way ANOVA test to assess whether the samples are drawn from populations with the same means.

1. Go to the Data menu and choose Data Analysis from the Analyze group.

2. Select ANOVA: Single Factor from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-52).
3. Choose **A1:C11** in the Input Range box for the range of weights for the three samples, including their label.
4. Choose the Columns option in the Grouped By section because the weights are listed in columns.
5. Place a check in the Labels in First Row checkbox because the first row of the input range contains data labels.
6. Type a value for Alpha, the significance level, for example, **0.05**.
7. Choose one of the Output Options to specify where you want the tool to output the results (for example, the Output Range **E1**).

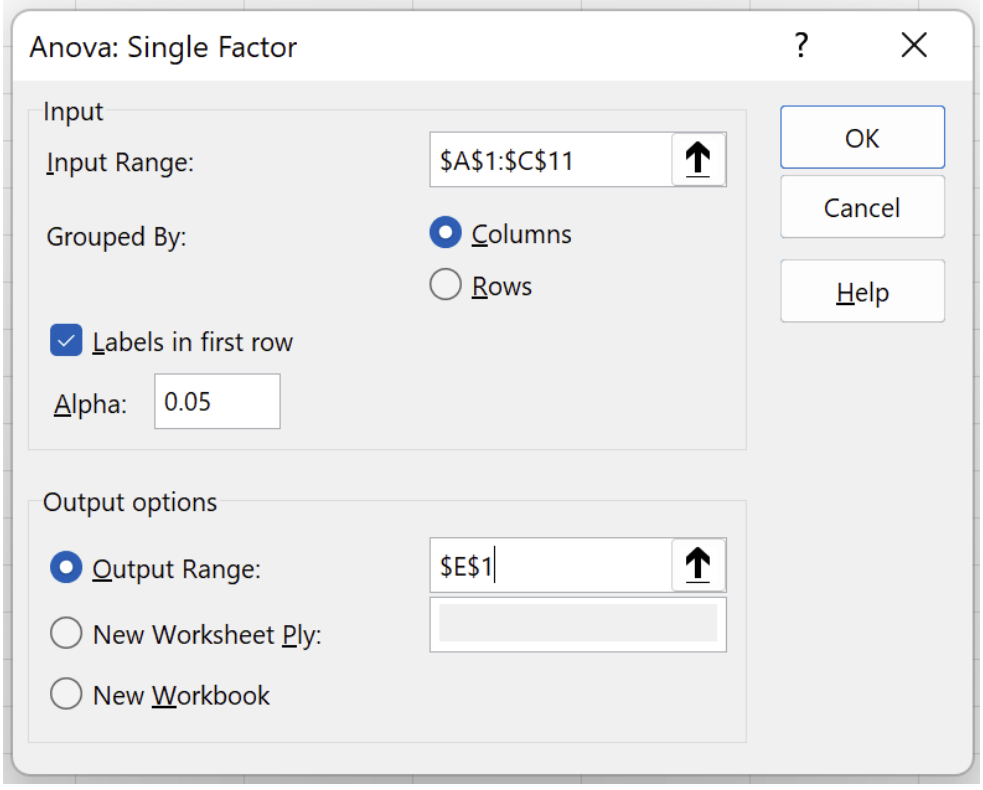


Figure 1-52. The ANOVA: Single Factor tool's dialog box.

When you click OK, the tool outputs the results (see Figure 1-53).

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Brand A	10	5013.79	501.379	139.8506989
Brand B	10	5000.88	500.088	93.72630667
Brand C	10	4864.48	486.448	164.9892844

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1368.836807	2	684.4184033	5.151602786	0.012730274	3.354130829
Within Groups	3587.09661	27	132.85543			
Total	4955.933417	29				

Figure 1-53. The ANOVA: Single Factor tool's output.

Discussion

This recipe uses multiple samples to test whether the means of at least two of the populations they're drawn from have different means. It tests the null hypothesis that the means are the same against the alternate hypothesis that they're not.

The Analysis ToolPak includes three separate ANOVA tools. This recipe uses the ANOVA: Single Factor tool, where the sample measurements (in this case, the weights) depend on a single factor

(the product brand). It's a more general form of "9.14 Performing a Two-Sample t-Test" except that you can have more than two samples.

