Using Excel’s Analysis ToolPak Add-In

S. Christian Albright, September 2013

Introduction

This document illustrates the use of Excel’s Analysis ToolPak add-in for data analysis. The document is aimed at users who prefer a free alternative to Palisade’s StatTools add-in. Although StatTools is provided for free to users of our books, it must be purchased separately by companies that want to use it. In contrast, Analysis ToolPak is bundled with Excel, so it is free for anyone who owns Excel.

Given that Analysis ToolPak is freely available in Excel, it is worth asking why the data analysis sections of our books are based on StatTools, not Analysis ToolPak. The answer requires a bit of history. Since the early days of Excel—at least 20 years ago—Analysis ToolPak has been part of Excel. Indeed, its current form is almost identical to its form then. Admittedly, Microsoft has recently revised many of Excel’s statistical functions to make them more accurate numerically and to provide a more consistent naming convention, but the functionality and user interface of Analysis ToolPak have changed hardly at all. This is somewhat curious, given Microsoft’s increasing attention to data analysis, but for whatever reasons, Microsoft has decided to focus on other data analysis features and keep Analysis ToolPak as is.

As Excel grew in importance in the 1990s, our students voiced strong opinions that we should perform all quantitative analysis, including statistical analysis, in Excel. At the time, this left only one option, Analysis ToolPak, for the statistical analysis. Frankly, I didn’t think it was up to the job. Therefore, in the mid-1990s, I wrote my own Excel statistical add-in, called StatPro, and it was actually the basis for the statistical sections in early editions of our books. Then around the year 2000, Palisade Corporation licensed StatPro from me and transformed it into the current StatTools add-in.[1]

The advantages of StatTools should be apparent to users of our books. It is powerful, it has a very simple user interface, and it is very easy to learn. Its one drawback is that it is not free. So, again, the purpose of this document is to acquaint you with the free Analysis ToolPak add-in—how it works and what it can do. Then you can decide which add-in is appropriate for you, StatTools or Analysis ToolPak. However, I make no attempt in this document to be unbiased. As I will discuss, Analysis ToolPak has some definite weaknesses that you should be aware of.

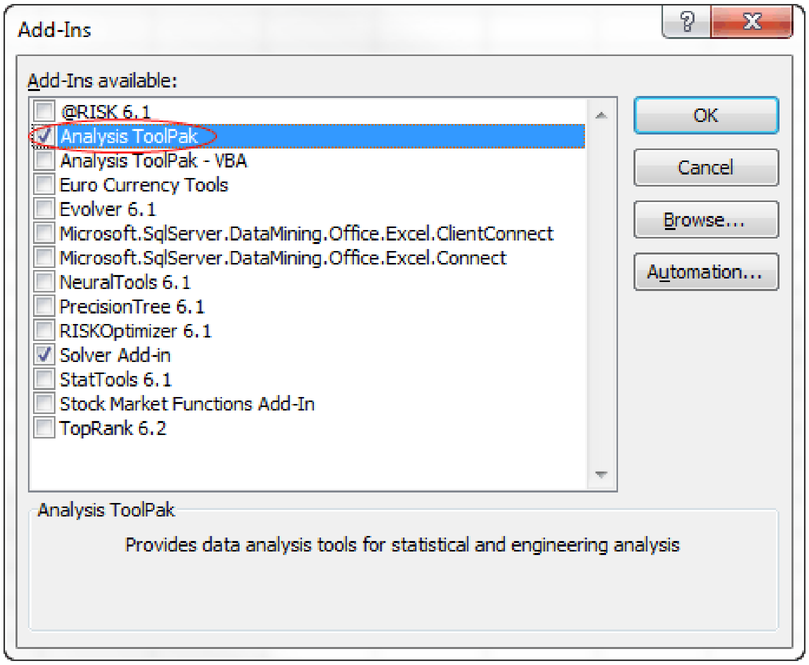
Loading Analysis ToolPak

Before using Analysis ToolPak, you must load it, just as you load other Excel add-ins:

1. Go to Excel Options. (In Excel 2010 or 2013, click the File tab and select Options. In Excel 2007, click the Office button and select Excel Options.)
2. Click the Add-Ins category.
3. Click the Go button toward the bottom.
4. Check the Analysis ToolPak item, as shown in Figure 1. (Note that there is also an Analysis ToolPak – VBA item. Unless you plan to automate Analysis ToolPak with VBA, which is very unlikely, there is no need to check this item.)

[1] StatPro is still freely available at http://www.kelley.iu.edu/albrightbooks/StatPro Add-In.htm. However, I no longer support it, and some of its features, especially graphical features, do not work properly in versions of Excel after 2003.

Figure 1 Add-Ins List



Add-Ins available:

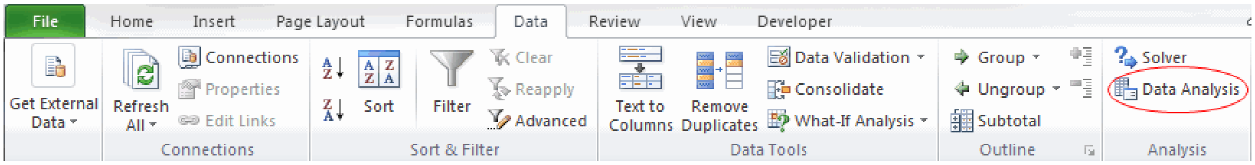
- ☐ @RISK 6.1
- ☑ Analysis ToolPak
- ☐ Analysis ToolPak - VBA
- ☐ Euro Currency Tools
- ☐ Evolver 6.1
- ☐ Microsoft.SqlServer.DataMining.Office.Excel.ClientConnect
- ☐ Microsoft.SqlServer.DataMining.Office.Excel.Connect
- ☐ NeuralTools 6.1
- ☐ PrecisionTree 6.1
- ☐ RISKOptimizer 6.1
- ☑ Solver Add-in
- ☐ StatTools 6.1
- ☐ Stock Market Functions Add-In
- ☐ TopRank 6.2

Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

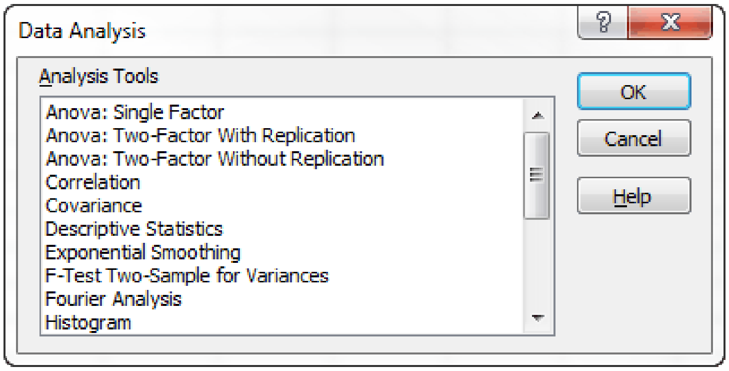
Once Analysis ToolPak is loaded, you will see a **Data Analysis** item on the Data ribbon. In fact, if you have also loaded the Solver add-in, the Data Analysis button is right below the Solver button, as shown in Figure 2.

Figure 2 Data Ribbon



When you click the Data Analysis button, you see the list of tools available, some of which appear in Figure 3. These tools are described in subsequent sections of this document.

Figure 3 Data Analysis Tools



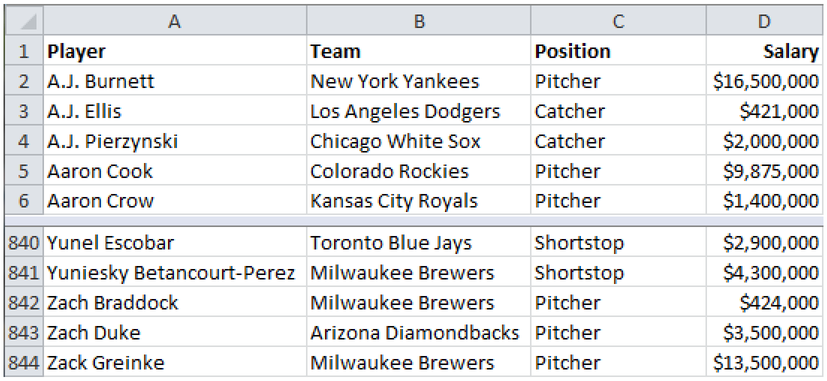
Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

Descriptive Statistics

You can obtain summary measures of numeric variables by selecting **Descriptive Statistics** from the Data Analysis Tools list in Figure 3. Here is an example based on the file **Baseball Salaries 2011.xlsx** (see Figure 4).

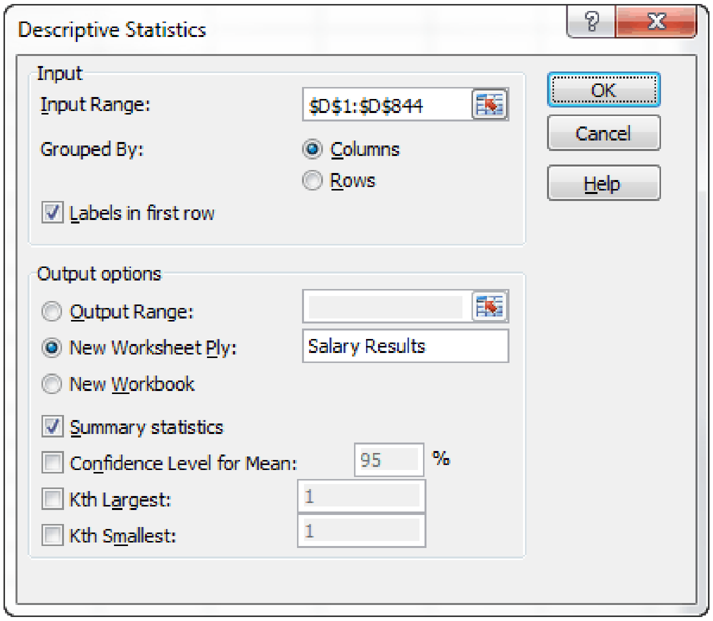
Figure 4 Baseball Salaries



	A	B	C	D
1	**Player**	**Team**	**Position**	**Salary**
2	A.J. Burnett	New York Yankees	Pitcher	$16,500,000
3	A.J. Ellis	Los Angeles Dodgers	Catcher	$421,000
4	A.J. Pierzynski	Chicago White Sox	Catcher	$2,000,000
5	Aaron Cook	Colorado Rockies	Pitcher	$9,875,000
6	Aaron Crow	Kansas City Royals	Pitcher	$1,400,000
840	Yunel Escobar	Toronto Blue Jays	Shortstop	$2,900,000
841	Yuniesky Betancourt-Perez	Milwaukee Brewers	Shortstop	$4,300,000
842	Zach Braddock	Milwaukee Brewers	Pitcher	$424,000
843	Zach Duke	Arizona Diamondbacks	Pitcher	$3,500,000
844	Zack Greinke	Milwaukee Brewers	Pitcher	$13,500,000

When you select Descriptive Statistics, you see the dialog box in Figure 5. It guesses correctly that the only numeric data are in the range D1:D844, although you have to check that labels are in the first row. The “Grouped By:” option should usually be “Columns,” meaning that each variable is in a column, not a row. There are three options for the location of the results, and if you choose the New Worksheet option, you can provide a name for this new worksheet. Finally, you can check any of the four options at the bottom, although none are checked by default.

Figure 5 Descriptive Statistics Dialog Box



Input

Input Range: $D$1:$D$844

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

○ Output Range:

● New Worksheet Ply: Salary Results

○ New Workbook

☑ Summary statistics

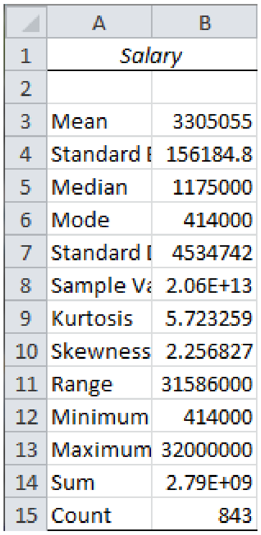
□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

The results appear in Figure 6. As you can see, one obvious weakness of the add-in is that results are not formatted very nicely. In particular, column widths are not changed to accommodate longer labels. You have to do this manually. Also, if you want to format the numbers, say, as currency or with more or fewer decimals, you have to do this manually. Finally, the results are static. For example, cell B2 contains the number 3305055, not a formula linked to the data.

