Introduction to Statistical Computing in Microsoft Excel

Tutorial 1

As data scientist, you 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:

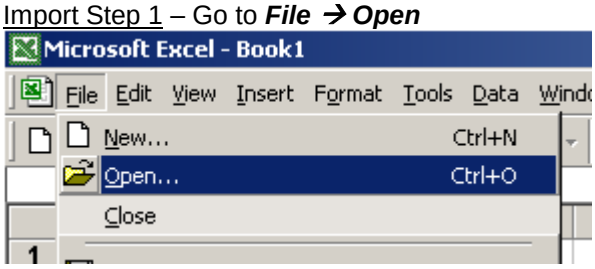
Count	Time	Brake
1	1.6	1.9
2	2.3	2.9
3	9.5	1.110
4	1.2	2.34

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.

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.

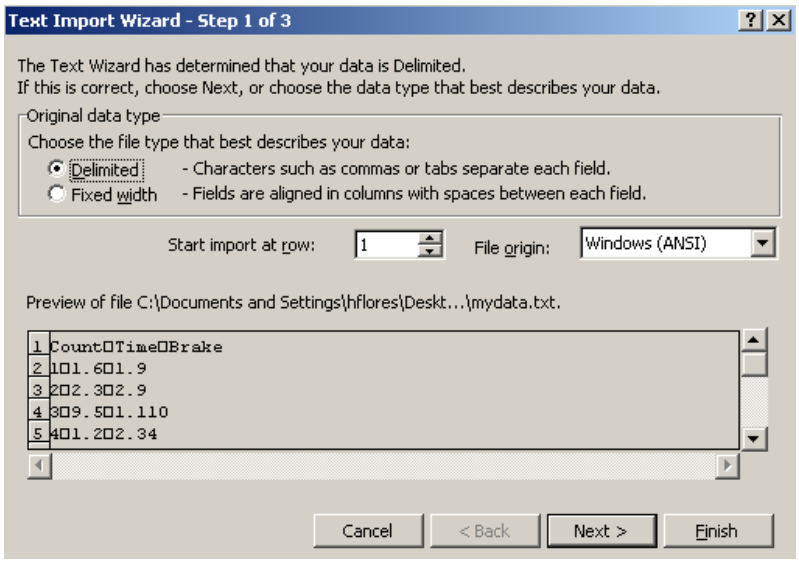
Now, assuming your data is "delimited" in some way, Excel loves you. You can import the data using the following steps (**Try this with excel_data.txt to be found in piazza**):

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

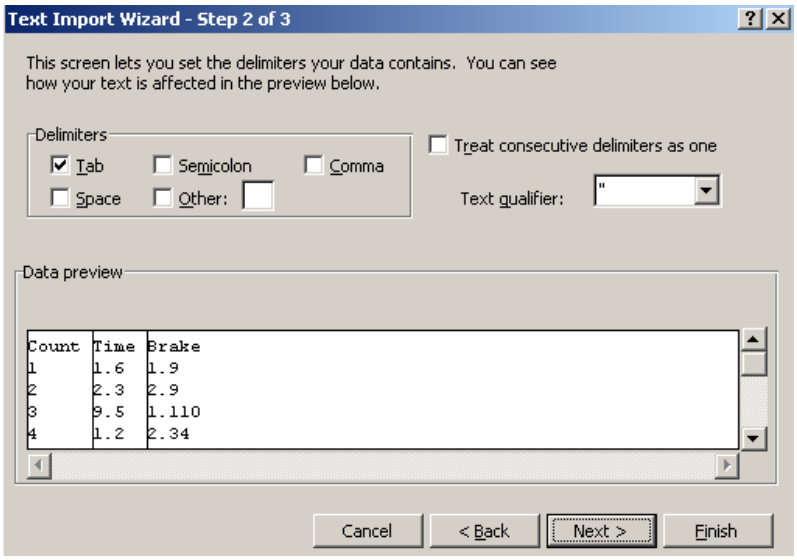


Import Step 2 – Find your file, and click *Open*. Note: You may have to change *Files of type* to *All Files* to see your file.