The most important statistic in the tool's output is the p-value. If this statistic is less than the significance level (Alpha) you chose in step 6, it means there's enough evidence to reject the null hypothesis that the samples are drawn from populations with the same means. In the example used in this recipe, the p-value (0.0127) is less than the significance level (0.05), so there's enough evidence to conclude that the mean weight of at least one brand differs from the others.

9.19 Performing a Two-Way ANOVA Test

Problem

You have a sample of measurements and want to test whether they're influenced by two factors, either individually or together.

Solution

Use the ANOVA: Two-Factor with Replication tool.

Suppose you want to test whether the starter culture brand, the yogurt maker model, or some interaction between the two influences the amount of time it takes to make yogurt. You have recorded the time taken in hours using three yogurt maker models and three brands of starter culture and repeated this four times for each combination. The range A2:D13 lists the results, with the yogurt maker models arranged in columns, the starter cultures arranged in rows, and data labels in the first row and column (see Figure 1-54).

	A	B	C	D
1		Model A	Model B	Model C
2	Culture 1	8.3	7.3	8.2
3		7.1	6.4	9.4
4		7.8	8.5	9.9
5		6.8	7.9	7.9
6	Culture 2	8.4	8.2	9.2
7		6.7	7.6	8.1
8		6.2	8	8.9
9		7.4	6.9	8.5
10	Culture 3	10.1	10.5	10.1
11		10.7	9.5	10.7
12		9.8	9.1	10.1
13		9.6	10.2	9.8

Figure 1-54. Yogurt maker and starter culture times.

Since the time taken depends on two factors—the starter culture brand and the yogurt maker model—and there are multiple times for each combination, you can use a two-way Analysis of Variance (ANOVA) test with replication to analyze the results.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select ANOVA: Two-Factor with Replication from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-55).
3. Choose **A1:D13** in the Input Range box for the entire range of the data, including labels.
4. Type **4** in the Rows per sample box because there are four recorded times for each combination, listed in separate rows.
5. Type a value for Alpha, the significance level, for example, **0.05**.
6. Choose one of the Output Options to specify where you want the tool to output the results (for example, the Output Range **F1**).

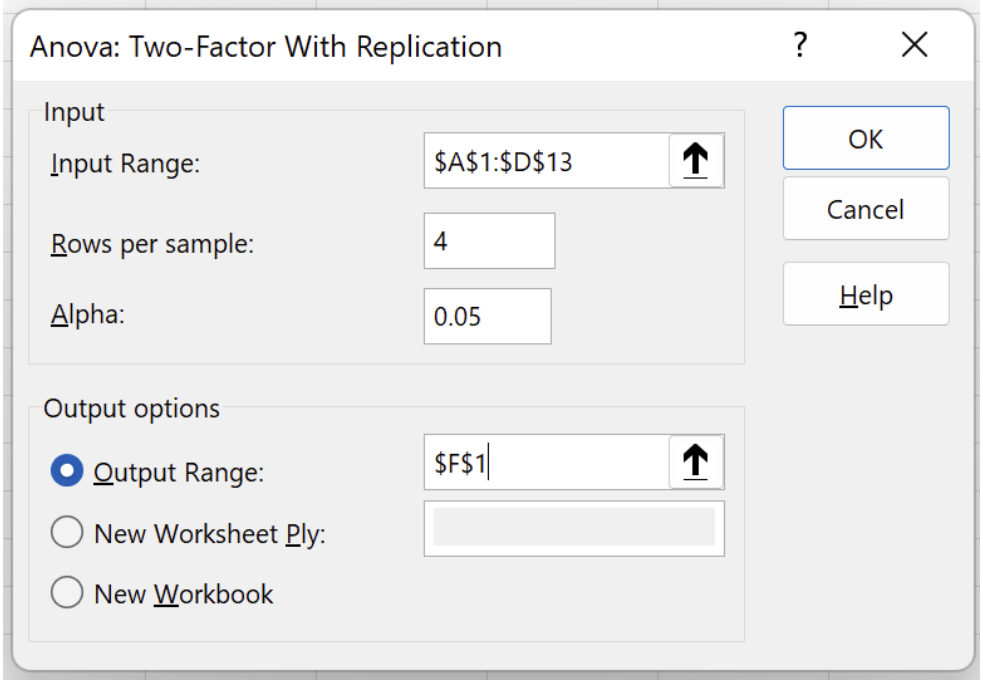


Figure 1-55. The ANOVA: Two-Factor with Replication tool's dialog box.