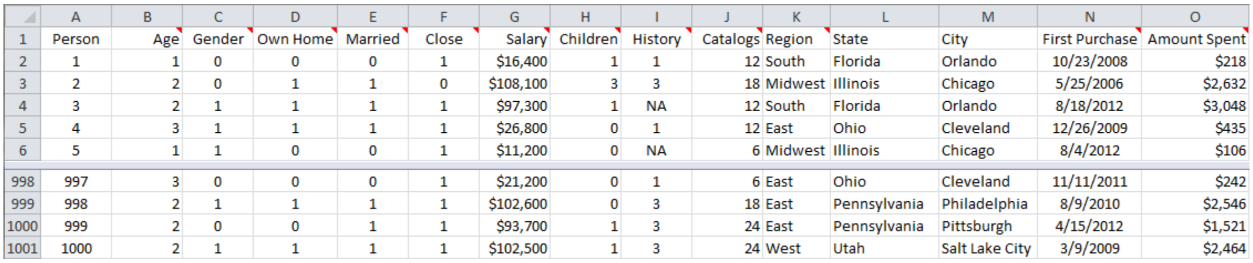
Figure 6 Summary Statistics of Salaries



	A	B
1	*Salary*	
2		
3	Mean	3305055
4	Standard E	156184.8
5	Median	1175000
6	Mode	414000
7	Standard I	4534742
8	Sample Va	2.06E+13
9	Kurtosis	5.723259
10	Skewness	2.256827
11	Range	31586000
12	Minimum	414000
13	Maximum	32000000
14	Sum	2.79E+09
15	Count	843

For another example, I opened the file **Catalog Marketing.xlsx** (see Figure 7) and again chose Descriptive Statistics from the Data Analysis Tools list. Two problems became apparent. The first is that Analysis ToolPak remembers the previous choices, so I saw *exactly* the same settings as in Figure 5, although they are no longer relevant.

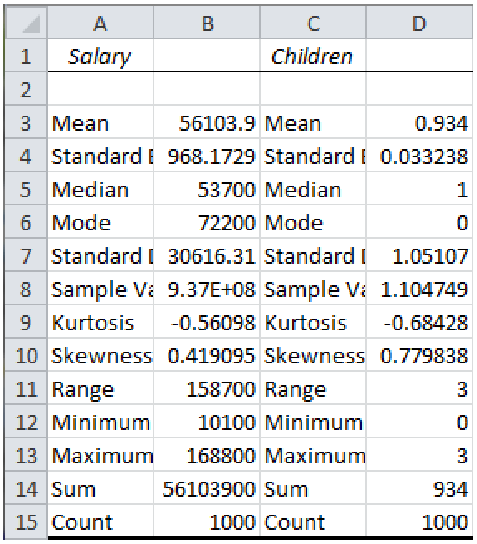
Figure 7 Catalog Marketing Data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2008	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2006	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2012	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2009	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2012	$106
998	997	3	0	0	0	1	$21,200	0	1	6	East	Ohio	Cleveland	11/11/2011	$242
999	998	2	1	1	1	1	$102,600	0	3	18	East	Pennsylvania	Philadelphia	8/9/2010	$2,546
1000	999	2	0	0	1	1	$93,700	1	3	24	East	Pennsylvania	Pittsburgh	4/15/2012	$1,521
1001	1000	2	1	1	1	1	$102,500	1	3	24	West	Utah	Salt Lake City	3/9/2009	$2,464

A bigger problem, however, is that the relevant columns for analysis, such as columns G and O, are not contiguous, and Analysis ToolPak will not allow you to choose noncontiguous columns. If you try to work around this by selecting all columns from G to O, you will be told that the analysis cannot be performed because this range contains non-numeric data. Your only choices, neither very convenient, are (1) to perform the analysis one variable (or several contiguous variables) at a time, or (2) to move the data around so that all numeric variables of interest are in contiguous columns. Furthermore, even if you do choose numeric variables in contiguous columns, the resulting output has redundant labels, as shown in Figure 8. Here, the Salary and Children variables were selected, and a separate column of labels is provided for each.

Figure 8 Summary Statistics for Salary and Children



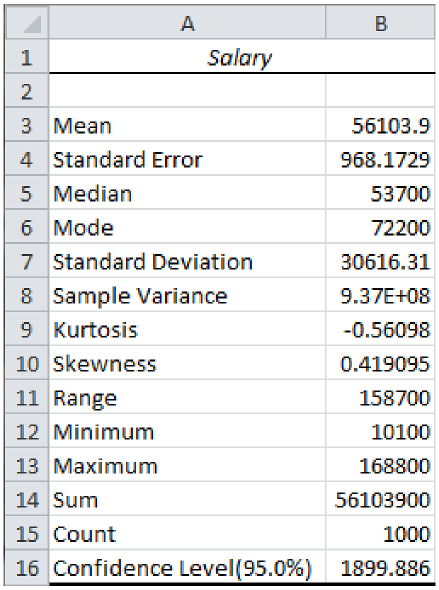
	A	B	C	D
1	*Salary*		*Children*	
2				
3	Mean	56103.9	Mean	0.934
4	Standard E	968.1729	Standard E	0.033238
5	Median	53700	Median	1
6	Mode	72200	Mode	0
7	Standard D	30616.31	Standard D	1.05107
8	Sample Va	9.37E+08	Sample Va	1.104749
9	Kurtosis	-0.56098	Kurtosis	-0.68428
10	Skewness	0.419095	Skewness	0.779838
11	Range	158700	Range	3
12	Minimum	10100	Minimum	0
13	Maximum	168800	Maximum	3
14	Sum	56103900	Sum	934
15	Count	1000	Count	1000

In my opinion, the requirement of contiguous columns for analysis is one of the biggest drawbacks of Analysis ToolPak. You will see it again in the section on regression, where it can be a real nuisance.

Two other comments about the Descriptive Statistics tool are worth mentioning. First, there is no way to break down summary statistics of Salary, say, by a categorical variable such as Region. (Recall that this is easy to do in StatTools with its Stacked Variables option.)

Second, suppose you check the Confidence Level for Mean option in Figure 5. What do you get? The output for the Salary variable appears in Figure 9. Unfortunately, no interpretation of the 1899.886 value in row 16 is provided (and good luck finding anything useful under Help). It turns out that this is the value that should be subtracted from and added to the sample mean to get a 95% confidence interval for the mean—that is, it is approximately 2 times the standard error of the mean.

Figure 9 Confidence Level for Mean

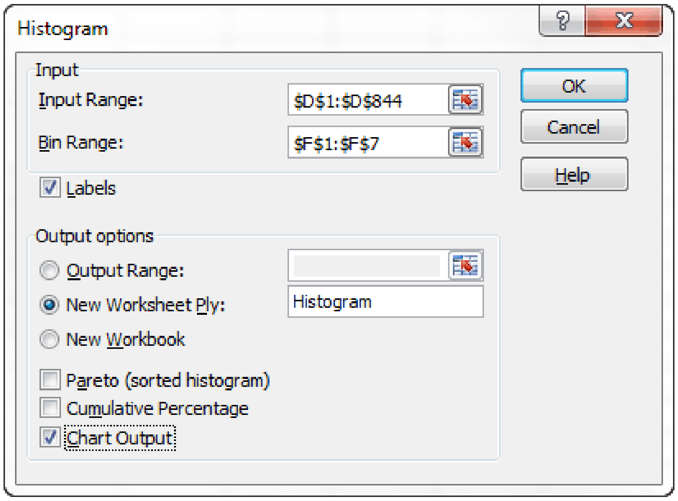


	A	B
1	*Salary*	
2		
3	Mean	56103.9
4	Standard Error	968.1729
5	Median	53700
6	Mode	72200
7	Standard Deviation	30616.31
8	Sample Variance	9.37E+08
9	Kurtosis	-0.56098
10	Skewness	0.419095
11	Range	158700
12	Minimum	10100
13	Maximum	168800
14	Sum	56103900
15	Count	1000
16	Confidence Level(95.0%)	1899.886

Histograms

The **Histogram** option in Analysis ToolPak allows you to create a frequency table and accompanying chart of a numeric variable. It is illustrated here for the Salary variable in the **Baseball Salaries 2011.xlsx** file. However, as you can see in Figure 10, the Histogram dialog box not only requires the range for the data variable, but it also requires a “Bins” range. Unlike in StatTools, there are no default bins; you have to choose the bins and place them somewhere in the worksheet, as shown in Figure 11. In this case, there will be 7 bins: less than or equal to 500,000, greater than 500,000 but less than or equal to 1,000,000, and so on, up to greater than 3,000,000.

Figure 10 Histogram Dialog Box



Histogram

Input

Input Range: $D$1:$D$844

Bin Range: $F$1:$F$7

☑ Labels

Output options

○ Output Range:

● New Worksheet Ply: Histogram

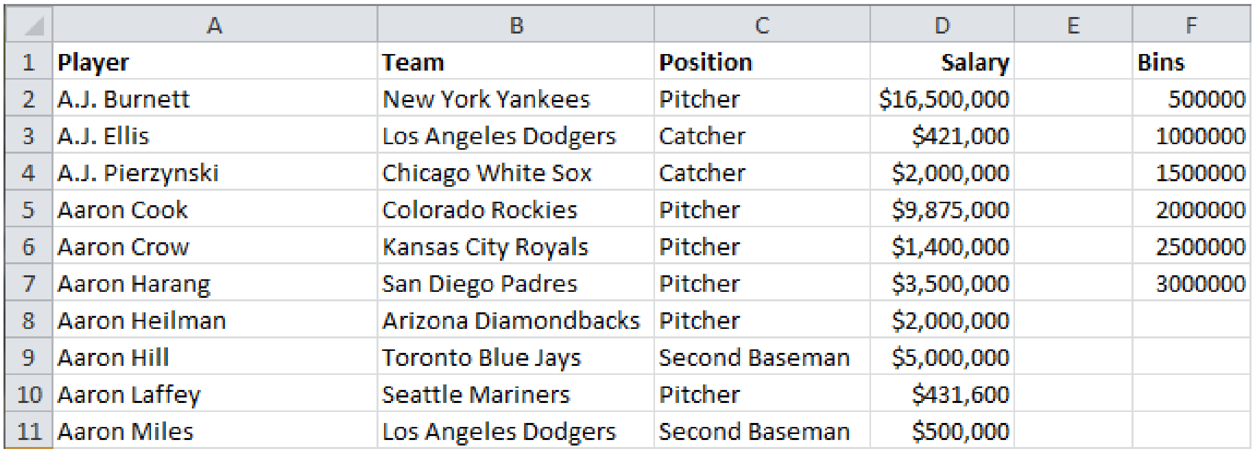
○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

☑ Chart Output

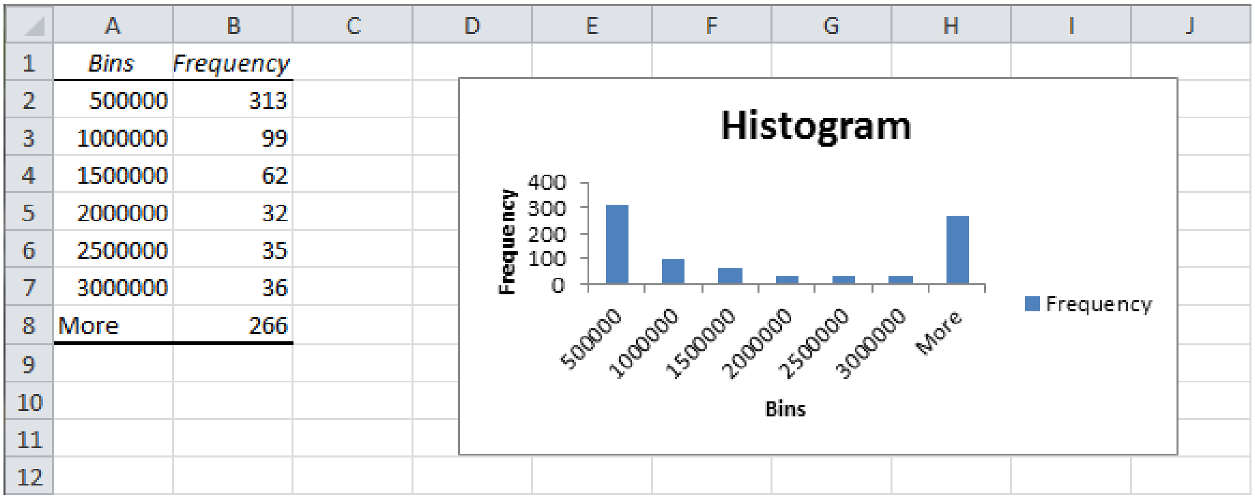
Figure 11 Salary Data with Bins



	A	B	C	D	E	F
1	**Player**	**Team**	**Position**	**Salary**		**Bins**
2	A.J. Burnett	New York Yankees	Pitcher	$16,500,000		500000
3	A.J. Ellis	Los Angeles Dodgers	Catcher	$421,000		1000000
4	A.J. Pierzynski	Chicago White Sox	Catcher	$2,000,000		1500000
5	Aaron Cook	Colorado Rockies	Pitcher	$9,875,000		2000000
6	Aaron Crow	Kansas City Royals	Pitcher	$1,400,000		2500000
7	Aaron Harang	San Diego Padres	Pitcher	$3,500,000		3000000
8	Aaron Heilman	Arizona Diamondbacks	Pitcher	$2,000,000		
9	Aaron Hill	Toronto Blue Jays	Second Baseman	$5,000,000		
10	Aaron Laffey	Seattle Mariners	Pitcher	$431,600		
11	Aaron Miles	Los Angeles Dodgers	Second Baseman	$500,000		

The results appear in Figure 12. They include the table of bin frequencies and the corresponding chart. If you prefer the bars in the histogram to be right next to one another, you can right-click any bar, select Format Data Series, and choose a Gap Width of 0. Also, you can delete the Frequency legend.

