Show All

About statistical analysis tools

Microsoft Excel provides a set of data analysis tools— called the Analysis ToolPak— that you can use to save steps when you develop complex statistical or engineering analyses. You provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

Related worksheet functions Excel provides many other statistical, financial, and engineering worksheet functions. Some of the statistical functions are built-in and others become available when you install the Analysis ToolPak.

Accessing the data analysis tools The Analysis ToolPak includes the tools described below. To access these tools, click **Data Analysis** on the **Tools** menu. If the **Data Analysis** command is not available, you need to load the Analysis ToolPak add-in program.

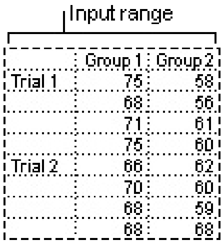
Anova

The Anova analysis tools provide different types of variance analysis. The tool to use depends on the number of factors and the number of samples you have from the populations you want to test.

Anova: Single Factor This tool performs a simple analysis of variance on data for two or more samples. The analysis provides a test of the hypothesis that each sample is drawn from the same underlying probability distribution against the alternative hypothesis that underlying probability distributions are not the same for all samples. If there were only two samples, the worksheet function, TTEST, could equally well be used. With more than two samples, there is no convenient generalization of TTEST and the Single Factor Anova model can be called upon instead.

Anova: Two-Factor With Replication This analysis tool is useful when data can be classified along two different dimensions. For example, in an experiment to measure the height of plants, the plants may be given different brands of fertilizer (for example, A, B, C) and might also be kept at different temperatures (for example, low, high). For each of the 6 possible pairs of {fertilizer, temperature} we have an equal number of observations of plant height. Using this Anova tool we can test:

1. Whether heights of plants for the different fertilizer brands are drawn from the same underlying population; temperatures are ignored for this analysis.
2. Whether heights of plants for the different temperature levels are drawn from the same underlying population; fertilizer brands are ignored for this analysis.
3. Whether having accounted for the effects of differences between fertilizer brands found in step 1 and differences in temperatures found in step 2, the 6 samples representing all pairs of {fertilizer, temperature} values are drawn from the same population. The alternative hypothesis is that there are effects due to specific {fertilizer, temperature} pairs over and above differences based on fertilizer alone or on temperature alone.


Input range

	Group 1	Group 2
Trial 1	75	58
	68	56
	71	61
	75	60
Trial 2	66	62
	70	60
	68	59
	68	68

Anova: Two-Factor Without Replication This analysis tool is useful when data are classified on two different dimensions as in the Two-Factor case With Replication. However, for this tool we assume that there is only a single observation for each pair (for example, each {fertilizer, temperature} pair in the example above. Using this tool we can apply the tests in steps 1 and 2 of the Anova: Two-Factor With Replication case but do not have enough data to apply the test in step 3.

Correlation

The CORREL and PEARSON worksheet functions both calculate the correlation coefficient between two measurement variables when measurements on each variable are observed for each of N subjects. (Any missing observation for any subject causes that subject to be ignored in the analysis.) The Correlation analysis tool is particularly useful when there are more than two measurement variables for each of N subjects. It provides an output table, a correlation matrix, showing the value of CORREL (or PEARSON) applied to each possible pair of measurement variables.

The correlation coefficient, like the covariance, is a measure of the extent to which two measurement variables “vary together.” Unlike the covariance, the correlation coefficient is scaled so that its value is independent of the units in which the two measurement variables are expressed. (For example, if the two measurement variables are weight and height, the value of the correlation coefficient is unchanged if weight is converted from pounds to kilograms.) The value of any correlation coefficient must be between -1 and +1 inclusive.

You can use the correlation analysis tool to examine each pair of measurement variables to determine whether the two measurement variables tend to move together— that is, whether large values of one variable tend to be associated with large values of the other (positive correlation), whether small values of one variable tend to be associated with large values of the other (negative correlation), or whether values of both variables tend to be unrelated (correlation near zero).

