Introduction to Excel 2016 with Data Analysis Toolpak

The Data Analysis Toolpak is a free Excel Add-in that provides extended data analysis beyond the built-in functions of Excel. These functions will help save steps and time to develop complex statistical analyses. The Data Analysis Toolpak is included when Microsoft Office or Excel is installed, but it is not enabled by default. To check if the Data Analysis Toolpak is already enabled, click on the Data tab (A below). The Data Analysis group (B below) will appear on the end of the Ribbon’s Data tab.

If the Data Analysis Toolpak does not appear as shown above, follow the steps in Load Analysis Toolpak.

Load Analysis Toolpak

1. Click the File tab (1a above) and then click Options (1b below) in the left pane.

2. The Options dialog box will open. Click the Add-Ins tab in the left pane (2a). Select Excel Add-ins from the Manage dropdown list (2b), click Go button (2c).

3. The Add-Ins dialog box will open. Check the Analysis ToolPak box (3a below) and then click OK.

4. If an alert dialog box states that Excel cannot run the selected add-in and prompts you to install it, click Yes.
Introduction to Excel 2016 with Data Analysis Toolpak: Common Procedures

Open an Analysis Tool Dialog Box
Open an existing data file or enter data into a worksheet. The data can be arranged into columns or rows. You can also include a text label in the first cell of a column or row or format your data as a table. Follow these steps to open a Data Analysis tool:

1. Select Data>Data Analysis. This will open the Data Analysis dialog box, which contains the list of statistical analysis tools (e.g., Anova:Single Factor, Descriptive Statistics, and Histogram).
2. Select an analysis tool from the list (right).
3. Click OK. A dialog box for the selected tool will appear.

Select Input Range (I)
Three alternative procedures may be used to select the range of cells containing data in Analysis Tools:

1. Place the cursor inside the Input Range box and highlight data with the mouse. Cell references will automatically appear in the Input Range box, I2.
2. Click the button at the right of the Input Range box to collapse window, select data with the mouse, and click the button at the right of window to return, I3.
3. Enter the cell references directly into the Input Range box.

TIP: If first cell contains labels, be sure to check the box next to “Labels in First Row”.

Designate Output Range (O)
There are three options for location of output in Analysis Tools:

1. To display the output in the same worksheet containing the data, select Output Range and use one of the steps above to select where to place the output.
2. To display the output in a new worksheet in the current workbook, select New Worksheet Ply. To name the new worksheet, place the cursor in the New Worksheet Ply box and enter the desired name.
3. To display the output in a new worksheet in a new workbook (that is, a new Excel file), select New Workbook.

Chart Tools
When working with graphs in Excel, format the graph’s colors, lines, layout, labels, and other options using Chart Tools: Design and Format menus.

- To access these menus, click anywhere on a chart. Double click on the chart to get the format dialog box on the right side of the screen (see right).
- Options available differ depending on the graph, but can be found at the top of the dialog box.
Generating Random Numbers

1. Select Data>Data Analysis>Random Number Generation.  
2. In the Number of Variables box, enter the number of columns containing the random numbers that will be generated.
3. In the Number of Random Numbers box, enter the number of random numbers desired in each column.
4. From the Distribution drop-down menu (right), select a distribution from which the random numbers will be generated.
5. Under Parameters, enter a parameter value(s) that characterizes the selected distribution. For example, you will need to enter a lower bound value and an upper bound value if the Uniform distribution is selected. For the Normal distribution, you will be prompted to enter a Mean value and a Standard deviation value.
6. In the Random Seed box, follow one of these options:  
a. Leave the box blank or enter 0. This will generate different sequences of random numbers each time you use the Random Number Generation tool.  
b. Enter a positive integer value. This will generate the same sequence of random numbers each time you use the Random Number Generation tool.
7. Follow the steps to Designate Output Range.  
8. Click OK. The dialog box will generate the output table of random numbers as shown below right.

TIP: Copy and paste random number values otherwise they will change each time the worksheet is updated.