Figure 12 Histogram of Salaries



	A	B
1	*Bins*	*Frequency*
2	500000	313
3	1000000	99
4	1500000	62
5	2000000	32
6	2500000	35
7	3000000	36
8	More	266

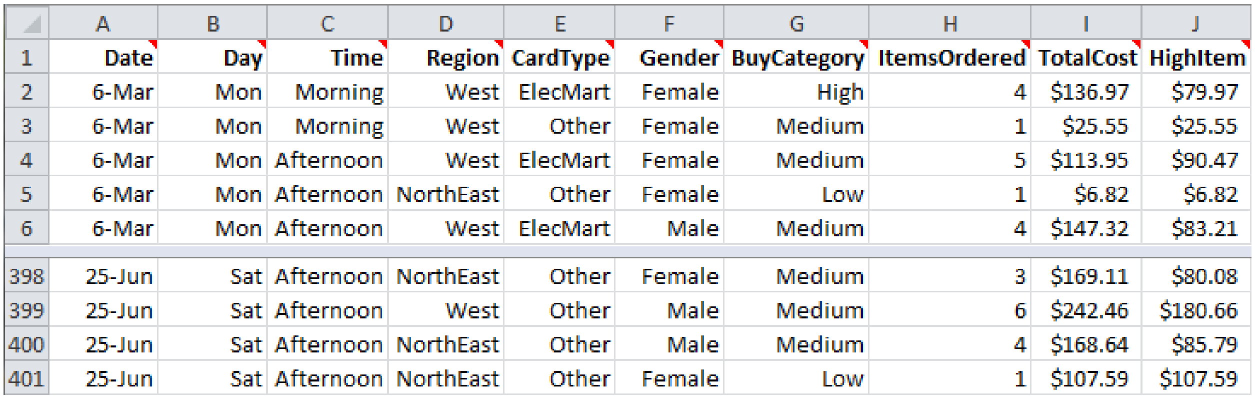
Other Common Charts

StatTools provides four basic statistical charts: histograms, box plots, scatterplots, and time series graphs. Analysis ToolPak provides only the first of these, histograms. Of course, scatterplots are fairly easy to create with Excel's built-in chart tools: just ask for a chart of the scatter type. Also, time series graphs are easy to create by requesting a line chart of a time series variable. Unfortunately, there is no easy way, using Excel built-in tools only, to create a box plot.

Correlation (and Covariance)

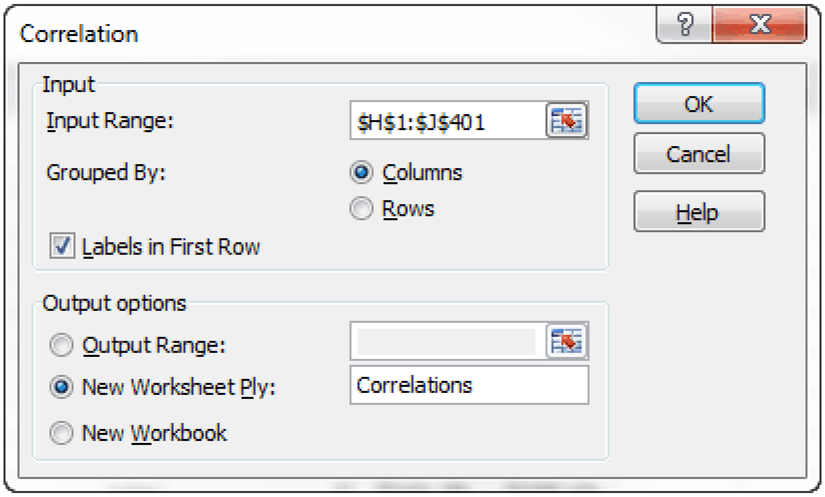
It is easy to create a table of correlations with Analysis ToolPak, as illustrated here with the last three columns of data in the file **Elecmart Sales.xlsx** (see Figure 13). You choose **Correlation** from the Data Analysis Tools list and fill out the dialog box as shown in Figure 14. The resulting table of correlations appears in Figure 15.

Figure 13 Elecmart Sales Data



	A	B	C	D	E	F	G	H	I	J
1	Date	Day	Time	Region	CardType	Gender	BuyCategory	ItemsOrdered	TotalCost	HighItem
2	6-Mar	Mon	Morning	West	ElecMart	Female	High	4	$136.97	$79.97
3	6-Mar	Mon	Morning	West	Other	Female	Medium	1	$25.55	$25.55
4	6-Mar	Mon	Afternoon	West	ElecMart	Female	Medium	5	$113.95	$90.47
5	6-Mar	Mon	Afternoon	NorthEast	Other	Female	Low	1	$6.82	$6.82
6	6-Mar	Mon	Afternoon	West	ElecMart	Male	Medium	4	$147.32	$83.21
398	25-Jun	Sat	Afternoon	NorthEast	Other	Female	Medium	3	$169.11	$80.08
399	25-Jun	Sat	Afternoon	West	Other	Male	Medium	6	$242.46	$180.66
400	25-Jun	Sat	Afternoon	NorthEast	Other	Male	Medium	4	$168.64	$85.79
401	25-Jun	Sat	Afternoon	NorthEast	Other	Female	Low	1	$107.59	$107.59

Figure 14 Correlations Dialog Box



Correlation

Input

Input Range: $H$1:$J$401

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

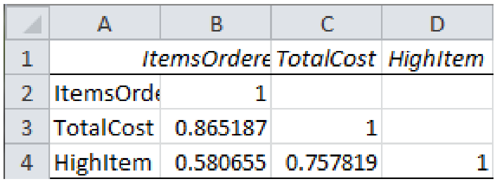
Output options

○ Output Range:

● New Worksheet Ply: Correlations

○ New Workbook

Figure 15 Table of Correlations



	A	B	C	D
1		*ItemsOrdere*	*TotalCost*	*HighItem*
2	ItemsOrd	1		
3	TotalCost	0.865187	1	
4	HighItem	0.580655	0.757819	1

Three of the limitations mentioned in the Descriptive Statistics section apply here as well: (1) the variables must be in contiguous columns, (2) the results are static, not formulas linked to the data, and (3) the column widths need to be adjusted to accommodate long labels.

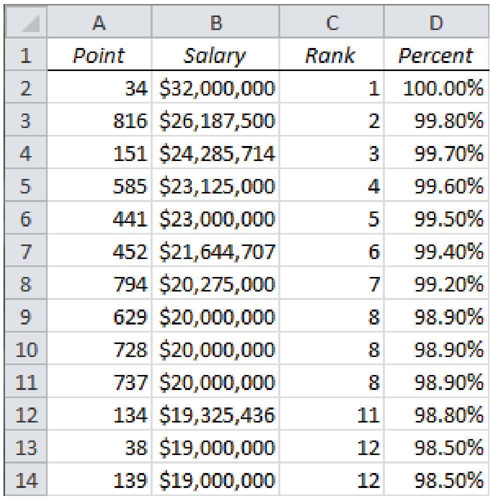
There is one other thing you should be aware of: missing values. If you open the file **Golf Stats.xlsx** and try to create a table of correlations for the variables in columns J to N (not shown here), you will get a message that this table can't be created because the range contains non-numeric data. Actually, it contains blanks—or at least cells that look blank. StatTools handles these blank cells with no problem, but Analysis ToolPak doesn't. However, if you select each of these blank cells and press Delete, then the Analysis ToolPak Correlation procedure works. This is probably more a peculiarity of Excel than of Analysis ToolPak: How can you tell whether a cell is really "blank"?

Analysis ToolPak also has a **Covariance** procedure for generating a table of covariances. It works exactly like the Correlation procedure. Surprisingly, however, the diagonal of this table contains *formulas* with the VARP (or the newer VAR.P) function, whereas the off-diagonal cells are simply numbers. In case you're interested, the covariances in the off-diagonal cells are the *population* covariances (denominator *n*, not *n*-1).

Rank and Percentile

Analysis ToolPak has a **Rank and Percentile** procedure that you might find useful. You select a column of numeric data, and the procedure essentially sorts the data from high to low. Figure 16 shows the results of doing this to the Salary variable in the file **Baseball Salaries 2011.xlsx**. The Rank column is equivalent to using Excel's RANK (or the newer RANK.EQ) function. The Percent column lists the approximate percentage of salaries at or below each given salary. It is equivalent to using the PERCENTRANK.INC function (available starting in Excel 2010).

Figure 16 Ranks of Baseball Salaries



	A	B	C	D
1	*Point*	*Salary*	*Rank*	*Percent*
2	34	$32,000,000	1	100.00%
3	816	$26,187,500	2	99.80%
4	151	$24,285,714	3	99.70%
5	585	$23,125,000	4	99.60%
6	441	$23,000,000	5	99.50%
7	452	$21,644,707	6	99.40%
8	794	$20,275,000	7	99.20%
9	629	$20,000,000	8	98.90%
10	728	$20,000,000	8	98.90%
11	737	$20,000,000	8	98.90%
12	134	$19,325,436	11	98.80%
13	38	$19,000,000	12	98.50%
14	139	$19,000,000	12	98.50%

Hypothesis Tests

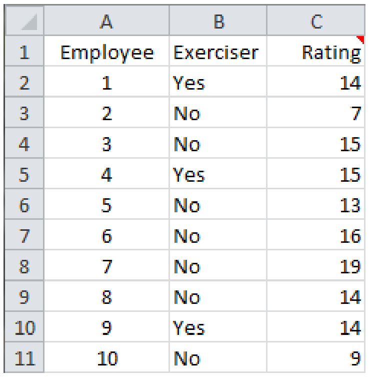
Analysis ToolPak has five separate procedures for implementing common hypothesis tests. (It also has several ANOVA procedures that will be discussed in a separate section.) Three of these are for testing the difference between two sample means when the two samples are independent. Another tests the difference between two sample means when the two samples are paired. Finally, there is a test for equality of two sample variances.

Tests for Difference between Two Sample Means: Independent Samples

These procedures are labeled **z-Test: Two Sample for Means**, **t-Test: Two-Sample Assuming Equal Variances**, and **t-Test: Two-Sample Assuming Unequal Variances**. The first assumes the population variances are known, whereas the last two make no such assumption. (There is no analog to the first test in StatTools, and StatTools performs the last two tests simultaneously.)

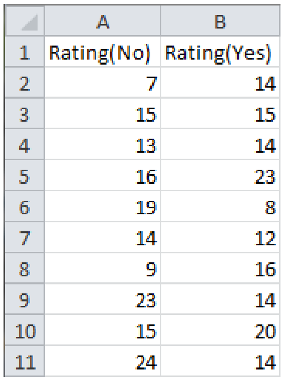
Unfortunately, unlike StatTools, Analysis ToolPak requires the data to be unstacked. As an example, the data in the file **Exercise & Productivity.xlsx** are stacked (see Figure 17). There is a categorical variable Exerciser and a numeric variable Rating. Indeed, this is the usual data arrangement in such data sets. However, to use any of the Analysis ToolPak tests for testing the mean rating across exercisers and non-exercisers, the data must first be unstacked, as shown in Figure 18, where the two column lengths for the unstacked variables are not necessarily the same. StatTools has a utility for unstacking, but presumably you are using Analysis ToolPak because you don't *have* StatTools. Therefore, you have to unstack the data manually (by sorting on Exerciser and then copying and pasting).