Covariance

The Correlation and Covariance tools can both be used in the same setting, when you have N different measurement variables observed on a set of individuals. The Correlation and Covariance tools each give an output table, a matrix, showing the correlation coefficient or covariance, respectively, between each pair of measurement variables. The difference is that correlation coefficients are scaled to lie between -1 and +1 inclusive, Corresponding covariances are not scaled. Both the correlation coefficient and the covariance are measures of the extent to which two variables “vary together.”

The Covariance tool computes the value of the worksheet function, COVAR, for each pair of measurement variables. (Direct use of COVAR rather than the Covariance tool is a reasonable alternative when there are only two measurement variables, i.e. N=2.) The entry on the diagonal of the Covariance tool’s output table in row i, column i is the covariance of the i-th measurement variable with itself; this is just the population variance for that variable as calculated by the worksheet function, VARP.

You can use the covariance tool to examine each pair of measurement variables to determine whether the two measurement variables tend to move together

— that is, whether large values of one variable tend to be associated with large values of the other (positive covariance), whether small values of one variable tend to be associated with large values of the other (negative covariance), or whether values of both variables tend to be unrelated (covariance near zero).

Descriptive Statistics

The Descriptive Statistics analysis tool generates a report of univariate statistics for data in the input range, providing information about the central tendency and variability of your data.

Exponential Smoothing

The Exponential Smoothing analysis tool predicts a value based on the forecast for the prior period, adjusted for the error in that prior forecast. The tool uses the smoothing constant *a*, the magnitude of which determines how strongly forecasts respond to errors in the prior forecast.

Note Values of 0.2 to 0.3 are reasonable smoothing constants. These values indicate that the current forecast should be adjusted 20 to 30 percent for error in the prior forecast. Larger constants yield a faster response but can produce erratic projections. Smaller constants can result in long lags for forecast values.

F-Test Two-Sample for Variances

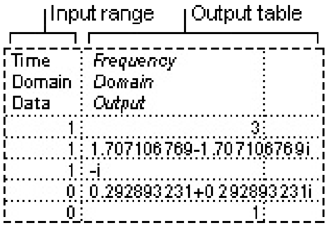
The F-Test Two-Sample for Variances analysis tool performs a two-sample F-test to compare two population variances.

For example, you can use the F-test tool on samples of times in a swim meet for each of two teams. The tool provides the result of a test of the null hypothesis that these two samples come from distributions with equal variances against the alternative that the variances are not equal in the underlying distributions.

The tool calculates the value f of an F-statistic (or F-ratio). A value of f close to 1 provides evidence that the underlying population variances are equal. In the output table, if f < 1 “P(F <= f) one-tail” gives the probability of observing a value of the F-statistic less than f when population variances are equal and “F Critical one-tail” gives the critical value less than 1 for the chosen significance level, Alpha. If f > 1, “P(F <= f) one-tail” gives the probability of observing a value of the F-statistic greater than f when population variances are equal and “F Critical one-tail” gives the critical value greater than 1 for Alpha.

Fourier Analysis

The Fourier Analysis tool solves problems in linear systems and analyzes periodic data by using the Fast Fourier Transform (FFT) method to transform data. This tool also supports inverse transformations, in which the inverse of transformed data returns the original data.


Input range
Output table

Time Domain Data	*Frequency Domain Output*	
1	3	
1	1.707106769-1.707106769i	
1	-i	
0	0.292893231+0.292893231i	
0	1	

Histogram

The Histogram analysis tool calculates individual and cumulative frequencies for a cell range of data and data bins. This tool generates data for the number of occurrences of a value in a data set.

For example, in a class of 20 students, you could determine the distribution of scores in letter-grade categories. A histogram table presents the letter-grade boundaries and the number of scores between the lowest bound and the current bound. The single most-frequent score is the mode of the data.