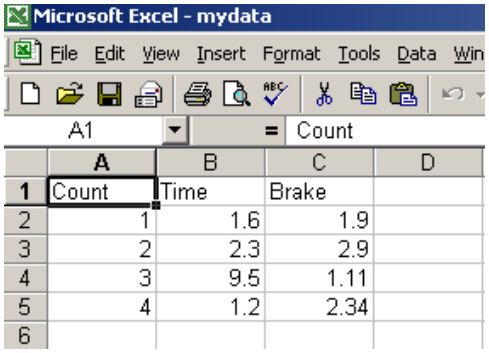
Import Step 3 – The import wizard will appear on the screen.



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.

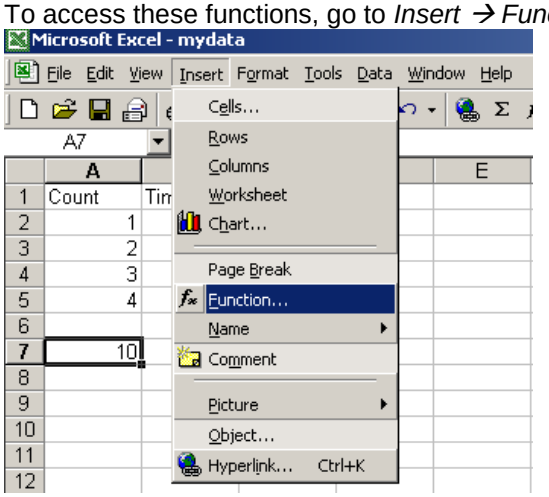
Trick #2: Suppose you want to add all the numbers in a particular column, but don't want to burden yourself with typing all the cell identifiers. Click on the cell you want to put the formula and type "= SUM(". Then, move your mouse to the first element, click-and-drag to the last element you wish to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

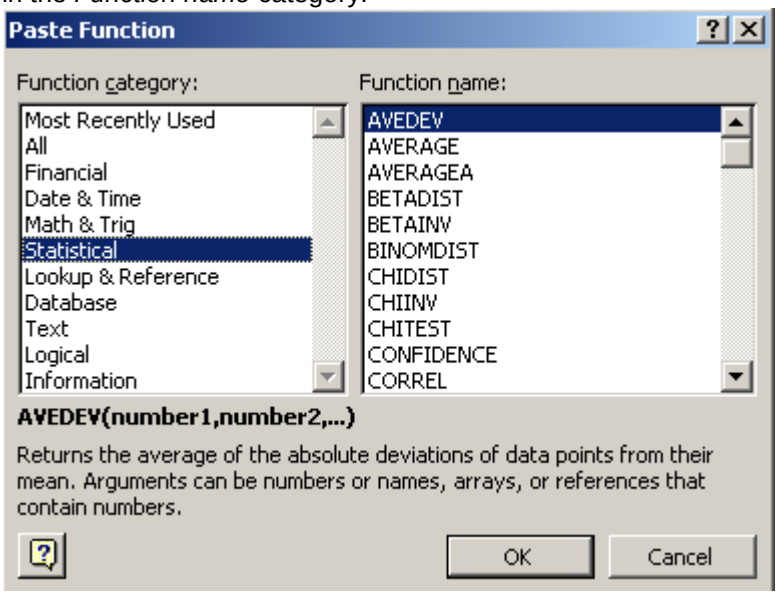
So where are the statistics?

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.



Lab 1:

a) **Apply the following functions to analyze the given data for a) the population of women b) the population of men, and c) combined the data for the two sexes**
b) **Interpret the statistics obtained based on the discussion below.**

AVERAGE
MAX [gives the largest number of all these cells in the population selected]
MIN [the smallest]
MEDIAN [the median]
MODE [The MODE function returns the mode (most frequently occurring number) in a group of supplied arguments. For example, =MODE(1,2,4,4,5,5,5,6) returns 5. Numbers can be supplied as numbers, ranges, named ranges, or cell references that contain numeric values.]

STDEV [standard deviation $\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$ where a is the number of data, x is the array of data, and x bar the average of the data]

Simple summary statistics

While simple numbers constitute the elements of data, inorder for them to be useful we need to look at the relationships between Statistics them, and perhaps combine them in some way. And this is where statistics comes in.

