Introduction to Statistical Computing in Microsoft Excel

By Hector D. Flores; hflores@rice.edu, and Dr. J.A. Dobelman

Statistics lab will be mainly focused on applying what you have learned in class with real (or simulated) data. As applied statisticians, we are commonly interested in 3 things: accessing data, analyzing it, and forming reasonable conclusions. Computer software packages, such as Excel, help us with the first and second items. The following brief tutorial will show you some fundamental tools that you will need in this course.

Importing/Accessing Data

Unless you enjoy the painful process of number crunching by hand, it's a good idea to get your data into a computer with programs to make these calculations easier.

Common problem: Most often data does not come in a format that is readily accessible to you. Since we are using Excel, the best-case scenario will be if the data is in Excel format already. However, for the sake of education, suppose we have the "next-best-case" scenario of the data in *delimited text file* format. An example of this is the following (see excel_sample.txt on the web page):

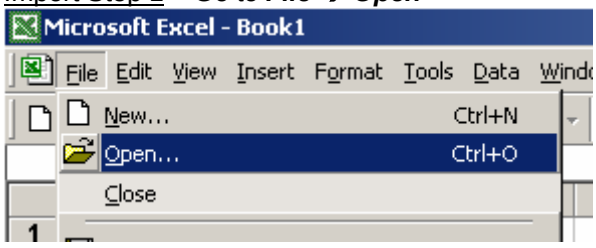
```
Count Time  Brake
1     1.6   1.9
2     2.3   2.9
3     9.5   1.110
4     1.2   2.34
```

This is the case where the text file is "tab-delimited". That is, "tab" spacing separates the data. There are other forms of delimiters: commas, semicolons, asterisks, etc. Typically, most reasonable people will separate their data in a logical fashion (and you should expect data like this in this course).

So what do you do if the data you get in real life is not one of these scenarios? Well, there are ways to deal with it, but that's beyond the scope of this tutorial. Ask me if you really want to know, or have a problem.

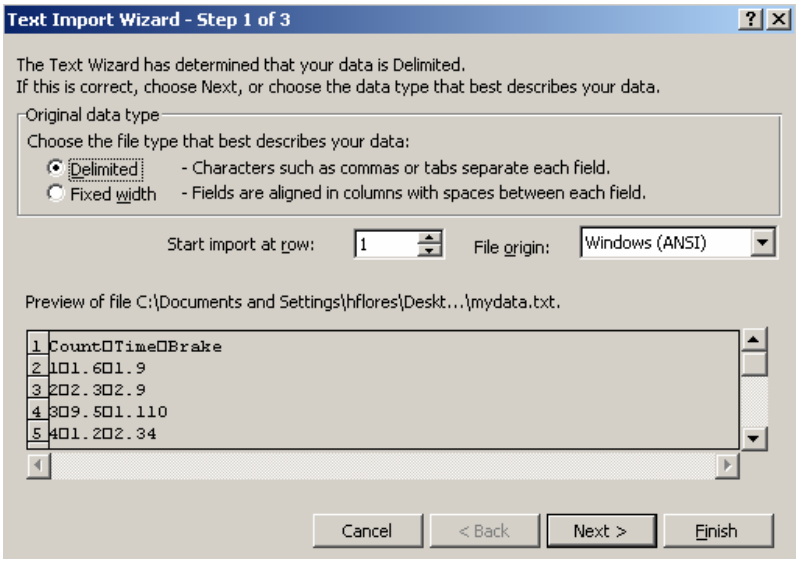
Now, assuming your data is "delimited" in some way, Excel loves you. You can import the data using the following steps (**Try this with excel_sample.txt**):