Moving Average

The Moving Average analysis tool projects values in the forecast period, based on the average value of the variable over a specific number of preceding periods. A moving average provides trend information that a simple average of all historical data would mask. Use this tool to forecast sales, inventory, or other trends. Each forecast value is based on the following formula.

$$F_{(t+1)} = \frac{1}{N}\sum_{j=1}^{N} A_{t-j+1}$$

where:

- N is the number of prior periods to include in the moving average
- A_j is the actual value at time j
- F_j is the forecasted value at time j

Random Number Generation

The Random Number Generation analysis tool fills a range with independent random numbers drawn from one of several distributions. You can characterize subjects in a population with a probability distribution.

For example, you might use a normal distribution to characterize the population of individuals' heights, or you might use a Bernoulli distribution of two possible outcomes to characterize the population of coin-flip results.

Rank and Percentile

The Rank and Percentile analysis tool produces a table that contains the ordinal and percentage rank of each value in a data set. You can analyze the relative standing of values in a data set. This tool uses the worksheet functions, RANK and PERCENTRANK. RANK does not account for tied values. If you wish to account for tied values, use the worksheet function, RANK, together with the correction factor suggested in the help file for RANK.

Regression

The Regression analysis tool performs linear regression analysis by using the "least squares" method to fit a line through a set of observations. You can analyze how a single dependent variable is affected by the values of one or more independent variables.

For example, you can analyze how an athlete's performance is affected by such factors as age, height, and weight. You can apportion shares in the performance measure to each of these three factors, based on a set of performance data, and then use the results to predict the performance of a new, untested athlete.

The Regression tool uses the worksheet function, LINEST.

Sampling

The Sampling analysis tool creates a sample from a population by treating the input range as a population. When the population is too large to process or chart, you can use a representative sample. You can also create a sample that contains only values from a particular part of a cycle if you believe that the input data is periodic.

For example, if the input range contains quarterly sales figures, sampling with a periodic rate of four places values from the same quarter in the output range.

t-Test

The Two-Sample t-Test analysis tools test for equality of the population means underlying each sample. The three tools employ different assumptions: that the population variances are equal, that the population variances are not equal, and that the two samples represent before treatment and after treatment observations on the same subjects.

For all three tools below, a t-Statistic value, t, is computed and shown as “t Stat” in the output tables. Depending on the data, this value, t, can be negative or non-negative. Under the assumption of equal underlying population means, if t < 0, “P(T <= t) one-tail” gives the probability that a value of the t-Statistic would be observed that is more negative than t. If t >=0, “P(T <= t) one-tail” gives the probability that a value of the t-Statistic would be observed that is more positive than t. “t Critical one-tail” gives the cutoff value so that the probability of observing a value of the t-Statistic greater than or equal to “t Critical one-tail” is Alpha.

“P(T <= t) two-tail” gives the probability that a value ot the t-Statistic would be observed that is larger in absolute value than t. “P Critical two-tail” gives the cutoff value so that the probability of an observed t-Statistic larger in absolute value than “P Critical two-tail” is Alpha.

t-Test: Two-Sample Assuming Equal Variances This analysis tool performs a two-sample student's t-test. This t-test form assumes that the two data sets came from distributions with the same variances. It is referred to as a homoscedastic t-test. You can use this t-test to determine whether the two samples are likely to have come from distributions with equal population means.

t-Test: Two-Sample Assuming Unequal Variances This analysis tool performs a two-sample student's t-test. This t-test form assumes that the two data sets came from distributions with unequal variances. It is referred to as a heteroscedastic t-test. As with the Equal Variances case above, you can use this t-test to determine whether the two samples are likely to have come from distributions with equal population means. Use this test when the there are distinct subjects in the two samples. Use the Paired test, described below,when there is a single set of subjects and the two samples represent measurements for each subject before and after a treatment.

