STATISTICS WITH SPREADSHEETS

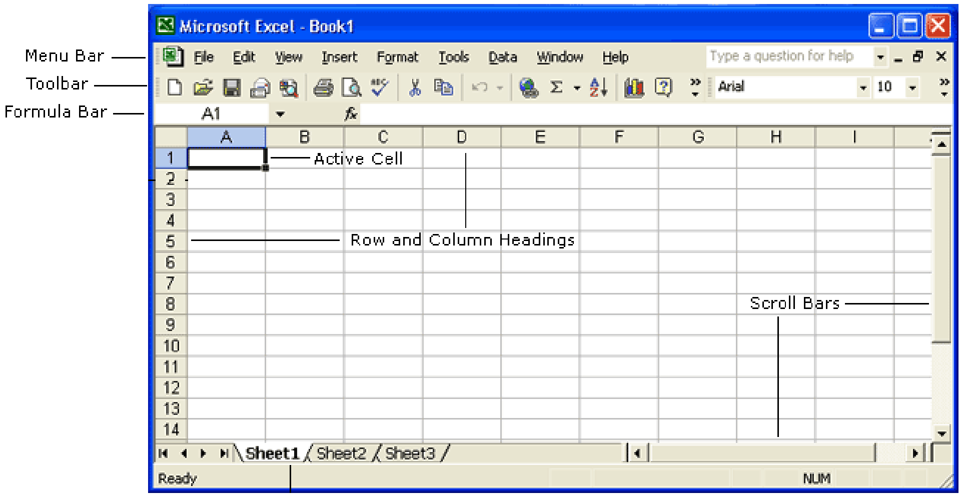
AMS 102: ELEMENTS OF STATISTICS

What is a spreadsheet? A spreadsheet is a computer application that presents data in a grid of multiple cells. Its primary purposes are to display data and aid in its analysis.

The most popular spreadsheet application is Miscrosoft Excel. Your computer most likely has it (if you have Microsoft Word on your computer, the likelihood is even greater). If you don't have access to Excel, other spreadsheet applications are available. For instance, one is provided as a part of the OpenOffice suite (available at `www.openoffice.org`).

The specific functions discussed below will be given in the Excel (for Windows) format. However, other spreadsheets are not very different; you can easily adapt this note for use with other applications.

Inputting data. As stated above, a spreadsheet presents data in a grid of cells:



Grid columns are labeled by letters and rows, by numbers. A cell is labeled by a letter (of its column) and a number (of its row).

Examples. 1. The cell at the intersection of column **B** and row **11** is labeled **B11**.
2. The cell **D8** is located at the intersection of column **D** and row **8**.

A cell can contain any kind of data: a number, a word, a date, or even a formula. To input anything into a cell, simply click on it and then enter whatever data you want to appear there. Then press `Enter`.

Once you click on a cell, its border becomes bolder (this is how you know what cell you're currenly modifying). But if you look closely, you will see that the cursor does not appear in the cell! In fact, you are not inputting the data directly into the cell. The input is happening above the grid in the so-called formula bar. (Note to Mac users: the formula bar may be floating somewhere or appear at the top of the screen, unless it's obscured by other windows.)

To edit the contents of a cell, click on it. The contents will appear in the cell and in the formula bar and you can modify as you would usually modify a string of text. Again, once the cell contents is to your liking, press `Enter`.

Of course, you don't have to enter the data from scratch every time you work with a spreadsheet. Data can be saved into a file (just as a text that you typed in an application like Microsoft Word) and date files can be opened providing you with a filled-in spreadsheet in a flash.

Functions: Basics. A spreadsheet is not just a nice way to present data; it can be used for computations too.

Pick a cell and enter 3+5 (press `Enter` to complete the input). The cell will display $3 + 5$. Now enter =3+5 (again, press `Enter`). The cell now shows 8. However, if you click on the cell again and look at the formula bar, it will show the formula $= 3 + 5$.

This is the general rule: to input a formula in the cell, start it with a "=" sign. After input, press `Enter`. The cell will display the result of the computation; the formula bar will display the formula.

A spreadsheet also allows you to perform operations with the numbers already in the cells. Suppose you have several numbers entered in a column, say in cells **A2, A3, A4, A5** and you want to sum them up. That is, you want to compute **A2+A3+A4+A5**. This you can do simply by entering in another cell =A2+A3+A4+A5. Note that there was no need to type out the numbers again. In a formula A2 stands for the number in the cell **A2**.

Now suppose you need to sum up a hundred numbers in column **A** that sit in rows from **1** to **100**. Typing out a hundred cell designators =A1+A2+... is impractical. We need to use a *function.*

Specifically, to sum up the numbers in the cells from **A1** to **A100**, pick a cell (where you want to display the sum) and type in the expression =SUM(A1:A100) (don't forget to press `Enter`). Here is the meaning of this expression:

= tells the spreadsheet that a formula is coming

SUM is the sum function; it tells the spreadsheet to sum up whatever follows the SUM in parentheses

A1:A100 is the *array* of data; it stands for all cells from **A1** to **A100**

Examples. 1. To sum up the numbers in column **D** in the rows **5-17**, input (into some other cell) =SUM(D5:D17).

2. To sum up the numbers in row **3** in the columns **B-G**, input (into some other cell) =SUM(B3:G3).

3. To sum up the numbers in column **D** in the rows **5-17** and the numbers in row **3** in the columns **B-G**, input (into some other cell) =SUM(B3:G3)+SUM(D5:D17).