When you click OK, the tool outputs the results in several tables. The most important results are in the ANOVA table (see Figure 1-56).

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	35.92388889	2	17.96194444	36.56030908	2.07178E-08	3.354130829
Columns	7.153888889	2	3.576944444	7.280625707	0.002958792	3.354130829
Interaction	2.536111111	4	0.634027778	1.290520166	0.298457563	2.727765306
Within	13.265	27	0.491296296			
Total	58.87888889	35				

Figure 1-56. The ANOVA: Two-Factor with Replication tool's output.

Discussion

This recipe uses the ANOVA: Two-Factor with Replication tool where the sample measurements (in this case, the time taken) depend on two factors (the starter culture brand and the yogurt maker model). The term *with replication* in the tool's name means multiple data points exist for each combination of factors.

The tool uses the data to test each factor's influence and how they interact. It tests the null hypothesis that the factors do not influence the outcomes against the alternate hypothesis that they do.

The most important statistics generated by the tool output are the p-values. The tool outputs three of these: one for each factor and another for the interaction between the two.

The first p-value to consider is the one labeled Interaction. If this statistic is less than the significance level (Alpha) you chose in step 5, the interaction between the two factors significantly influences the results. In the example used in this recipe, the p-value for Interaction is 0.298 (see Figure 1-56); there's not enough evidence to suggest that interactions between the starter culture brand and the yogurt maker model influence the time taken to culture yogurt.

If the p-value for Interaction is higher than the significance level, you next need to consider the p-values for Sample and Columns. If either value is less than the significance level, that factor significantly influences the outcomes. In the example used in this recipe, both p-values are less than the significance level of 0.05, meaning that the starter culture brand and the yogurt maker model influence the time taken to culture yogurt.

WARNING

The ANOVA: Two-Factor with Replication tool works with data sets with multiple data points for each combination of factors. If there's a single data point for each combination, use the ANOVA: Two-Factor without Replication tool instead.

9.20 Running a Fourier Analysis

Problem

You have periodic data set and want to transform it to another domain using the Fast Fourier Transform (FFT) method.

Solution

Use the Fourier Analysis tool.

Suppose you've recorded the value of a signal every 0.1 seconds, and you want to convert the results from a time series to the frequency domain. The range A2:A33 lists the index of each data point, B2:B33 lists the elapsed seconds, C2:C33 lists the signal data, and the first row contains labels (see Figure 1-57).

	A	B	C
1	Index	Time	Signal
2	0	0	0.220584502
3	1	0.1	-0.302384133
4	2	0.2	-0.538197598
5	3	0.3	0.145506613
6	4	0.4	0.570915726
7	5	0.5	-0.21725583
32	30	3	0.367596646
33	31	3.1	1.441062978

Figure 1-57. Signal data.

1. Go to the Data menu and choose Data Analysis from the Analyze group.
2. Select Fourier Analysis from the list of analysis tools, then click OK to open the tool's dialog box (see Figure 1-58).
3. Choose **C1:C33** in the Input Range box for the range of the signal data, including the label in C1.

4. Place a check in the Labels in First Row checkbox because the first row of the input range contains the data's label.
5. Choose an Output Range of **D2** so that the FFT for each row is output next to the original data.