The following formula is used to determine the statistic value *t*.

$$t' = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

The following formula is used to calculate the degrees of freedom, df. Because the result of the calculation is usually not an integer, the value of df is rounded to the nearest integer to obtain a critical value from the t table. The Excel worksheet function, TTEST, uses the calculated df value without rounding since it is possible to compute a value for TTEST with a non-integer df. Because of these different approaches to determining degrees of freedom, results of TTEST and this t-Test tool will differ in the Unequal Variances case.

$$df = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{\left(S_1^2 / m\right)^2}{m - 1} + \frac{\left(S_2^2 / n\right)^2}{n - 1}}$$

t-Test: Paired Two Sample For Means You can use a paired test when there is a natural pairing of observations in the samples, such as when a sample group is tested twice— before and after an experiment. This analysis tool and its formula perform a paired two-sample student's t-test to determine whether observations taken before a treatment and observations taken after a treatment are likely to have come from distributions with equal population means. This t-test form does not assume that the variances of both populations are equal.

Note Among the results generated by this tool is pooled variance, an accumulated measure of the spread of data about the mean, derived from the following formula.

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$$

z-Test

The z-Test: Two Sample for Means analysis tool performs a two-sample z-test for means with known variances. This tool is used to test the null hypothesis that there is no difference between two population means against either one-sided or two-sided alternative hypotheses . If variances are not known, the worksheet function, ZTEST, should be used instead.

When using the z-Test tool, one should be careful to understand the output. “P(Z <= z) one-tail” is really P(Z >= ABS(z)), the probability of a z-value further from 0 in the same direction as the observed z value when there is no difference between the population means. “P(Z <= z) two-tail” is really P(Z >= ABS(z) or Z <= -ABS(z)), the probability of a z-value further from 0 in either direction than the observed z-value when there is no difference between the population means. The two-tailed result is just the one-tailed result multiplied by 2. The z-Test tool can also be used for the case where the null hypothesis is that there is a specific non-zero value for the difference between the two population means.

For example, you can use this test to determine differences between the performances of two car models.

Show All

Perform a statistical analysis

1. On the **Tools** menu, click **Data Analysis**.

   If **Data Analysis** is not available, load the Analysis ToolPak.

   How?

   1. On the **Tools** menu, click **Add-Ins**.

   2. In the **Add-Ins available** list, select the **Analysis ToolPak** box, and then click **OK**.

   3. If necessary, follow the instructions in the setup program.
2. In the **Data Analysis** dialog box, click the name of the analysis tool you want to use, then click **OK**.
3. In the dialog box for the tool you selected, set the analysis options you want.

   You can use the **Help** button on the dialog box to get more information about the options.

A bibliography of statistical methods and algorithms

The following book provides detailed information about the algorithms used to create the Microsoft Excel analysis tools and functions.

- Strum, Robert D., and Donald E. Kirk. *First Principles of Discrete Systems and Digital Signal Processing*. Reading, Mass.: Addison-Wesley Publishing Company, 1988.

The following books provide detailed information about statistical methods or algorithms used to create the Microsoft Excel statistical tools and functions.

- Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Washington, D.C.: U.S. Government Printing Office, 1972.
- Box, George E.P., William G. Hunter, and J. Stuart Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: John Wiley and Sons, 1978.
- Devore, Jay L. *Probability and Statistics for Engineering and the Sciences*. 4th ed. Wadsworth Publishing, 1995.
- McCall, Robert B. *Fundamental Statistics for the Behavioral Sciences*. 5th ed. New York: Harcourt Brace Jovanovich, 1990.
- Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. New York: Cambridge University Press, 1992.
- Sokal, Robert R., and F. James Rohlf. *Biometry: The Principles and Practice of Statistics in Biological Research*. 2nd ed. New York: W. H. Freeman, 1995.

Troubleshoot data analysis

Applies to tools in the Analysis ToolPak.

Data appears only on the first worksheet

The data analysis functions can be used on only one worksheet at a time. When you perform data analysis on grouped worksheets, results will appear on the first worksheet and empty formatted tables will appear on the remaining worksheets. To perform data analysis on the remainder of the worksheets, recalculate the analysis tool for each worksheet.