4. Alternatively to sum up the numbers in arrays **B3:G3** and **D5:D17**, you can enter =SUM(B3:G3, D5:D17).

5. To double up the sum of numbers in column **D** in the rows **5-17**, input (into some other cell) =SUM(D5:D17)*2.

6. The function =SUM(D4:K4)*3 returns the sum of entries in row **4** in the columns from **D** to **K** times three.

There is another way to input a function. If you click on the f_x symbol next to the formula bar, a list of functions will appear. Choose the appropriate function there (for instance, SUM) and enter the required data. For a simple function such as SUM, this procedure is tedious but if a function has many parameters (whose order you may not remember), inputting it entirely by hand may lead to an error.

Statistical functions: Measuring the center. There are two ways to measure the center of the distribution: the mean and the median.

The Excel function that computes the mean is called AVERAGE. For instance, to compute the mean of the data in cells from **B2** to **B11**, enter =AVERAGE(B2:B11). (Just as in the SUM function, the expression B2:B11 describes data in all cells from **B2** to **B11**).

To compute the median, use the function MEDIAN. For instance, =MEDIAN(B2:B11) computes the median value of the data in the cells between **B2** and **B11**.

With both AVERAGE and MEDIAN, you can enter a list of several arrays or single cells separated by a comma: =AVERAGE(B2:B11, C3:G3, F8). The function will then compute the mean (or the median) of all the data in the list.

Statistical functions: Measuring the spread. To compute the standard deviation for the data in the array B2:B11 use =STDEV(B2:B11).

To compute the appropriate quartiles in the 5-number summary for the array B2:B11, use the function =QUARTILE(B2:B11, □), where □ should be replaced with the quartile you want to compute. For instance, =QUARTILE(B2:B11, 1) returns the 1st quartile.

More generally, □ can be replaced with 0,1,2,3,4,5. Here 1 stands for 1st quartile, 3 for the 3rd, 2 for the median (that we can also compute through the MEDIAN function), 0 for the minimum, and 5 for the maximum.

To compute the minimum, we also have the function MIN (as in =MIN(B2:B11)). For the maximum, we have the special function MAX.

Correlation and regression. If you want to compute the correlation between two sets of data, use the function CORREL.

To compute the correlation for several data points of the form (x_i, y_i), arrange the data in two columns or rows: the first column corresponds to the x's and the second, to the y's. As an example, if your data is contained in arrays A2:A11 and B2:B11, enter =CORREL(A2:A11, B2:B11) to compute the correlation. It is important that you work with two arrays of the same size, otherwise the function will return an error.

Note that unlike in the SUM function, two arrays inside the CORREL function play different roles.

To compute the coefficients a and b in the linear regression formula $y = a + bx$, use the functions =SLOPE(A2:A11, B2:B11) for b (the slope of the regression line) and =INTERCEPT(A2:A11, B2:B11) for a (the y-intercept of the regression line).

A spreadsheet application can also graph the regression line together with the original data. However, procedures for making a chart in a spreadsheet differ from one application to another (and even from one version of Excel to another). If you wish, you can learn them on your own from your spreadsheet's help file.

Normal distribution. The function =NORMSDIST(z) is an alternative to the table of standard normal distribution. It returns the proportion $P(Z < z)$ under the standard normal distribution, that is the proportion of the data in the distribution taking values less than the given number z. Here z can be entered as either a number or a cell value.

Example. Entering =NORMSDIST(1) returns 0.8413 (if you get the answer 0.84, your spreadsheet is set up to return only two decimal places; change the setting under the format tab).

If you are dealing with a non-standard normal distribution, you can standartize the values (that is, compute the z-scores) by applying the function =STANDARTIZE(x, μ, σ). Here x is the number whose z-score you're looking for; μ is the mean of the given normal distribution; and σ is its standard deviation.

Also, for a non-standard normal distribution, you can obtain the proportion $P(X < x)$ directly without standartizing first. Use the function =NORMDIST(x, μ, σ,TRUE). Here x is the given value, μ the mean of the distribution, σ its standard deviation, and TRUE is the word "TRUE" (if you put "FALSE" there, the function computes something else).

Example. The data is distributed normally as $N(10, 3)$. What proportion of data is less than 11?

Solution: enter =NORMDIST($11, 10, 3$,TRUE) in the spreasheet.

Even more statistics. Spreadsheet applications can perform even more intricate statistical calculations, such as determining confidence intervals and p-values. However, not all of them do it correctly. For instance, Excel is notorious for producing wrong confidence intervals.

For involved statistical tasks, the best approach is to perform them step-by-step (perhaps using spreadsheet applications and calculators at some steps) or use specifically designated statistical software.

Summary of statistical functions. Below is the summary of basic spreadsheet functions. Here "list" stands for a combination of array(s) and/or single cell(s):

AVERAGE(list)	Mean value of the data in the list
MEDIAN(list)	Median value of the data in the list
STDEV(list)	Standard deviation of the data in the list
QUARTILE(list, N)	N-th quartile of the data in the list
MIN(list)	The minimum value on the list
MAX(list)	The maximum value on the list
STANDARTIZE(x, μ, σ)	The z-score of x under the distribution $N(\mu, \sigma)$
NORMSDIST(z)	The proportion of values less than z under the standard normal distribution
NORMDIST(x, μ, σ,TRUE)	The proportion of values less than x under the distribution $N(\mu, \sigma)$