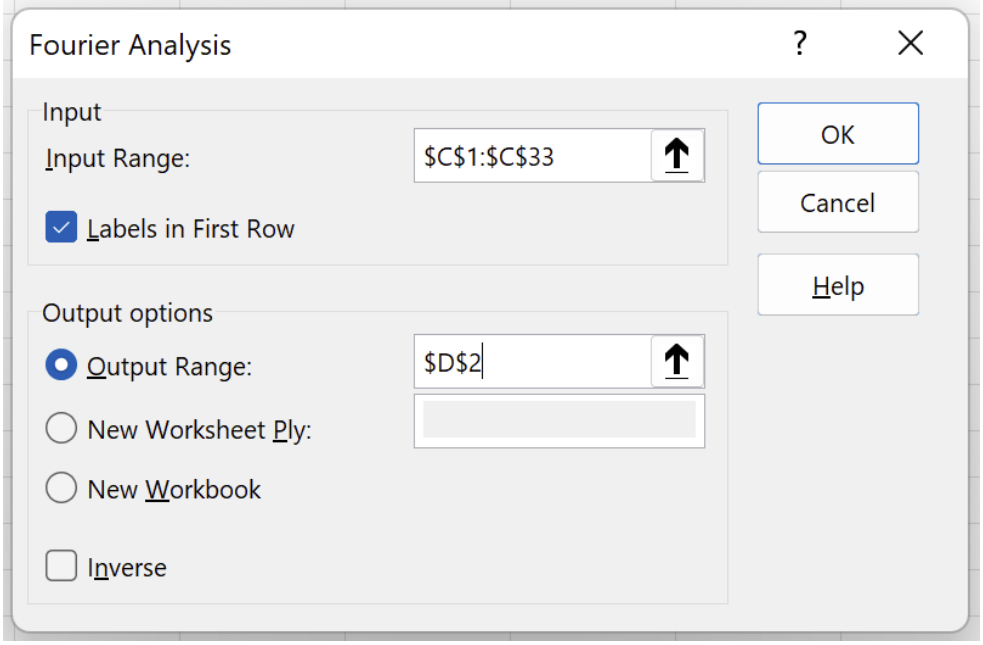


Figure 1-58. The Fourier Analysis tool's dialog box.

When you click OK, the tool outputs the FFT values as complex numbers (see Figure 1-59).

	A	B	C	D
1	Index	Time	Signal	Fourier
2	0	0	0.220584502	18.6784862485348
3	1	0.1	-0.302384133	-7.99226886665048+6.19937416958388i
4	2	0.2	-0.538197598	-0.985540594832308+4.57066502113923i
5	3	0.3	0.145506613	8.1843853508566-7.71795940452832i
6	4	0.4	0.570915726	-2.41131761533812+5.95845960634738i
7	5	0.5	-0.21725583	-1.18786774996966+3.38678196583616i
32	30	3	0.367596646	-0.98554059483232-4.57066502113923i
33	31	3.1	1.441062978	-7.9922688666505-6.19937416958386i

Figure 1-59. The Fourier Analysis tool's output[5].

Discussion

The Fourier Analysis tool transforms discrete, periodic data using the FFT method. It's an algorithm used in many areas, including engineering, music, science, and mathematics.

A common use of the FFT method is to transform a time series into the frequency domain and use the output to plot magnitude against frequency.

- To find the frequencies, take the index of each data point, then divide it by the number of data points multiplied by the sample rate (see Figure 1-60).
- To find the magnitudes, use the `IMABS` function to return the absolute value of the FFT output. Then divide each value by the number of data points divided by 2.

	A	B	C	D	E	F
1	Index	Time	Signal	Fourier	Frequency (Hz)	Magnitude
2	0	0	0.22058450	18.678486248534	=A2/($B$35*$B$36)	=IMABS(D2)/($B$35/2)
3	1	0.1	-0.30238413	-7.992268866650	=A3/($B$35*$B$36)	=IMABS(D3)/($B$35/2)
4	2	0.2	-0.53819759	-0.985540594832	=A4/($B$35*$B$36)	=IMABS(D4)/($B$35/2)
5	3	0.3	0.14550661	8.1843853508566	=A5/($B$35*$B$36)	=IMABS(D5)/($B$35/2)
6	4	0.4	0.57091572	-2.411317615338	=A6/($B$35*$B$36)	=IMABS(D6)/($B$35/2)
7	5	0.5	-0.21725583	-1.187867749969	=A7/($B$35*$B$36)	=IMABS(D7)/($B$35/2)
32	30	3	0.36759664	-0.985540594832	=A32/($B$35*$B$36)	=IMABS(D32)/($B$35/2)
33	31	3.1	1.44106297	-7.992268866650	=A33/($B$35*$B$36)	=IMABS(D33)/($B$35/2)
34						
35	Samples	32				
36	Interval	0.1				

Figure 1-60. Formulas calculating the frequency and magnitude of each point.

Once you have these results, you can use a scatter chart to plot magnitude against frequency (see Figures 1-61 and 1-62).