Figure 17 Stacked Exercise Data



	A	B	C
1	Employee	Exerciser	Rating
2	1	Yes	14
3	2	No	7
4	3	No	15
5	4	Yes	15
6	5	No	13
7	6	No	16
8	7	No	19
9	8	No	14
10	9	Yes	14
11	10	No	9

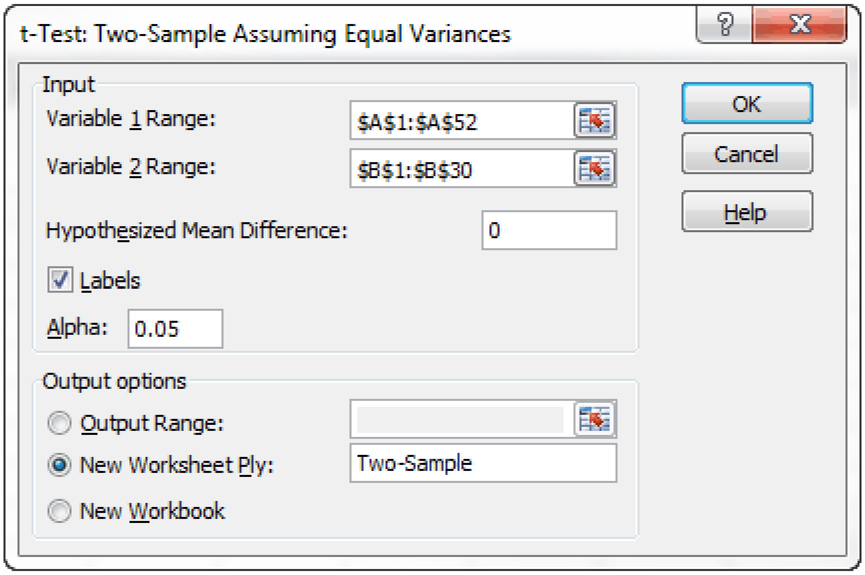
Figure 18 Unstacked Exercise Data



	A	B
1	Rating(No)	Rating(Yes)
2	7	14
3	15	15
4	13	14
5	16	23
6	19	8
7	14	12
8	9	16
9	23	14
10	15	20
11	24	14

In any case, once you have the unstacked data, all three of the procedures are straightforward and similar. For example, the dialog box for the **t-Test: Two-Sample Assuming Equal Variances** procedure is shown in Figure 19.

Figure 19 Two-Sample Test Dialog Box



t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range: $A$1:$A$52

Variable 2 Range: $B$1:$B$30

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

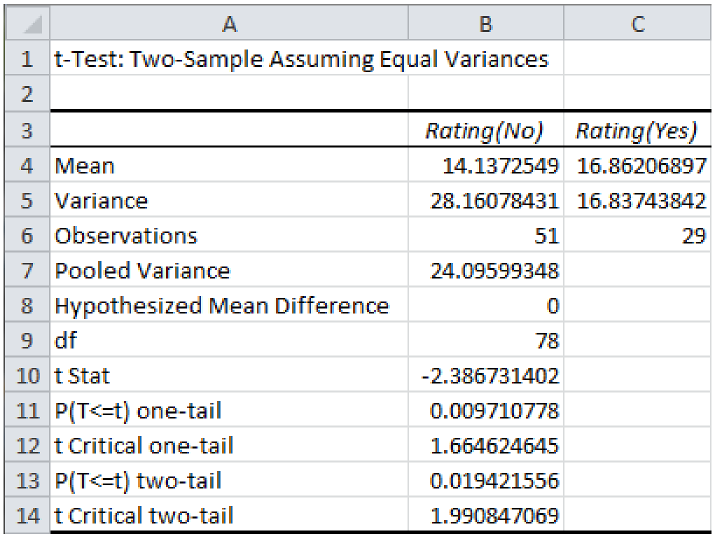
○ Output Range:

● New Worksheet Ply: Two-Sample

○ New Workbook

After widening the columns appropriately, the results appear in Figure 20. Interestingly, even though the dialog box asks for a significance level (alpha), it is not used in the results at all. However, you can mentally compare your alpha level to the p-value shown in cell B11 for a one-tailed test or in cell B13 for a two-tailed test.

Figure 20 Two-Sample Test Results

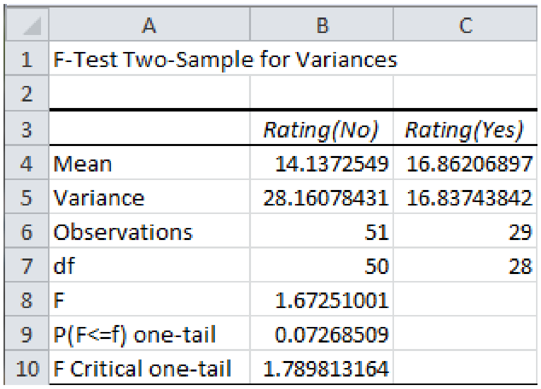


	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		*Rating(No)*	*Rating(Yes)*
4	Mean	14.1372549	16.86206897
5	Variance	28.16078431	16.83743842
6	Observations	51	29
7	Pooled Variance	24.09599348	
8	Hypothesized Mean Difference	0	
9	df	78	
10	t Stat	-2.386731402	
11	P(T<=t) one-tail	0.009710778	
12	t Critical one-tail	1.664624645	
13	P(T<=t) two-tail	0.019421556	
14	t Critical two-tail	1.990847069	

Test for Equality of Variances

Given the test results in Figure 20, you might want to check whether the equal-variance assumption is reasonable. You can do this with the **F-Test Two-Sample for Variances** procedure, again using the unstacked data. The dialog box is filled out exactly as in Figure 19, and the results appear in Figure 21. The p-value of about 0.07 indicates that there is evidence, but not totally convincing evidence, that the two variances are *not* equal.

Figure 21 Equal Variance Test Results

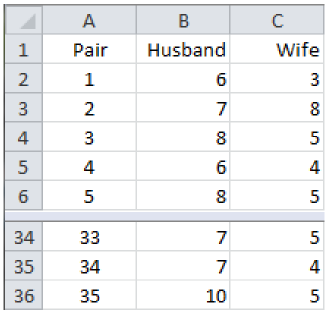


	A	B	C
1	F-Test Two-Sample for Variances		
2			
3		*Rating(No)*	*Rating(Yes)*
4	Mean	14.1372549	16.86206897
5	Variance	28.16078431	16.83743842
6	Observations	51	29
7	df	50	28
8	F	1.67251001	
9	P(F<=f) one-tail	0.07268509	
10	F Critical one-tail	1.789813164	

Test for Difference between Two Sample Means: Paired Samples

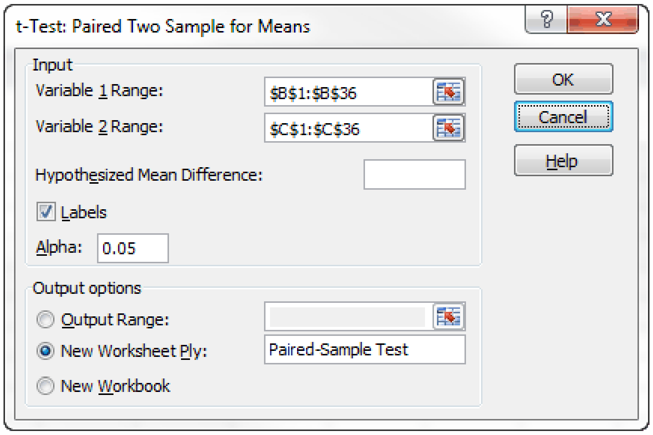
If you are comparing two samples that are paired in some natural way, you should use the **t-Test: Paired Two Sample for Means** procedure. As an example, the husband and wife ratings in the file **Sales Presentation Ratings.xlsx** are naturally paired, assuming that the reactions of husbands and wives to sales presentations are correlated (see Figure 22). These data are already unstacked, as Analysis ToolPak requires, so the Paired Sample dialog box can be filled in directly, as shown in Figure 23. The results then appear in Figure 24. For example, if you want this to be a two-tailed test, the negligible p-value in cell B13 indicates that the husbands react differently, on average, than their wives. (Actually, husbands rate higher, on average.)

Figure 22 Paired Sales Presentation Ratings



	A	B	C
1	Pair	Husband	Wife
2	1	6	3
3	2	7	8
4	3	8	5
5	4	6	4
6	5	8	5
34	33	7	5
35	34	7	4
36	35	10	5

Figure 23 Paired-Sample Test Dialog Box



t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $B$1:$B$36

Variable 2 Range: $C$1:$C$36

Hypothesized Mean Difference:

☑ Labels

Alpha: 0.05

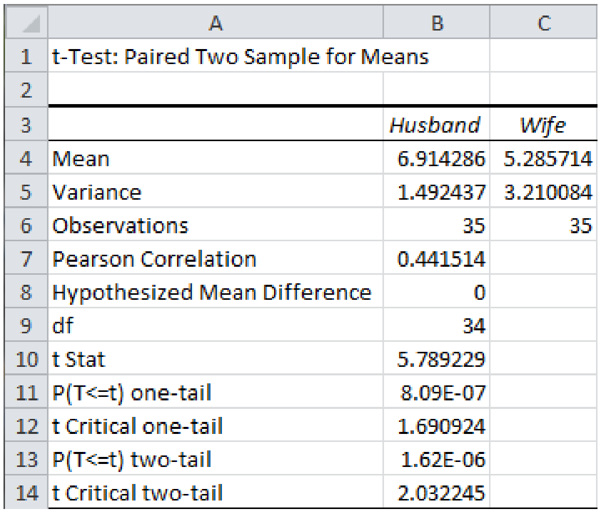
Output options

○ Output Range:

● New Worksheet Ply: Paired-Sample Test

○ New Workbook

Figure 24 Paired-Sample Test Results



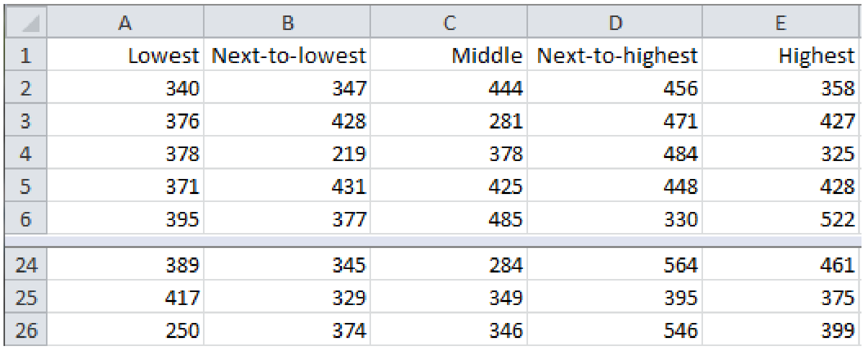
	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		*Husband*	*Wife*
4	Mean	6.914286	5.285714
5	Variance	1.492437	3.210084
6	Observations	35	35
7	Pearson Correlation	0.441514	
8	Hypothesized Mean Difference	0	
9	df	34	
10	t Stat	5.789229	
11	P(T<=t) one-tail	8.09E-07	
12	t Critical one-tail	1.690924	
13	P(T<=t) two-tail	1.62E-06	
14	t Critical two-tail	2.032245	

Analysis of Variance (ANOVA) Procedures

Single-Factor ANOVA

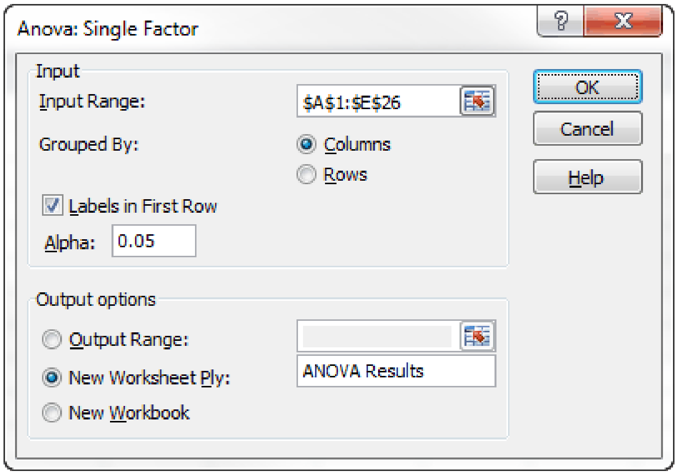
Single-factor ANOVA, also called one-way ANOVA, is an extension of the two-sample t-test (with independent samples) to more than two samples. It tests whether the means of all samples are equal. The Analysis ToolPak's **Anova: Single Factor** procedure implements this test, again assuming unstacked data. As an example, the file **Cereal Sales.xlsx** lists cereal sales at a supermarket chain for five different shelf heights (see Figure 25). To run the analysis, you fill out the dialog box as shown in Figure 26.