Here we look at some of the most straightforward ways: simply look at information and insights which can be extracted from relationships between values measured on the same variable.

For example, we might have recorded the ages of the applicants for a place at a university, the luminosity of the stars in a cluster, the monthly expenditures of families in a town, the weights of cows in a herd at the time of sending them to market, and so on. In each case, a single numerical value is recorded for each 'object' in a population of objects.

The individual values in the collection, when taken together, are said to form a '**distribution**' of values.

Summary statistics are ways of characterizing that distribution: of saying whether the values are very similar, whether there are some exceptionally large or small values, what a 'typical' value is like, and so on.

Averages

One of the most basic kinds of descriptions, or summary statistics, of a set of numbers is an '**average'**. An average is a representative value; it is close, in some sense, to the numbers in the set. The need for such a thing is most apparent when the set of numbers is large. For example, suppose we had a table recording the ages of each of the people in a large city – perhaps with a million inhabitants. For administrative and business purposes it would obviously be useful to know the average age of the inhabitants. Very different services would be needed and sales opportunities would arise if the average age was 16 instead of 60. We could try to get a ball-park feel for the general size of the numbers in the table, the ages, by looking at each of the values. But this would clearly be a tough exercise. Indeed, if it took only one second to look at each number, it would take over 270 hours to look through a table of a million numbers, and that's ignoring the actual business of trying to remember and compare them. But we can use our computer to help us.

First, we need to be clear about exactly what we mean by 'average', because the word has several meanings. Perhaps the most widely used type of average is the arithmetic mean, or just mean for short. If people use the word 'average' without saying how they interpret it, then they probably intend the arithmetic mean. Before I show how to calculate the arithmetic mean, imagine another table of a million numbers. Only, in this second table, suppose that all the numbers are identical to each other. That is, suppose that they all have the same value. Now add up all the numbers in the first table, to find their total (this takes but a split second using a computer). And add up all the numbers in the second table, to find their total. If the two totals are the same, then the number which is repeated a million times in the second table is capturing some sort of essence of the numbers in the first table.
This single number, for which a million copies add up to the same total as the first table, is called the arithmetic mean (of the numbers in the first table). In fact, the arithmetic mean is most easily calculated simply by dividing the total of the million numbers in the first table by amillion. In general, the arithmetic mean of a set of numbers is found by adding all the numbers up and dividing by how many there are. Here is a further example. In a test, the percentage scores for five students in a class were 78, 63, 53, 91, and 55. The total is 78 + 63 + 53 + 91 + 55 = 340. The arithmetic mean is then simply given by dividing 340 by 5. It is 68. We would get the same total of 340 if all five students each scored the mean value, 68. The arithmetic mean has many attractive properties. It always takes a value between the largest and smallest values in the set of numbers. Moreover, it balances the numbers in the set, in the sense that the sum of the differences between the arithmetic mean and those values larger than it is exactly equal to the sum of the differences between the arithmetic mean and those values smaller than it. In that sense, it is a 'central' value.

Those of a mechanical turn of mind might like to picture a set of 1kg weights placed at various positions along a (weightless) plank of wood. The distances of the weights from one end of the plank represent the values in the set of numbers. The mean is the distance from the end such that a pivot placed there would perfectly balance the plank.

The arithmetic mean is a statistic. It summarizes the entire set of values in our collection to a single value. It follows from this that it also throws away information: we should not expect to represent a million (or five, or however many) different numbers by a single number without sacrificing something. We shall explore this sacrifice later. But since it is a central value in the sense illustrated above it can be a useful summary. We can compare the average class size in different schools, the average test score of different students, the average time it takes different people to get to work, the average daily temperature in different years, and so on.

The arithmetic mean is one important statistic, a summary of a set of numbers.

Median

Another important summary is the median. The mean was the pivotal value, a sort of central point balancing the sum of differences between it and the numbers in the set. The median balances the set in another way: it is the value such that half the numbers in the data set are larger and half are smaller. Returning to the class of five students illustrated above, their scores, in order from smallest to largest, are 53, 55, 63, 78, and 91. The middle score here is 63, so this is the median.