Import Step 1 – Go to ***File → Open***

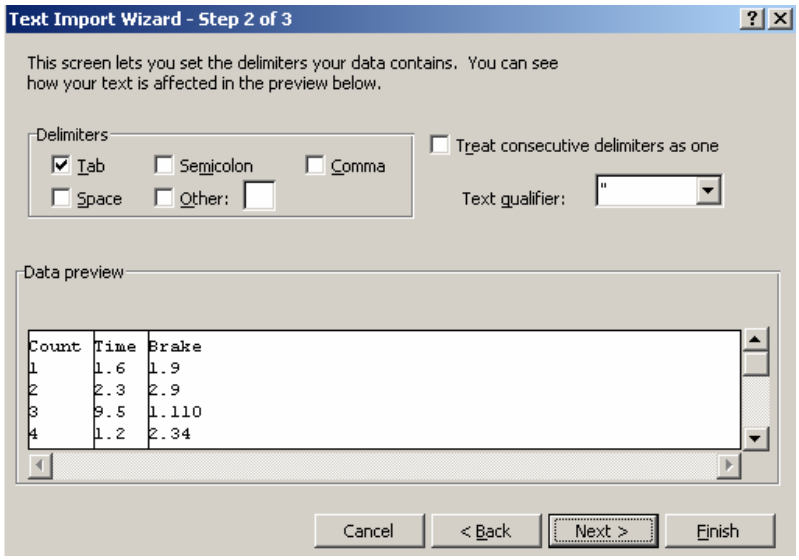


Import Step 2 – Find your file, and click *Open*. Note: You may have to change *Files of type* to *All Files* to see your file.

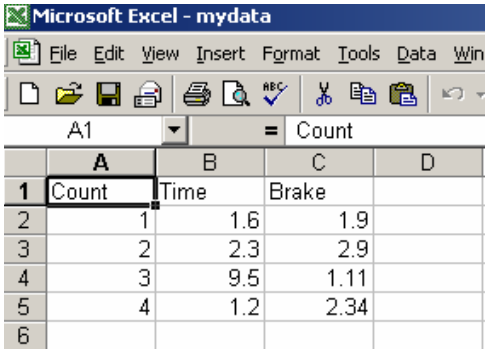
Import Step 3 – The import wizard will appear on the screen.



Honestly, you can mess around with the settings here till you get the desired result in the *Preview* window. Since I know my file is delimited, I make sure it is selected and click *Next*.



You should see the columns line up correctly (see above picture) in the *Data preview*. Clicking on *Next* or *Finish* here will import the data.



Yeah.

Data Analysis

Now that you have data in Excel, what do we do with it?

Answer: Compute statistics with relative ease.

First some notes about Excel. Each cell can hold an object: a character string, a number, an equations, picture, etc. We will mostly be concerned with equations. To enter an equation in any empty cell, first type "=" and then type the desired expression.

Example: To add cells A2 and A3, click on an open cell (where you want the result to be) and enter "= A2 + A3" (and hit the enter key or click away from the cell). The result should be there. Failure to type the "=" will result in the text "A2 + A3". Try making other equations yourself.

Trick #1: Suppose we wanted to add cell 2 and cell 3 from each column (not just A as in the above example). Assuming that you've tried the above example, click on the cell with the "=A2+A3" equation in it. Copy this equation (*CTRL*+*C*) and paste the equation (*CTRL*+V) in the next cell to the right. Now look at the equation in the equation bar. It should read "=B2+B3". Experiment with this idea, moving to other cells.

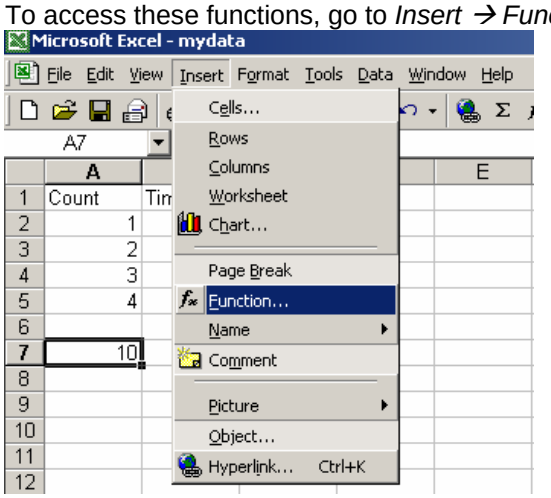
Trick #2: Suppose you want to add all the numbers in a particular column, but don't want to burden yourself with typing all the cell identifiers. Click on the cell you want to put the formula and type "= SUM(". Then, move your mouse to the first element, click-and-drag to the last element you wish to add, and type <*enter*>. Experiment with this "click-and-drag" technique with other formulas.

Trick #3: Suppose that you want to move formulas back and forth as in Trick 1, but you don't want one of the values to move. For example, suppose you want to copy the formula over, and keep the formula saying "=A2+A3" (instead of the default, which changes the letters and numbers). Simply place "$" in front of those letters that you want to remain constant (e.g. "=A2+A3" can becomes "=$A$2+$A$3" to hold the entire equation constant). Experiment with this to get the hang of it.