Figure 25 Cereal Sales Data



	A	B	C	D	E
1	Lowest	Next-to-lowest	Middle	Next-to-highest	Highest
2	340	347	444	456	358
3	376	428	281	471	427
4	378	219	378	484	325
5	371	431	425	448	428
6	395	377	485	330	522
24	389	345	284	564	461
25	417	329	349	395	375
26	250	374	346	546	399

Figure 26 Single-Factor ANOVA Dialog Box



Anova: Single Factor

Input

Input Range: $A$1:$E$26

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Alpha: 0.05

Output options

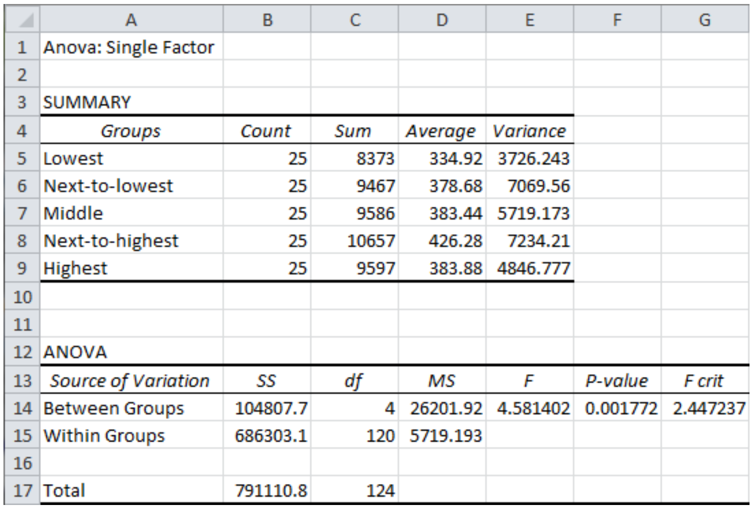
○ Output Range:

● New Worksheet Ply: ANOVA Results

○ New Workbook

The results appear in Figure 27. The sample mean and variance for each shelf height are listed, followed by the ANOVA table for the test. In this case, its very small p-value indicates that the means are *not* all equal. Unfortunately, the output does not indicate which means are significantly different from which others. (StatTools provides confidence intervals for this purpose.)

Figure 27 ANOVA Results for Cereal Data

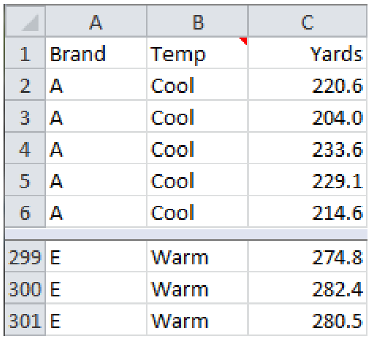


	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Lowest	25	8373	334.92	3726.243		
6	Next-to-lowest	25	9467	378.68	7069.56		
7	Middle	25	9586	383.44	5719.173		
8	Next-to-highest	25	10657	426.28	7234.21		
9	Highest	25	9597	383.88	4846.777		
10							
11							
12	ANOVA						
13	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
14	Between Groups	104807.7	4	26201.92	4.581402	0.001772	2.447237
15	Within Groups	686303.1	120	5719.193			
16							
17	Total	791110.8	124				

Two-Factor ANOVA with Replication

Analysis ToolPak's **Anova: Two-Factor With Replication** procedure is an extension of the single-factor ANOVA procedure. Now there are two factors, and observations are made for each combination of the two factor levels. (As in StatTools, there must be an *equal* number of observations for each combination—that is, it must be a balanced design.) The test is again basically a test of equal means, or equivalently, of equal factor-level effects. As an example, the file **Golf Ball.xlsx** contains 20 observations for each of five brands of golf ball and each of three temperatures. Each observation is the length of a drive (see Figure 28). Arguably, this data arrangement, where there are two categorical variables for the factors and one numeric variable, is the most natural arrangement, and this is the arrangement expected by the StatTools Two-Way ANOVA procedure.

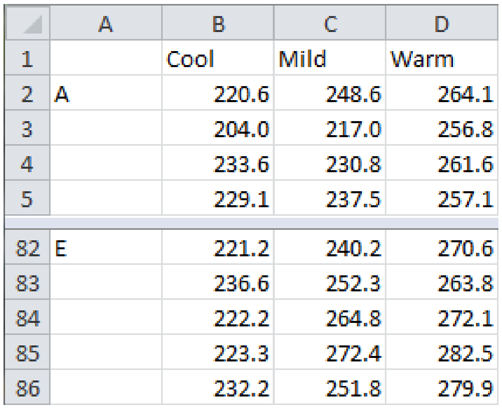
Figure 28 Golf Ball Data



	A	B	C
1	Brand	Temp	Yards
2	A	Cool	220.6
3	A	Cool	204.0
4	A	Cool	233.6
5	A	Cool	229.1
6	A	Cool	214.6
299	E	Warm	274.8
300	E	Warm	282.4
301	E	Warm	280.5

However, as indicated in Analysis ToolPak help, the arrangement required for its ANOVA procedure is shown in Figure 29. For each temperature, there are 20 observations for brand A, 20 for brand B, and so on. If you start with the setup in Figure 28, there is no easy way to rearrange the data as required other than by copying and pasting.

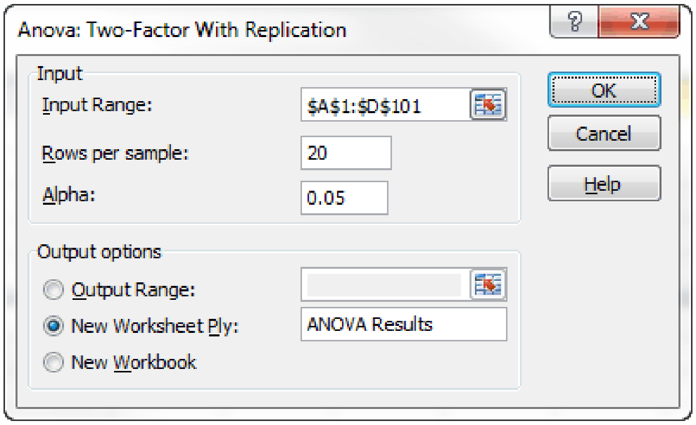
Figure 29 Rearranged Golf Ball Data



	A	B	C	D
1		Cool	Mild	Warm
2	A	220.6	248.6	264.1
3		204.0	217.0	256.8
4		233.6	230.8	261.6
5		229.1	237.5	257.1
82	E	221.2	240.2	270.6
83		236.6	252.3	263.8
84		222.2	264.8	272.1
85		223.3	272.4	282.5
86		232.2	251.8	279.9

With this required data setup, the ANOVA dialog box should be filled out as shown in Figure 30, where "Labels" indicates the labels in column A and row 1.

Figure 30 Two-Factor ANOVA with Replications Dialog Box



Anova: Two-Factor With Replication

Input

Input Range: $A$1:$D$101

Rows per sample: 20

Alpha: 0.05

Output options

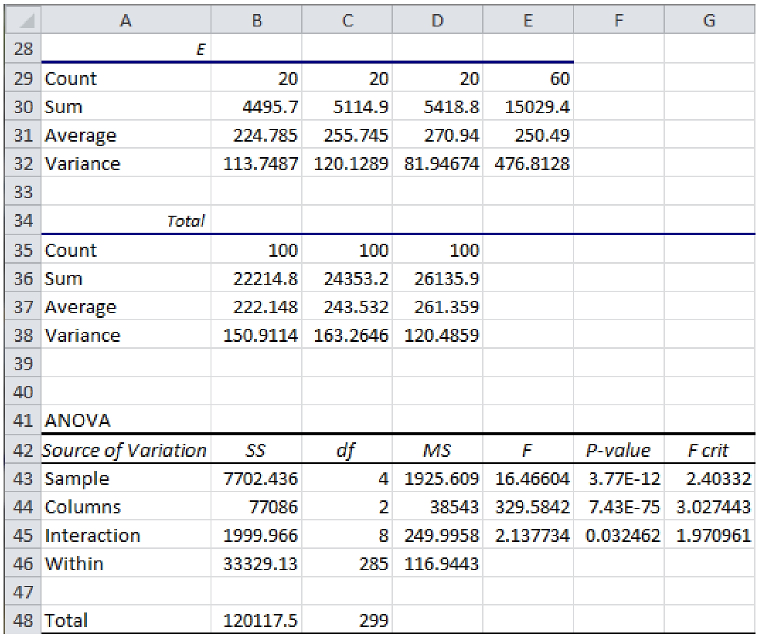
○ Output Range:

● New Worksheet Ply: ANOVA Results

○ New Workbook

Some of the results appear in Figure 31. There is actually a summary measure section for each brand; only the section for brand E is shown. More importantly, the ANOVA table shows tests for equal brand effects in row 43, equal temperature effects in row 44, and interaction effects in row 45. Usually the latter is examined first. If interactions are significant—and the small p-value in row 45 indicates that they are, at least at the 5% significance level—this means that the effect of one factor depends on the level of the other factor. For example, one brand of golf ball might go farther in cool temperatures, and another brand might go farther in warm temperatures. (This is only one of several possible types of interaction effects.) These interactions—or the lack of them—are made more apparent in the graphs of the sample means provided by StatTools. Unfortunately, such graphs are not provided by Analysis ToolPak. You can create the graphs manually, or you can look carefully at the tables of sample means.

Figure 31 Two-Factor ANOVA with Replications Results

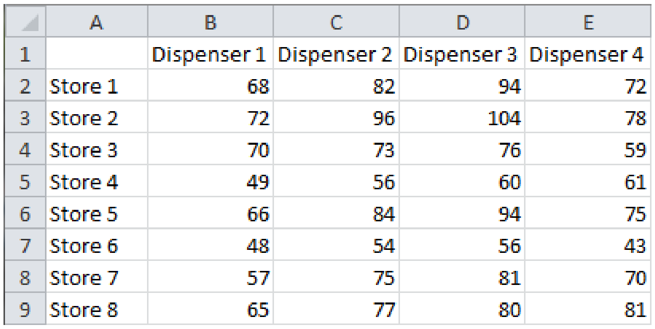


	A	B	C	D	E	F	G
28	*E*						
29	Count	20	20	20	60		
30	Sum	4495.7	5114.9	5418.8	15029.4		
31	Average	224.785	255.745	270.94	250.49		
32	Variance	113.7487	120.1289	81.94674	476.8128		
33							
34	*Total*						
35	Count	100	100	100			
36	Sum	22214.8	24353.2	26135.9			
37	Average	222.148	243.532	261.359			
38	Variance	150.9114	163.2646	120.4859			
39							
40							
41	ANOVA						
42	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
43	Sample	7702.436	4	1925.609	16.46604	3.77E-12	2.40332
44	Columns	77086	2	38543	329.5842	7.43E-75	3.027443
45	Interaction	1999.966	8	249.9958	2.137734	0.032462	1.970961
46	Within	33329.13	285	116.9443			
47							
48	Total	120117.5	299				

Two-Factor ANOVA without Replication

Analysis ToolPak also has an **Anova: Two-Factor Without Replication** procedure. It is exactly like the "With Replication" procedure except that there is only *one* observation per factor-level combination. As an example, the file **Soap Sales.xlsx** has one observation on soap sales for each of eight stores and each of four soap dispensers. The data in the file are arranged as in Figure 28 of the golf ball example, but they again need to be rearranged, as shown in Figure 32. With this arrangement, the ANOVA dialog box is filled out exactly as in Figure 30, except that Rows per sample is now 1.