Simple descriptions

In any case, once again the median is a representative value in some sense, although in a different sense from the mean. Because of this difference, we should expect it to take a value different from the mean. Obviously the median is easier to calculate than the mean. We do not even have to add up any values to reach it, let alone divide by the number of values in the set. All we have to do is order the numbers, and locate the one in the middle.

Question: Presented with these two summary statistics, both providing representative values, how should we choose which to use in any particular situation?

Since they are defined in different ways, combining the numerical values differently, they are likely to produce different values, so any conclusions based on them may well be different. A short answer is that the choice will depend on the precise details of the question one wishes to answer.

Here is an illustration. Suppose that a small company has five staff, each in a different grade and earning, respectively, $10,000, $10,001, $10,002, $10,003, and $99,999. The mean of these is $28,001, while the median is $10,002. Now suppose that the company intends to recruit five new employees, one to each grade. The employer might argue that in this case, 'on average', she would have to pay the newcomers a salary of $28,001, so that this is the average salary she states in the advertisement. The employees, however, might feel that this is dishonest, since as many new employees will be paid less than $10,002 as will be paid more than $10,002. They might feel it is more honest to put this figure in the advertisement. Sometimes it requires careful thought to decide which measure is appropriate. (And in case you think this argument is contrived, Figure 1 shows the distribution of American baseball players' salaries prior to the 1994 strike. The arithmetic mean was $1.2 million, but the median was only $0.5 million.)

This example also illustrates the relative impact of extreme values on the mean and the median. In the pay example above, the mean is nearly three times the median. But suppose the largest value had been $10,004 instead of $99,999. Then the median would remain as $10,002 (half the values above and half below), but the mean would shrink to $10,002. The size of just a single value can have a dramatic effect on the mean, but leave the median untouched.

This sensitivity of the mean to extreme values is one reason why the median may sometimes be chosen in preference to the mean. The mean and the median are not the only two representative value summaries.

Another important one is the **mode**.

This is the value taken most frequently in a sample. For example, suppose that I count the number of children per family for families in a certain population. I might find that some families have one child, some two, some three, and so on, and, in particular, I might find that more families have two children than any other value. In this case, the mode of the number of children per family would be two.

Dispersion

Averages, such as the mean and the median, provide single numerical summaries of collections of numerical values. They are useful because they can give an indication of the general size of the values in the data. But, as we have seen in the example above, single summary values can be misleading. In particular, single values might deviate substantially from individual values in a set of numbers. To illustrate, suppose that we have a set of a million and one numbers, taking the values 0, 1, 2, 3, 4, ..., 1,000,000. Both the mean and the median of this set of values are 500,000.

But it is readily apparent that this is not a very '**representative'** value of the set. At the extremes, one value in the set is half a million larger and one value is half a million smaller than the mean (and median).

What is missing when we rely solely on an average to summarize a set of data is some indication of how widely dispersed the data are around that average. Are some data points much larger than the average? Are some much smaller? Or are they all tightly bunched about the average? In general, how different are the values in the data set from each other? Statistical measures of dispersion provide precisely this information, and as with averages there is more than one such measure.

The simplest measure of dispersion is the range. This is defined as the difference between the largest and smallest values in the data set. In our data set of a million and one numbers, the range is 1,000,000 − 0 = 1,000,000. In our example of five salaries, the range is \$99,999 − \$10,000 = \$89,999. Both of these examples, with large ranges, show that there are substantial departures from Statistics the mean.

For example, if the employees had been earning the respective salaries of \$27,999, \$28,000, \$28,001, \$28,002, \$28,003 then the mean would also be \$28,001, but the range would be only \$4. This paints a very different picture, telling us that the employees with these new salaries earn much the same as each other. The large range of the earlier example, \$89,999, immediately tells us that there are gross differences. The range is all very well, and has many attractive properties as a measure of dispersion, not least its simplicity and ready interpretability. However, we might feel that it is not ideal. After all, it ignores most of the data, being based on only the largest and smallest values. To illustrate, consider two data sets, each consisting of a thousand values. One data set has one value of 0, 998 values of 500, and one value of 1000. The other data set has 500 values of 0 and 500 values of 1000. Both of these data sets have a range of 1000 (and, incidentally, both also have a mean of 500), but they are clearly very different in character.