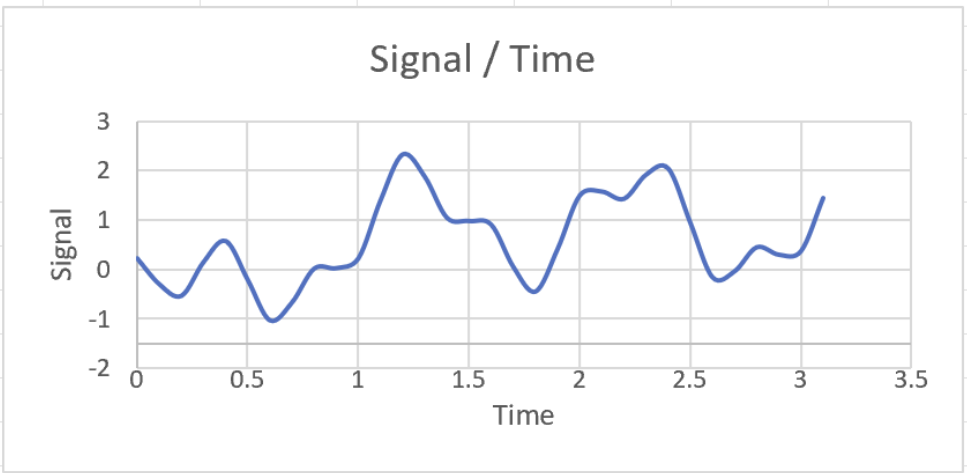


Figure 1-61. A chart showing signal plotted against time.

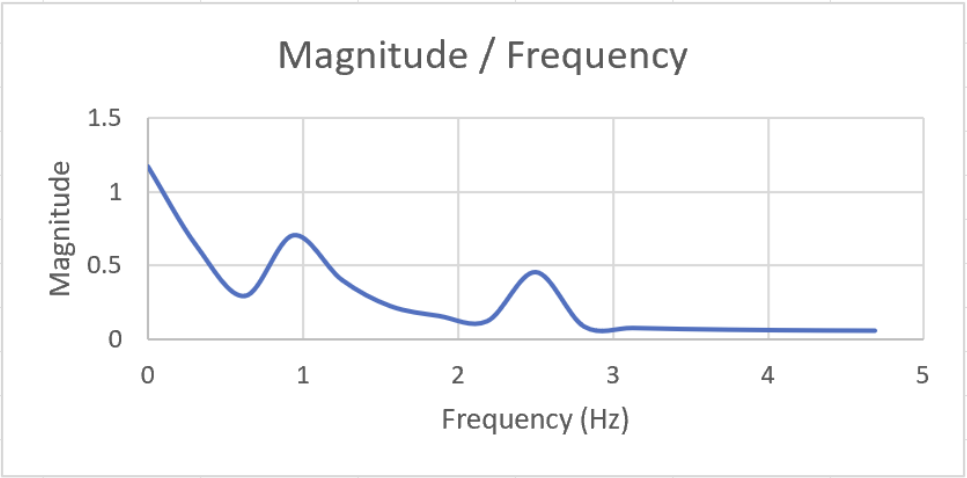


Figure 1-62. A chart showing magnitude plotted against frequency.

The Fourier Analysis tool also supports inverse transformations, which transform the FFT values back to the original data. To use this, place a check in the Inverse checkbox in the tool's dialog box.

WARNING

Due to the nature of the FFT algorithm, you must provide the Fourier Analysis tool with a range of values whose size is a power of 2 (2, 4, 8, 16, 32, and so on). If you don't have 2^n values, you must pad the end of the series with zeroes until the next power of 2.

[1] Extra Moving Average and Standard Error labels have been added to the first row.

[2] Extra Damping Factor and Standard Error labels have been added to the first row.

[3] An extra label has been added to G1.

[4] An extra lavel has been added to G1.

[5] An extra Fourier label has been added to the first row.

About the Author

Dawn Griffiths is an author and trainer with over 20 years experience using Excel. She has written various books, including *Head First Statistics*, *Head First Android Development*, *Head First Kotlin*, *Head First C* and *React Cookbook*, and is a contributing author to *97 Things Every Java Programmer Should Know*. Dawn also developed the animated video course *The Agile Sketchpad* with her husband, David, as a way of teaching key concepts and techniques in a way that keeps your brain active and engaged.

Dawn regularly runs live, online training classes for Excel on the O'Reilly learning platform. You can find out more and register for her upcoming classes here.