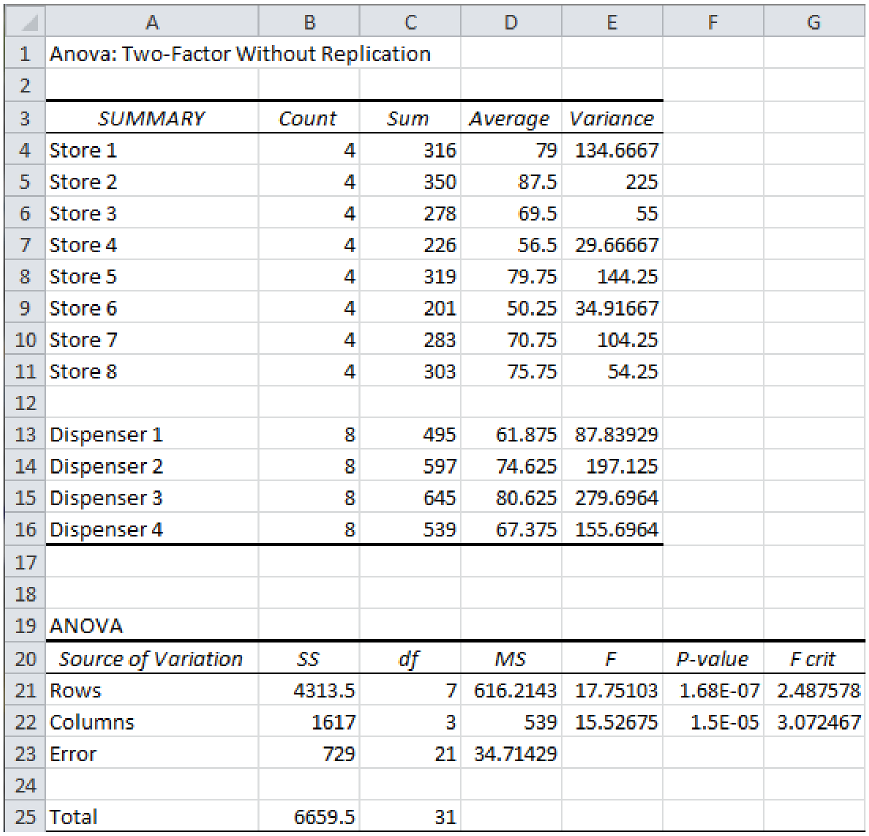
Figure 32 Rearranged Soap Sales Data



	A	B	C	D	E
1		Dispenser 1	Dispenser 2	Dispenser 3	Dispenser 4
2	Store 1	68	82	94	72
3	Store 2	72	96	104	78
4	Store 3	70	73	76	59
5	Store 4	49	56	60	61
6	Store 5	66	84	94	75
7	Store 6	48	54	56	43
8	Store 7	57	75	81	70
9	Store 8	65	77	80	81

The results appear in Figure 33. Summary statistics are listed for each level of each of the two factors, and the ANOVA table shows the results of the tests. However, because there is only one observation for each factor-level combination, it is impossible to test for interactions, which is why the Interactions row is missing. The implicit assumption is that there are *no* interactions, only main effects. In this particular example, the Store factor acts as a "blocking" variable, and the main interest is in the effect of the Dispenser factor. Therefore, the most important row is row 22. Its low p-value indicates that dispenser type has a significant effect on sales, even after controlling for different stores.

Figure 33 ANOVA Results for Soap Sales

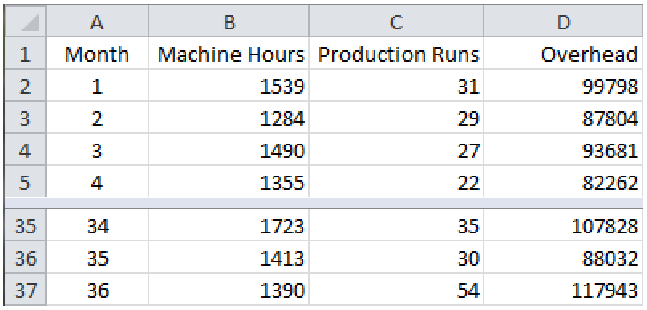


	A	B	C	D	E	F	G
1	Anova: Two-Factor Without Replication						
2							
3	*SUMMARY*	*Count*	*Sum*	*Average*	*Variance*		
4	Store 1	4	316	79	134.6667		
5	Store 2	4	350	87.5	225		
6	Store 3	4	278	69.5	55		
7	Store 4	4	226	56.5	29.66667		
8	Store 5	4	319	79.75	144.25		
9	Store 6	4	201	50.25	34.91667		
10	Store 7	4	283	70.75	104.25		
11	Store 8	4	303	75.75	54.25		
12							
13	Dispenser 1	8	495	61.875	87.83929		
14	Dispenser 2	8	597	74.625	197.125		
15	Dispenser 3	8	645	80.625	279.6964		
16	Dispenser 4	8	539	67.375	155.6964		
17							
18							
19	ANOVA						
20	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
21	Rows	4313.5	7	616.2143	17.75103	1.68E-07	2.487578
22	Columns	1617	3	539	15.52675	1.5E-05	3.072467
23	Error	729	21	34.71429			
24							
25	Total	6659.5	31				

Regression Analysis

One of the favorite Analysis ToolPak procedures is its **Regression** procedure. This offers basically the same interface and results as StatTools and other statistical software packages, but there are some limitations that will be noted below. As a typical example, the file **Overhead Costs.xlsx** contains data on monthly overhead costs, the dependent variable, as well as data on machine hours and production runs, the independent variables (see Figure 34).

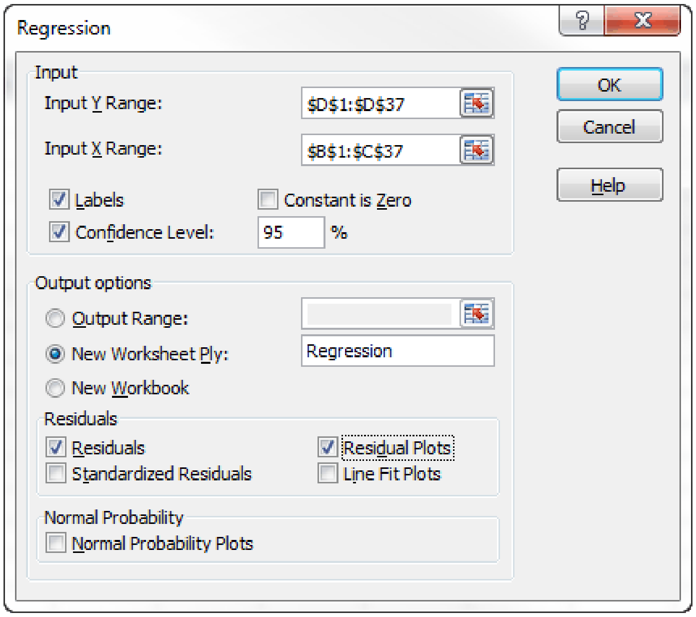
Figure 34 Overhead Cost Data



	A	B	C	D
1	Month	Machine Hours	Production Runs	Overhead
2	1	1539	31	99798
3	2	1284	29	87804
4	3	1490	27	93681
5	4	1355	22	82262
35	34	1723	35	107828
36	35	1413	30	88032
37	36	1390	54	117943

These data are fortunately in the form Analysis ToolPak requires—the independent variables are in contiguous columns—so the Regression dialog box can be filled out as shown in Figure 35. You can decide which of the five check boxes at the bottom to check (including none of them) for diagnostic analysis of the residuals.

Figure 35 Regression Dialog Box



Regression

Input

Input Y Range: $D$1:$D$37

Input X Range: $B$1:$C$37

☑ Labels □ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply: Regression

○ New Workbook

Residuals

☑ Residuals ☑ Residual Plots

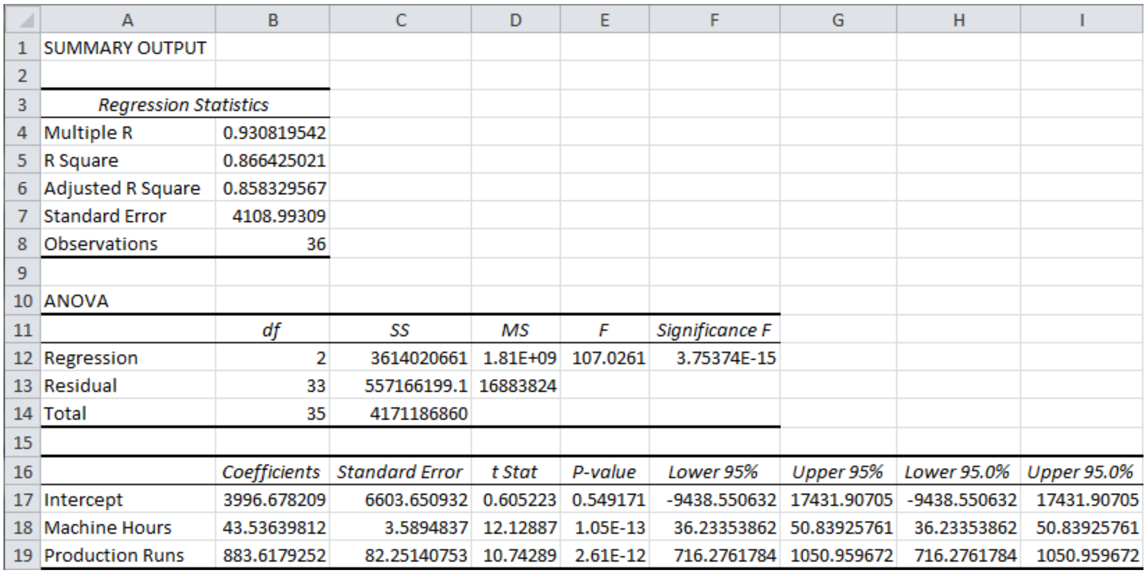
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

The regression output shown in Figure 36 is standard. It includes the regression summary statistics at the top, the ANOVA table for checking whether the regression has any significance as a whole, and the information on the individual regression coefficients. One curious feature is that you automatically get *two* versions of the confidence intervals for the coefficients—and you get them *regardless* of whether you check the Confidence Level box in Figure 35. There is an explanation for this somewhat strange behavior. The first confidence interval is always at the 95% confidence level. The second is at a confidence level of your choice, such as 90%, but only if you check the Confidence Level box.

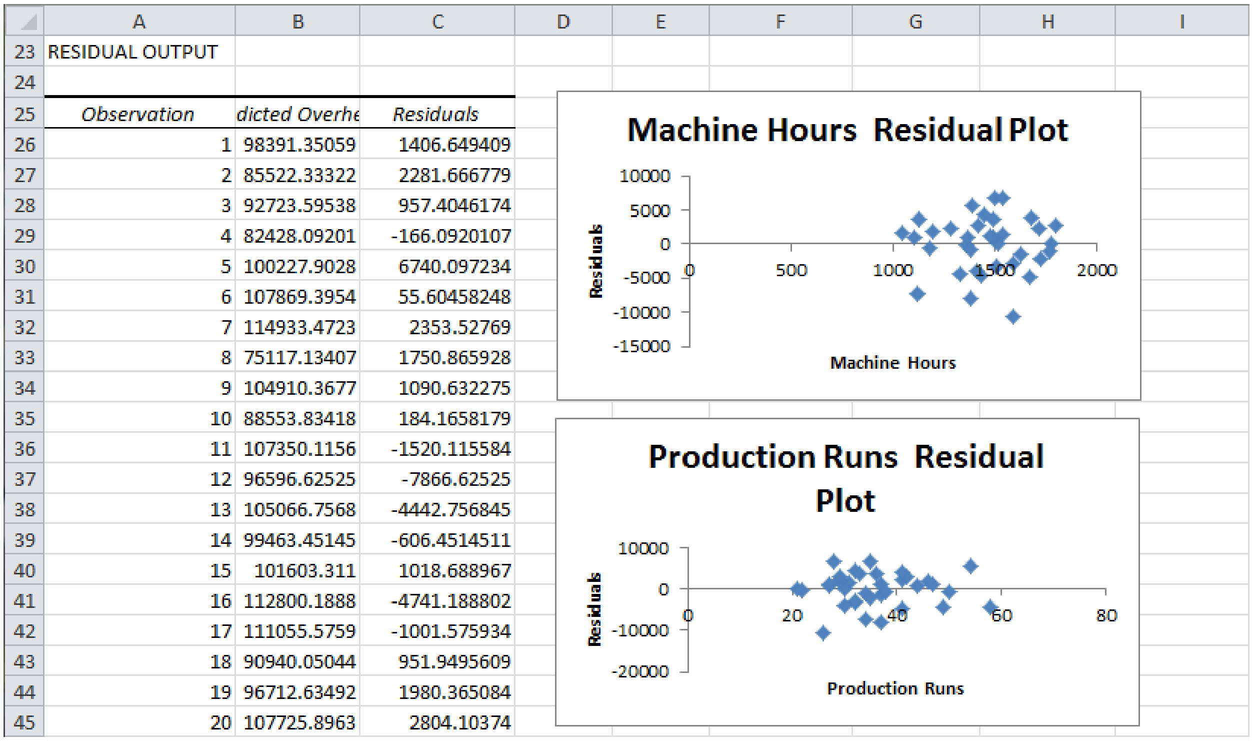
Figure 36 Regression Output



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.930819542							
5	R Square	0.866425021							
6	Adjusted R Square	0.858329567							
7	Standard Error	4108.99309							
8	Observations	36							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	3614020661	1.81E+09	107.0261	3.75374E-15			
13	Residual	33	557166199.1	16883824					
14	Total	35	4171186860						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	3996.678209	6603.650932	0.605223	0.549171	-9438.550632	17431.90705	-9438.550632	17431.90705
18	Machine Hours	43.53639812	3.5894837	12.12887	1.05E-13	36.23353862	50.83925761	36.23353862	50.83925761
19	Production Runs	883.6179252	82.25140753	10.74289	2.61E-12	716.2761784	1050.959672	716.2761784	1050.959672

The residuals and residual plots requested in Figure 35 are shown in Figure 37. (Actually, all charts overlap one another, so you will probably want to move them around.) These residual plots let you see whether there are any obvious violations of the regression assumptions.