By focusing solely on the largest and smallest values, the range has failed to detect the fact that the first data set is mostly densely concentrated around the mean.
This shortcoming can be overcome by using a measure of dispersion which takes all of the values into account

One common way to do this is to take the differences between the (arithmetic) mean and each number in the data set, square these differences, and then find the mean of these squared differences. (Squaring the differences makes the values all positive, otherwise positive and negative differences would cancel out when we calculated the mean.) If the resulting mean of the squared differences is small, it tells us that, on average, the numbers are not too different from their mean. That is, they are not widely dispersed. This mean squared difference measure is called the **variance of the data** – or, in some disciplines, simply the mean squared deviation. Illustrating with our five students, their test scores were 78, 63, 53, 91, and 55 and their mean was 68. The squared differences between the first score and the mean is $(78 - 68)^2$ = 100, and so on. The sum of the squared differences is 100 + 25 + 225 + 529 + 169 = 1048, so that the mean of the squared differences is 1048 ÷ 5=209.6. **This is the variance.**

One slight complication arises from the fact that the variance involves squared values. This implies that the variance itself is measured in 'square units'. If we measure the productivity of farms in terms of tons of corn, the variance of the values is measured in 'tons squared'. It is not obvious what to make of this. Because of this difficulty, it is common to take the square root of the variance. This changes the units back to the original units, and produces the measure of dispersion called the standard deviation. In the example above, the standard deviation of the students' test scores is the square root of 209.6, or 14.5.

The standard deviation overcomes the problem that we identified with the range: it uses all of the data. If most of the data points are clustered very closely together, with just a few outlying points, this will be recognized by the standard deviation being small. In contrast, if the data points take very different values, even if they have the same largest and smallest value, the standard deviation will be much larger.

Skewness (λοξότητα)

Measures of dispersion tell us how much the individual values deviate from each other. But they do not tell us in what way they deviate. In particular, they do not tell us if the larger deviations tend to be for the larger values or the smaller values in the data set. Recall our example of the five company employees, in which four earned about $10,000 per year, and one earned around ten times that. A measure of dispersion (the standard deviation, for example) would tell us that the values were quite widely spread out, but would not tell us that one of the values was much larger than the others. Indeed, the standard deviation for the five values $90,000, $89,999, $89,998, $89,997, and $1 is exactly the same as for the original five values. What is different is that the anomalous value (the $1 value) is now very small instead of very large. To detect this difference, we need another statistic to summarize the data, one which picks up on and measures the asymmetry in the distribution of values. One kind of asymmetry in distributions of values is called **skewness**. Our original employee salary example, with one anomalously large value of $99,999, is right skewed because the distribution of values has a long 'tail' stretching out to the single very large value of $99,999. This distribution has many smaller values and very few larger values. In contrast, the distribution of values given above, in which $1 is the anomaly, is **left skewed**, because the bulk of the values bunch together and there is a long tail stretching down to the single very small value.

Right skewed distributions are very common. A classic example is the distribution of wealth, in which there are many individuals with small sums and just a few individuals with many billions of dollars.

Quantiles

Averages, measures of dispersion, and measures of skewness provide overall summary statistics, condensing the values in a distribution down to a few convenient numbers. We might, however, be interested in just parts of a distribution. For example, we might be concerned with just the largest or smallest few – say, the largest 5% – values in the data set. We have already met the median, the value which is in the middle of the data in the sense that 50% of the values are larger and 50% are smaller. This idea can be generalized. For example, the upper quartile of a set of numbers is that value such that 25% (i.e. a quarter) of the data values are larger, and the lower quartile is that value such that 25% of the data values are smaller. This is taken further to produce deciles (dividing the data set into tenths, from the lowest tenth through to the highest tenth) and percentiles (dividing the data into 100ths). Thus, someone might be described as scoring above the 95th percentile, meaning that they are in the top 5% of the set of scores. The general term, including quartiles, deciles, percentiles, etc., as special cases, is quantile.

Reference: **Statistics: A Very Short Introduction**, David J. Hand, Oxford University Press, 2008.