Sampling

1. Select Data>Data Analysis>Sampling>OK.
2. Follow the steps to Select Input Range. The data must be numerical and arranged in either a single column or single row containing at least four cells of data. The selected data range is treated as a population.
3. If the column (or row) of data contains a label, select the Labels checkbox.
4. Under Sampling Method, select Random and enter a number for a sample size in the Number of Samples box.
5. Follow the steps to Designate Output Range.  
6. Click OK. This will generate a column containing a random sample of the size specified.

TIP: Random Sampling is with replacement, values may be selected more than once.
Descriptive Statistics and Confidence Intervals for a Mean

1. Select Data Analysis>Descriptive Statistics>OK.
2. Follow the steps to Select Input Range.
3. Follow one of these procedures:
   a. If the data are arranged in columns, select Columns. If the data contain a label in the first row of each column, select the Labels in First Row checkbox.
   b. If the data are arranged in rows, select Rows. If the data contain a label in the first column of each data row, select the Labels in First Column checkbox.
4. Follow the steps to Designate Output Range.
5. Check the Summary statistics box.
6. Select other summaries to include in a row in the output table by checking the box next to:
   a. Confidence Level for Mean, enter a confidence level (in percentage).
   b. Kth Largest, enter a number to use for K. For example, K=1 will provide the maximum value of the data set.
   c. Kth Smallest, enter a number to use for K. For example, K=1 will provide the minimum value of the data set.
7. Click OK. Note that in the output the Confidence Level (95%) value is a margin of error used to construct a 95% confidence interval. This value should be subtracted and added to the mean in order to obtain a confidence interval.

Histogram

1. Select Data Analysis>Histogram>OK. This will open the Histogram dialog box.
2. Follow the steps to Select Input Range.
3. If the data contain a label, select the Labels checkbox.
4. Follow one of these procedures:
   a. To specify bins (ranges) automatically, leave the Bin Range box blank. Excel will divide the range of the data from the maximum to the minimum into equally spaced bins.
   b. To manually specify bins, enter a set of boundary values in a column or row, and in ascending order. If the Labels box was checked in step 3, you must have a label for the set of boundary values. Note the ranges are exclusive (>) at the minimum value and inclusive (>=) at the maximum value.
5. Follow the steps to Designate Output Range.
6. Select the Chart Output checkbox. This will generate a histogram and a frequency table (right). Notice that the frequency of 10 (right) is the number of data points whose values are greater than 1000 and less than or equal to 2000.
7. Select one or both of the following checkboxes as an option(s):
   a. Pareto (sorted histogram): This will generate an output table and a histogram. The output table contains bin frequencies arranged in descending order. The histogram bars are sorted in descending order of frequency.
   b. Cumulative Percentage: This will generate a column of cumulative percentages in the frequency table. If the selection is made together with Pareto (sorted histogram), the histogram chart will include a cumulative percentage line.
8. Click OK. Click anywhere on the histogram to edit using the Chart Tools, Design and Format tabs (below).
   • To remove the gaps between bars, right-click on the bars in the chart and choose Format Data Series from the menu. This will open the Format Data Series dialog (below, right). Move the Gap Width slider to “0%”.
   • Many other options are available in the Format Data Series dialog (below, far right).
Scatterplot and Regression Line

1. Highlight two adjacent columns containing numerical data and labels in the first row. The first column will be on the x-axis, the second on the y.
2. From the Charts group in the Insert menu, select the Scatter option to generate a scatter plot (below left).
3. To add a regression line, click anywhere on the scatter plot to show the Chart Tools.
4. Click on the Design tab. From the Add Chart Element menu, select Trendline, then Linear (below center).
5. To make changes to the trendline, open the Format Trendline dialog box by double clicking on the trend line in the graph. Find Trendline Options in the sub-menu of the trendline dialog (below right).
   a. Rename the regression line
   b. Display a regression equation on the chart
   c. Display the R^2 value on the chart