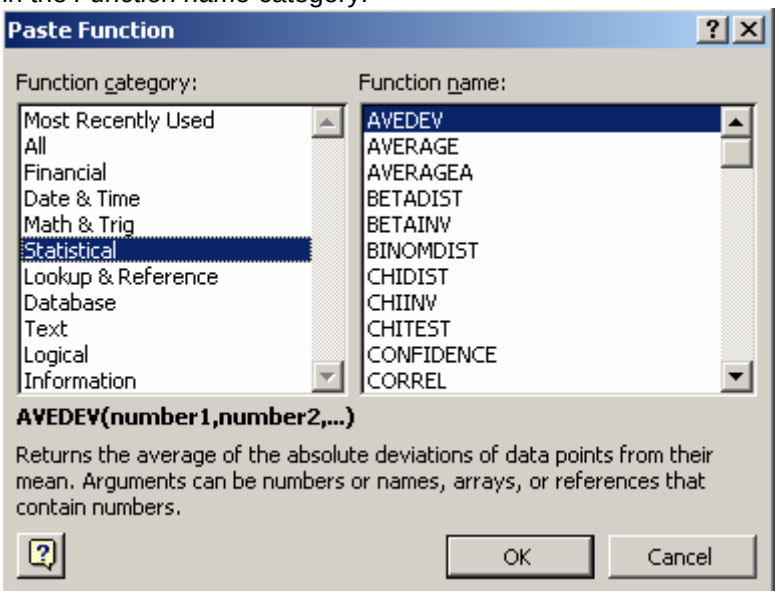
So where are the statistics?

Well, hopefully by now you've learned how to compute such statistics as the mean, median, mode, range, IQR, etc. You can physically enter these formulas into particular cells manually…or…you can cheat and use the built-in functions provided by Excel.

To access these functions, go to *Insert* → *Function*



Then in the *Function Category*, select *Statistical*. You can then choose from any of the functions in the *Function name* category.



Examples:

AVERAGE(A1,A2)
AVERAGE(A1:A10) [average rows 1-10 in column A]
MAX(A2:A5,B2:B5,C2:C5) [gives the largest number of all these cells]
MIN(A2:A5,B2:B5,C2:C5) [the smallest]
MEDIAN(A1:A9) [the median]
MODE(B1:B100) [the mode]
STDEV(B1:B100) [standard deviation]

Data Analysis – Statistics Add-ins. Excel also has an add-in called "Data Analysis" which performs various mathematical tasks such as:

Analysis ToolPak: Adds financial, statistical, and engineering analysis tools and functions.

Analysis ToolPak VBA: Allows users to publish financial, statistical, and engineering analysis tools and functions using Analysis ToolPak syntax.

Conditional Sum Wizard: Creates a formula that sums data in a list if the data matches criteria you specify.

Euro Currency Tools: Formats values as euros, and provides the EUROCONVERT worksheet function to convert currencies.

Internet Assistant VBA: Allows developers to publish Excel data to the Web by using Internet Assistant syntax.

Lookup Wizard: Creates a formula to look up data in a list by using another known value in the list.

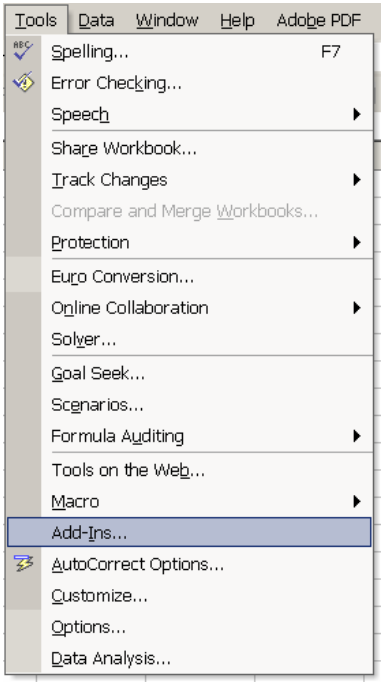
Solver Add-In: Calculates solutions to what-if scenarios based on adjustable cells and constraint cells.