Figure 37 Residuals and Residual Plots

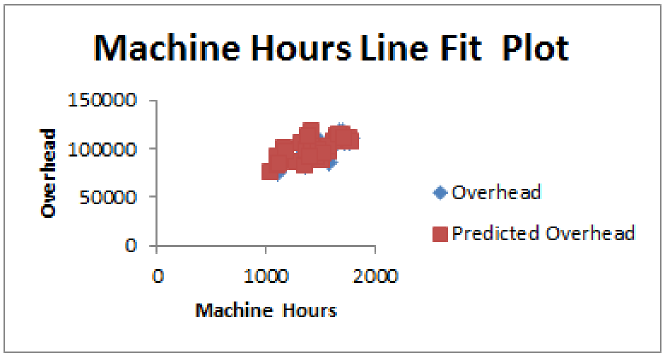


RESIDUAL OUTPUT

Observation	*dicted Overhe*	*Residuals*
1	98391.35059	1406.649409
2	85522.33322	2281.666779
3	92723.59538	957.4046174
4	82428.09201	-166.0920107
5	100227.9028	6740.097234
6	107869.3954	55.60458248
7	114933.4723	2353.52769
8	75117.13407	1750.865928
9	104910.3677	1090.632275
10	88553.83418	184.1658179
11	107350.1156	-1520.115584
12	96596.62525	-7866.62525
13	105066.7568	-4442.756845
14	99463.45145	-606.4514511
15	101603.311	1018.688967
16	112800.1888	-4741.188802
17	111055.5759	-1001.575934
18	90940.05044	951.9495609
19	96712.63492	1980.365084
20	107725.8963	2804.10374

In addition, if you check the Line Fit Plots option in Figure 35, you get plots such as the one in Figure 38. They let you see graphically whether the actual overhead values match the predicted values. This is not apparent from the chart in its present form, but you can see it by reducing the sizes of the points.

Figure 38 Line Fit Plot



Because regression is probably one of the Analysis ToolPak procedures you will use most often, it is worth listing its positive and negative features. On the positive side, it provides the same basic outputs (Figures 36-38) in the same basic formats as StatTools and other statistical software packages. But there are definitely some negatives:

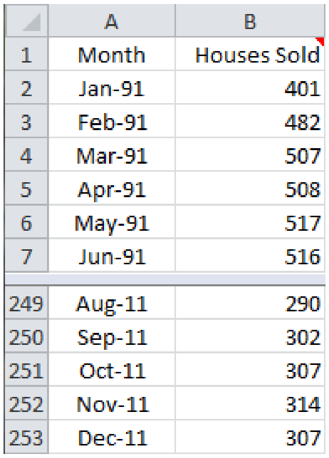
- The independent variables need to be in contiguous columns. This is not a big deal if you plan to run only one regression. In this case, you can rearrange the columns once and for all, if necessary. However, it is a real nuisance if you want to make several regression runs, each with different selections of the independent variables. You will likely need to do a lot of rearranging.
- As always with Analysis ToolPak, you will need to widen the columns in the regression output to see the full labels. Also, the charts, such as the one in Figure 38, will need some reformatting to be of much use.
- Suppose you start with a text variable such as Gender, with values Male and Female, and you want to include Gender as an independent variable. You will need to create dummy variables for such categorical variables with Excel formulas. StatTools also requires dummies for its regression analysis, but at least it has Data Utilities for creating dummies. The same comments apply to nonlinear transformations and interaction variables. You have to create them with Excel formulas if you plan to use them in the regression analysis.
- Unlike StatTools and many other statistical software packages, Analysis ToolPak has no stepwise regression options. This can be a real deal-breaker for many users who rely on stepwise procedures.

Time Series Analysis

Analysis ToolPak provides two standard tools for time series analysis, moving averages and exponential smoothing. (It also includes Fourier Analysis, an advanced engineering tool that isn't discussed here.) Unfortunately, the only exponential smoothing method provided is *simple* exponential smoothing, not the Holt's or Winters' versions provided in StatTools. Therefore, the Analysis ToolPak time series methods apply only to time series that aren't really "going anywhere." They won't apply very well to time series with obvious trends, and they are not relevant at all for time series with seasonality.

These methods will be illustrated with the monthly data in the file **House Sales.xlsx** (see Figure 39). Each value is the number (in thousands) of single new-family homes sold in the United States.

Figure 39 House Sales Data



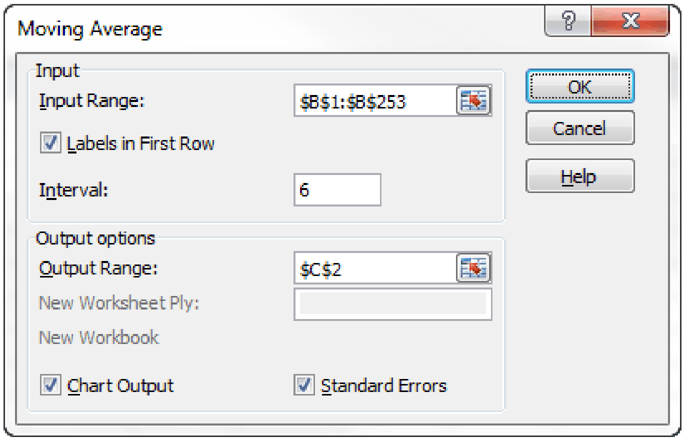
	A	B
1	Month	Houses Sold
2	Jan-91	401
3	Feb-91	482
4	Mar-91	507
5	Apr-91	508
6	May-91	517
7	Jun-91	516
249	Aug-11	290
250	Sep-11	302
251	Oct-11	307
252	Nov-11	314
253	Dec-11	307

Moving Averages

The idea in the moving averages is that you choose a number, called an "interval" in Analysis ToolPak, such as 6 months, and you repeatedly average this many observations. As an example, the average of the values for Jan-91 to Jun-91 is then used as the forecast for the next month, Jul-91. To do this in

Analysis ToolPak, you fill out the Moving Averages dialog box as shown in Figure 40. Again, the “Interval” is the number of monthly values in each average. Note that you have only one option for the location of the output. It turns out that the output will be monthly values, so in this case, it is a good idea to start the output in row 2 and column C, right next to the data.

Figure 40 Moving Averages Dialog Box



Moving Average

Input

Input Range: $B$1:$B$253

☑ Labels in First Row

Interval: 6

Output options

Output Range: $C$2

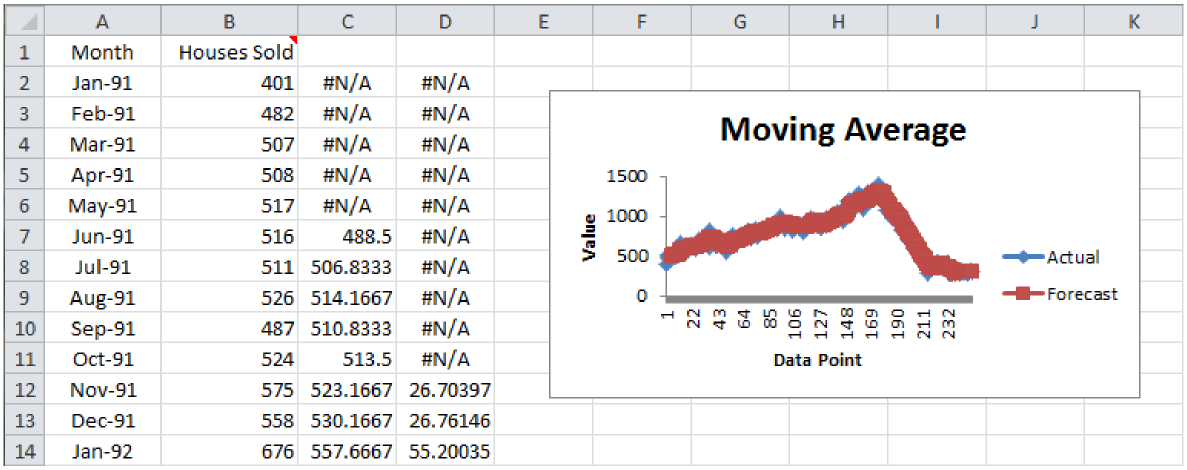
New Worksheet Ply:

New Workbook

☑ Chart Output ☑ Standard Errors

The output appears in Figure 41. Analysis ToolPak doesn’t even provide labels for the outputs in columns C and D, but it is easy to see that the moving averages are in column C and the standard errors are in column D. The first average that can be calculated is the one for Jun-91, the average of Jan-91 to Jun-91. But it is arguably placed in the wrong row. It should really be placed in the Jul-91 row because it is the forecast for Jul-91, not for Jun-91. The standard errors are also questionable. For example, the formula for the first standard error, in row 12, is =SQRT(SUMXMY2(B7:B12,C7:C12)/6). This seems reasonable—the square root of the average of the squared differences between actual sales and forecasts—but again, the forecasts used are the wrong ones. They’re one row off, as just explained. Finally, there is no measure such as RMSE or MAPE for how well the forecasts match the actual values for the entire historical period, and there is no option for obtaining *future* forecasts.[2]

Figure 41 Moving Averages Output



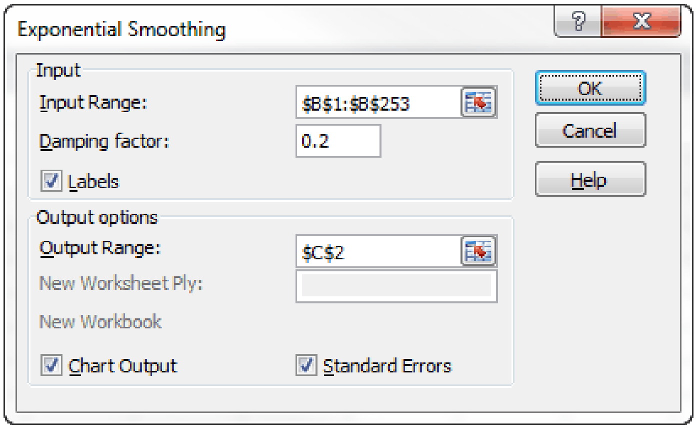
	A	B	C	D
1	Month	Houses Sold		
2	Jan-91	401	#N/A	#N/A
3	Feb-91	482	#N/A	#N/A
4	Mar-91	507	#N/A	#N/A
5	Apr-91	508	#N/A	#N/A
6	May-91	517	#N/A	#N/A
7	Jun-91	516	488.5	#N/A
8	Jul-91	511	506.8333	#N/A
9	Aug-91	526	514.1667	#N/A
10	Sep-91	487	510.8333	#N/A
11	Oct-91	524	513.5	#N/A
12	Nov-91	575	523.1667	26.70397
13	Dec-91	558	530.1667	26.76146
14	Jan-92	676	557.6667	55.20035

[2] RMSE and MAPE stand for root mean square error and mean absolute percentage error, respectively.

Exponential Smoothing

The Exponential Smoothing procedure in Analysis ToolPak is very similar to the Moving Averages procedure. Its dialog box, shown in Figure 42, is identical to the Moving Averages dialog box, except that you now enter a damping factor between 0 and 1. However, this damping factor is actually one minus the "smoothing constant" requested by StatTools and most other statistical software packages. So if you want a low smoothing constant such as 0.2, which is usually recommended, you need to enter a damping factor of 0.8.