Simple Regression, Multiple Regression, and Residuals Plots

1. Select Data>Data Analysis>Regression>OK. The Regression dialog box will appear:
2. For the Input Y (dependent variable) Range box, follow the steps to Select Input Range.
3. For the Input X (independent variable) Range box, follow the steps to Select Input Range. Note: For multiple regression, the data for the independent variables must be arranged in adjacent columns.
4. If the data columns contain labels in the first row, select the Labels checkbox.
5. To exclude an intercept in a regression model, select the Constant is Zero checkbox.
6. By default the summary output table includes 95% confidence intervals for the slope parameter and the intercept parameter. To include additional confidence intervals with a specified confidence level in the summary output table, select the Confidence Level checkbox and enter a confidence level.
7. Follow the steps to Designate Output Range.
8. To generate an output table containing predicted values and residuals, select the Residuals checkbox.
9. To generate an output table containing predicted values, residuals, and standardized residuals select the Standardized Residuals checkbox.
10. To generate a chart for each independent variable versus the residuals, select the Residual Plots checkbox. This will also generate an output table containing predicted values and residuals.
11. To generate a chart for predicted values versus the observed values, select the Line Fit Plots checkbox. This will also generate an output table containing predicted values and residuals.
Correlation Matrix and Variance-Covariance Matrix

1. To create a correlation matrix, select Data Analysis>Correlation>OK. This will open the Correlation dialog box (right).
2. To create a variance-covariance matrix, select Data Analysis>Covariance>OK. This will open the Covariance dialog box.
3. Data for both procedures (Correlation and Covariance) must contain at least two adjacent ranges of data arranged in columns or rows; indicate how they are grouped next to Grouped By:
4. Follow the steps to Select Input Range and Designate Output Range.
5. Check the box if data has Labels in First Row.

Time Series Analysis

Moving Average
1. Select Data>Data Analysis>Moving Average>OK. The Moving Average dialog box will appear (right).
2. Follow the steps to Select Input Range. Dependent data must be numerical and contain at least four cells of data.
3. Check the box if Labels in First Row.
4. In the Interval box, enter the moving average length (the number in the preceding series of values used to compute the average). If left blank, the default interval of three will be used.
5. Follow the steps to Designate Output Range.
6. To display a chart comparing the actual and forecast values, select the Chart Output checkbox.

Exponential Smoothing
1. Select Data>Data Analysis>Exponential Smoothing>OK. The Exponential Smoothing dialog box will appear (right).
2. Follow the steps to Select Input Range. Data must be numerical and contain at least four cells of data.
3. Check the box if Labels in First Row.
4. In the Damping factor box, enter a number that will be the exponential smoothing coefficient (weight). If left blank, the default damping factor of 0.3 will be used. A forecast is calculated based on prior period data and prior forecast data. The new forecast is adjusted for the error in the prior forecast. The result of one minus the number used for the damping factor is the coefficient (weight) put on the prior forecast data. To generate smoother and more stable forecasts, choose a value for the Damping factor that is closer to zero.
5. Follow the steps to Designate Output Range.
6. To display a chart comparing the actual and forecast values, select the Chart Output checkbox.
One-Way ANOVA

1. Arrange numerical data into columns or rows that represent different groups (factor levels). The number of the groups must be at least two. To the right is an example of location as a single factor consisting of three areas (three levels).
2. Select Data Analysis> Anova: Single Factor>OK. This will open the Anova: Single Factor dialog box.
3. Follow the steps to Select Input Range.
4. Follow one of these procedures:
   a. If the groups are arranged in columns, select Columns in Grouped By; If the groups are arranged in rows, select Rows.
   b. If the data contain a label in the first row of each group, select the Labels in first row checkbox.
5. Enter a significance level value in the Alpha box.
6. Follow the steps to Designate Output Range.
7. Click OK. This will generate Summary and ANOVA tables.

Two-Way ANOVA

Two-Way ANOVA with Replication

1. Arrange each numerical data sample into a column factor level (e.g., Location) and a row factor level (e.g., Pool). Each factor must have at least two levels with labels and each sample contains more than one replication (e.g., four replications, or rows, for each sample in the table, top right). Data columns and rows must be adjacent.
2. Select Data Analysis> Anova: Two Factor Replication>OK. This will open the Anova: Two-Factor With Replication dialog box (right, center).
3. Follow the steps to Select Input Range.
4. Enter the number of replications in each sample; note that each sample must be equal in size. (For the example above, each sample contains four replications.)
5. Enter a significance level value in the Alpha box.
6. Follow the steps to Designate Output Range.
7. This will generate several Summary of statistics and ANOVA tables (right, center).