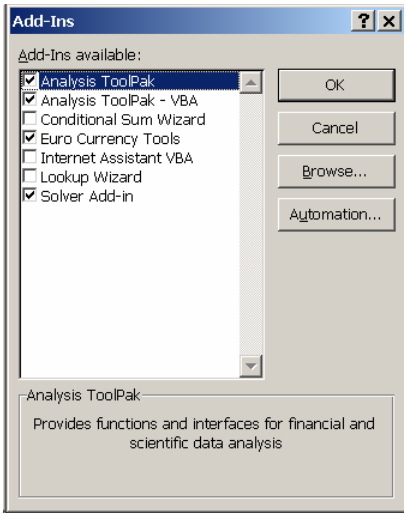
Accessing the data analysis tools The Analysis ToolPak includes the tools described below. To access these tools, click Data Analysis on the Tools menu. If the Data Analysis command is not available, you need to load the Analysis ToolPak add-in program.

To Perform a statistical analysis:

On the Tools menu, click Data Analysis. If Data Analysis is not available, load the Analysis ToolPak.

How? On the Tools menu, click Add-Ins.
In the Add-Ins available list, select the Analysis ToolPak box, and then click OK.



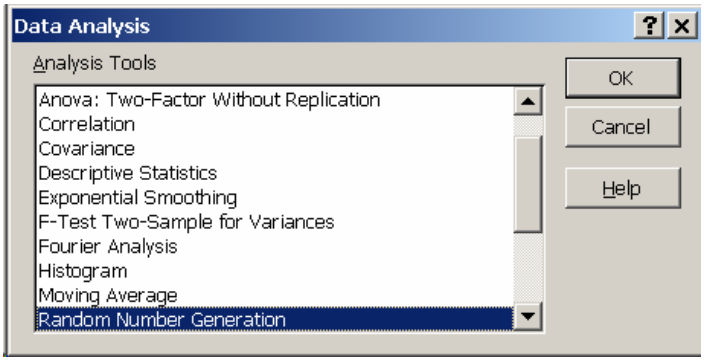


If necessary, follow the instructions in the setup program.

You might have to insert your original Distribution CD, but most likely not.

Once this is done, then you should be able to access the data analysis without having to restart your computer or having to restart Excel.

In the Data Analysis dialog box, click the name of the analysis tool you want to use, then click OK.



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.

The **Analysis ToolPak** saves 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.

Anova

Correlation

Covariance

Descriptive Statistics

Exponential Smoothing

F-Test Two-Sample for Variances

Fourier Analysis

Histogram

Moving Average

Random Number Generation

Rank and Percentile

Regression

Sampling

t-Test

z-Test

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, testing the hypothesis that means from two or more samples are equal (drawn from populations with the same mean). This technique expands on the tests for two means, such as the t-test.

Anova: Two-Factor With Replication This analysis tool performs an extension of the single-factor anova that includes more than one sample for each group of data.

Anova: Two-Factor Without Replication This analysis tool performs a two-factor anova that does not include more than one sampling per group, testing the hypothesis that means from two or more samples are equal (drawn from populations with the same mean). This technique expands on tests for two means, such as the t-test.

Correlation. The Correlation analysis tool measures the relationship between two data sets that are scaled to be independent of the unit of measurement. The population correlation calculation returns the covariance of two data sets divided by the product of their standard deviations based on the following formulas. You can use the correlation analysis tool to determine whether two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive correlation), whether small values of one set are associated with large values of the other (negative correlation), or whether values in both sets are unrelated (correlation near zero).

Note To return the correlation coefficient for two cell ranges, use the CORREL worksheet function.

Covariance. Covariance is a measure of the relationship between two ranges of data. The Covariance analysis tool returns the average of the product of deviations of data points from their respective means, based on the following formula. You can use the covariance tool to determine whether two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive covariance), whether small values of one set are associated with large values of the other (negative covariance), or whether values in both sets are unrelated (covariance near zero).

Note To return the covariance for individual data point pairs, use the COVAR worksheet function.

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 an F-test to determine whether the time scores in a swimming meet have a difference in variance for samples from two teams.

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.

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.

Random Number Generation. (Simulation, or monte carlo). 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.

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.

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 t-Test analysis tools test the means of different types of populations.

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 means of both data sets are equal; it is referred to as a homoscedastic t-test. You can use t-tests to determine whether two sample means are equal.

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 variances of both ranges of data are unequal; it is referred to as a heteroscedastic t-test. You can use a t-test to determine whether two sample means are equal. Use this test when the groups under study are distinct. Use a paired test when there is one group before and after a treatment.

t-Test: Paired Two Sample For Means This analysis tool and its formula perform a paired two-sample student's t-test to determine whether a sample's means are distinct. This t-test form does not assume that the variances of both populations are equal. 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.

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 hypotheses about the difference between two population means. For example, you can use this test to determine differences between the performances of two car models.