Figure 42 Exponential Smoothing Dialog Box



Exponential Smoothing

Input

Input Range: $B$1:$B$253

Damping factor: 0.2

☑ Labels

Output options

Output Range: $C$2

New Worksheet Ply:

New Workbook

☑ Chart Output ☑ Standard Errors

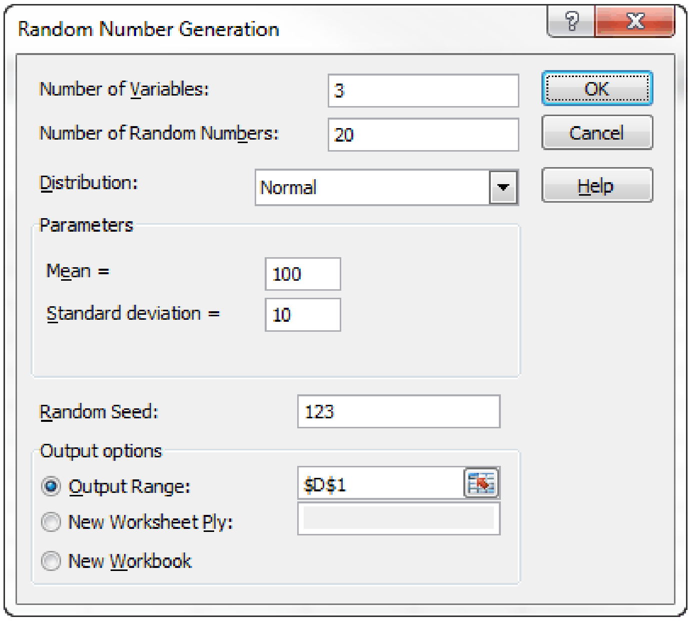
The exponential smoothing output appears in Figure 43. Again, forecasts are listed in column C and standard errors are listed in column D. The first forecast (the "initialization"), for Feb-91, is the actual Jan-91 value. Then the typical formula in column C, the one for Mar-91, is =0.2*B3+0.8*C3, which is copied down. It is a weighted average of the actual Feb-91 value and the forecast for Feb-91. This is the *correct* forecast for Mar-91. It is not off by one row as with moving averages. The typical standard error formula, the one for May-91, is =SQRT(SUMXMY2(B3:B5,C3:C5)/3). (No rationale is provided for averaging *three* squared differences.) Finally, there is again no measure such as RMSE or MAPE for how well the forecasts match the actual values for the entire historical period, and there is no option for obtaining *future* forecasts.

Random Number Generation

Analysis ToolPak has a **Random Number Generation** tool for generating one or more columns of random numbers from any of several probability distributions: Uniform, Normal, Bernoulli, Binomial, Poisson, Patterned, and Discrete.[3] It isn't absolutely clear why you would want to use this tool. The problem is that the random numbers are static; they don't change when you press the F9 key to recalculate. Therefore, they aren't useful for simulation models, the usual situation where you require random numbers. In any case, Figures 43-46 illustrate how you can use this tool.

[3] The Patterned option is a strange one to include here. Its results aren't random at all. It is more like what you get from Excel's Fill Series command.

Figure 43 Random Number Dialog Box for Normal Distribution



Random Number Generation

Number of Variables: 3

Number of Random Numbers: 20

Distribution: Normal

Parameters

Mean = 100

Standard deviation = 10

Random Seed: 123

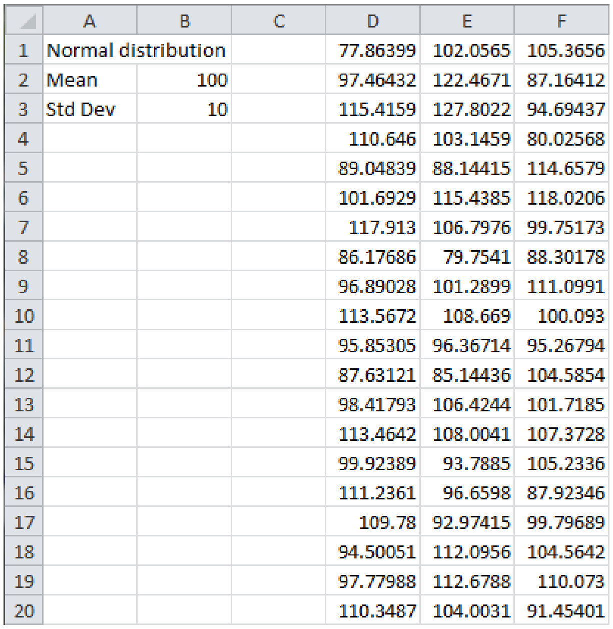
Output options

● Output Range: $D$1

○ New Worksheet Ply:

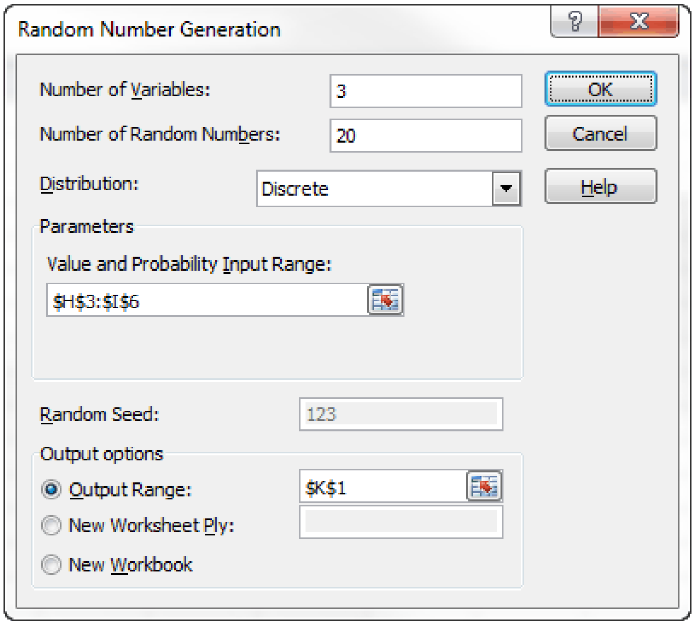
○ New Workbook

Figure 44 Normal Random Numbers



	A	B	C	D	E	F
1	Normal distribution			77.86399	102.0565	105.3656
2	Mean	100		97.46432	122.4671	87.16412
3	Std Dev	10		115.4159	127.8022	94.69437
4				110.646	103.1459	80.02568
5				89.04839	88.14415	114.6579
6				101.6929	115.4385	118.0206
7				117.913	106.7976	99.75173
8				86.17686	79.7541	88.30178
9				96.89028	101.2899	111.0991
10				113.5672	108.669	100.093
11				95.85305	96.36714	95.26794
12				87.63121	85.14436	104.5854
13				98.41793	106.4244	101.7185
14				113.4642	108.0041	107.3728
15				99.92389	93.7885	105.2336
16				111.2361	96.6598	87.92346
17				109.78	92.97415	99.79689
18				94.50051	112.0956	104.5642
19				97.77988	112.6788	110.073
20				110.3487	104.0031	91.45401

Figure 45 Random Number Dialog Box for Discrete Distribution


Random Number Generation

Number of Variables: 3

Number of Random Numbers: 20

Distribution: Discrete

Parameters

Value and Probability Input Range: $H$3:$I$6

Random Seed: 123

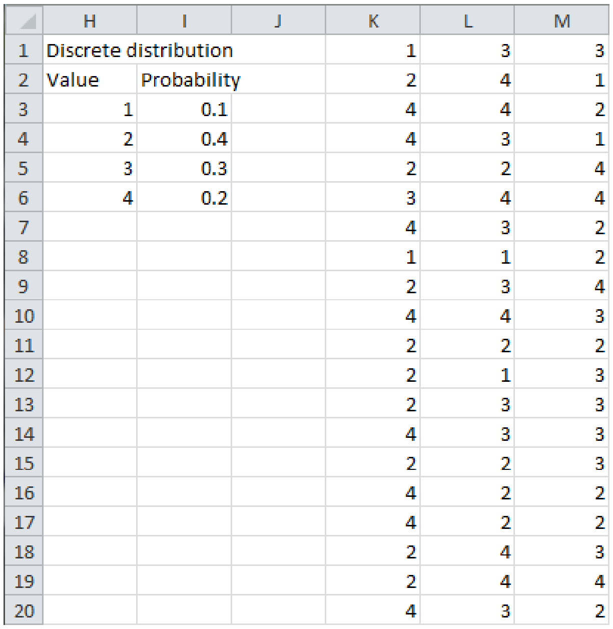
Output options

● Output Range: $K$1

○ New Worksheet Ply:

○ New Workbook

Figure 46 Discrete Random Numbers



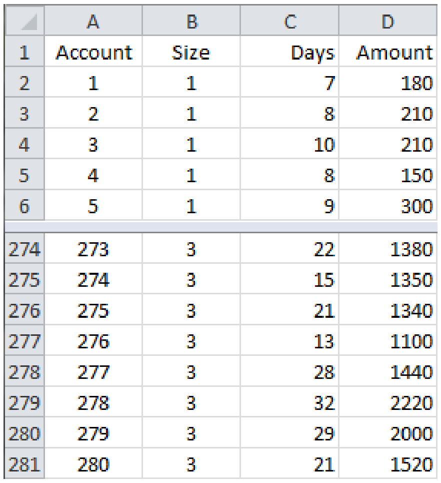
	H	I	J	K	L	M
1	Discrete distribution			1	3	3
2	Value	Probability		2	4	1
3	1	0.1		4	4	2
4	2	0.4		4	3	1
5	3	0.3		2	2	4
6	4	0.2		3	4	4
7				4	3	2
8				1	1	2
9				2	3	4
10				4	4	3
11				2	2	2
12				2	1	3
13				2	3	3
14				4	3	3
15				2	2	3
16				4	2	2
17				4	2	2
18				2	4	3
19				2	4	4
20				4	3	2

Sampling

The **Sampling** tool in Analysis ToolPak allows you to choose a random sample from a larger "population" of values.[4] As an example, the file **Accounts Receivable.xlsx** contains four columns and 280 records of data (see Figure 47). Of course, this "population" could be considerably larger. In any case, you might want to choose a random sample from these 280 records. The Sampling tool lets you do so—sort of.

[4] It also allows you to select a periodic sample, such as every 10th observation.

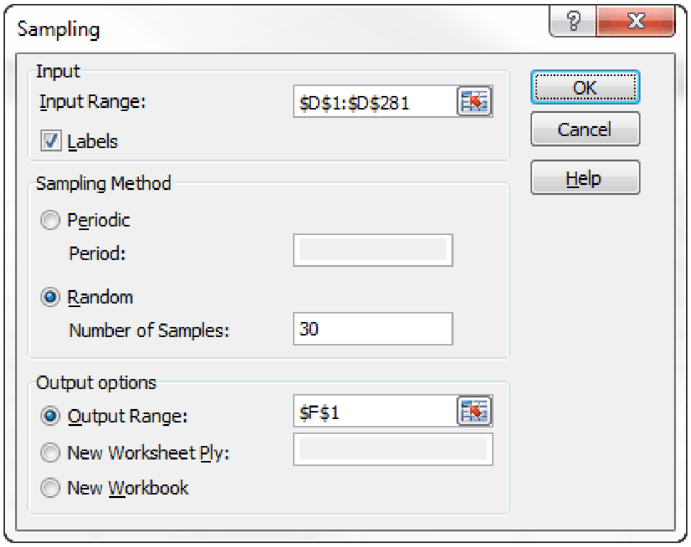
Figure 47 Accounts Receivable Data



	A	B	C	D
1	Account	Size	Days	Amount
2	1	1	7	180
3	2	1	8	210
4	3	1	10	210
5	4	1	8	150
6	5	1	9	300
274	273	3	22	1380
275	274	3	15	1350
276	275	3	21	1340
277	276	3	13	1100
278	277	3	28	1440
279	278	3	32	2220
280	279	3	29	2000
281	280	3	21	1520

The Sampling dialog box appears in Figure 48. If you want 30 randomly selected full records—all four columns—you would naturally specify the Input Range as $A$1:$D$281. However, this not only doesn't work, but it gives the misleading error message "Sampling – Input range contains non-numeric data." Evidently, Analysis ToolPak requires a *single* column for its input range, as in Figure 48. Then the output, not shown here, is a list of 30 randomly selected numbers from column D. There are no "IDs" for these 30 numbers, so you don't know which records they came from and you don't what the corresponding Size and Days values are.

Figure 48 Sampling Dialog Box



Sampling

Input

Input Range: $D$1:$D$281

☑ Labels

Sampling Method

○ Periodic

Period:

● Random

Number of Samples: 30

Output options

● Output Range: $F$1

○ New Worksheet Ply:

○ New Workbook

Conclusion

Perhaps I have been too negative about the Analysis ToolPak add-in. It clearly lacks many features and the overall professional look of other statistical software packages, including StatTools. However, if you want to perform standard statistical analyses and you have only Excel, you can certainly get by with Analysis ToolPak. You might have to rearrange data, widen output columns, reformat graphs, and possibly a few other things, but you will be able to get the basic results you need fairly quickly and easily. The curious aspect of Analysis ToolPak—and I don’t have an answer for this—is why Microsoft hasn’t spent the minimal development time it would take to improve this add-in. Will it look exactly the same in another 10 or 20 years? Hopefully not, but time will tell.