Two-Way ANOVA without Replication

Arrange your numerical data in the same fashion as the Two-Way ANOVA with Replication but with only one replication in each sample. From Data Analysis, select the ANOVA: Two-Factor without Replication procedure and follow steps 3–6 for ANOVA: Two-Factor with Replication. This will generate Summary of statistics and ANOVA tables (bottom right).
Hypothesis Tests for Differences between Two Means

**t-Test: Two-Sample Assuming Unequal Variances**

1. Select **Data>Data Analysis>t-Test: Two-Sample Assuming Unequal Variances**.

2. Here, the Input Range box is replaced with the Variable 1 Range and the Variable 2 Range boxes. The variable that is assumed to have the larger population mean should be identified as Variable 1.

3. Follow the steps to **Select Input Range** for each Variable.

4. In the Hypothesized Mean Difference box, enter a nonnegative number for the hypothesized difference in the two population means.

5. If the column (or row) of data contains a label, select the **Labels** checkbox.

6. In the **Alpha** box, enter a value between 0 and 1. This value is a level of significance desired.

7. Follow the steps to **Designate Output Range**.

8. The output shown here will be generated (top right). The \( P(T<=t) \) one-tail value is the \( p \)-value appropriate for an upper one-sided hypothesis test, whereas the \( P(T<=t) \) two-tail value is the \( p \)-value appropriate for a two-sided hypothesis test. Positive (upper tail) critical values are given. The observed value of the t-test statistic is identified as t Stat.

**z-test: Two Sample for Means**

A z-test for two means can be performed by selecting **Data>Data Analysis>z-test: Two Sample for Means**. Follow steps 2–7 for **t-test: Two-Sample Assuming Unequal Variances**. Two additional boxes will appear in the dialog. Variable 1 and Variable 2 variance: known population variance should be entered here.

**t-Test: Two-Sample Assuming Equal Variances**

1. Select **Data>Data Analysis>t-Test: Two-Sample Assuming Equal Variances**. This will open the **t-Test: Two-Sample AssumingEqual Variances** dialog box.

2. Follow steps 2–7 from **t-Test: Two-Sample Assuming Unequal Variances**.

3. With the exception of an additional line labeled Pooled Variance, the output will be similar to the unequal variances output. Note that the \( p \)-values are associated with the t-test statistic calculated using the estimate of Pooled Variance under the assumption of equal variances.

**t-Test: Paired Two Sample for Means**

1. Select **Data>Data Analysis>t-Test Paired Two Samples for Means**. This will open the **t-Test Paired Two Samples for Means** dialog box.

2. Follow steps 2–7 in **t-Test: Two-Sample Assuming Unequal Variances**. Note that in step 2 the number of data points in each range must be the same. Except for an additional line in the output labeled Pearson Correlation, the output will be similar to the **t-Test: Two-Sample Assuming Unequal Variances** output.

**z-test: Two Sample for Means**

A z-test for two means can be performed by selecting **Data>Data Analysis>z-test: Two Sample for Means**. Follow steps 2–7 for **t-test: Two-Sample Assuming Unequal Variances**. Two additional boxes will appear in the dialog. Variable 1 and Variable 2 variance: known population variance should be entered here.

2. In the **Variable 1 Range** box, select the data range drawn from the population with the assumed larger variance.

3. In the **Alpha** box, enter a value between 0 and 1. This value is a level of significance associated with the **F Critical one-tail** value provided in the output.

4. Click **OK**. This will generate output that includes the \( P(F<=f) \) one-tail value, which is the \( p \)-value for the upper one-sided hypothesis test. The alternative hypothesis states that the variance of population 1 is larger than the variance of population 2, where Variable 1 is data selected from population 1, and Variable 2 is data from population 2.

Hypothesis Tests for Differences between Two Variances (Two-Sample F-Tests)

1. Select **Data>Data Analysis>F-test Two-Sample for Variances**. This will open the **F-test Two-Sample for Variances** dialog box.

2. In the **Variable 1 Range** box, select the data range drawn from the population with the assumed larger variance.

Note: The F statistic is the ratio of sample variances with the sample variance of Variable 1 as the numerator and sample variance of Variable 2 as the denominator.