

# Introduction to Excel

Excel is a spreadsheet which allows you to enter data into cells, and do calculations on that data. Excel also comes with some graphing and statistics abilities, which are what concern us most in this course.

## ***Referring to Cells and Ranges***

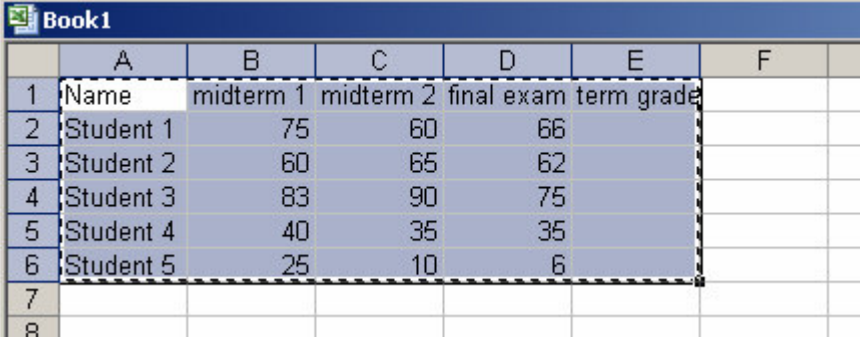
When you start Excel you will see a rectangular grid of cells, with letters across the top, and numbers down the left-hand side. Cells in Excel are referred to by the letter followed by the number, for example, B4. Because Excel formulas often take cell references as arguments, and because you often want to use the same formula in several locations, but with different cell references, when you cut and paste formulas, the cell references change. If you want the column not to change, prefix the letter by a \$, and if you want the row not to change, prefix the number by a \$. **\$B3** will fix the column, but the row may change, **\$B\$5** will fix both column and row (and hence the cell).

To refer to a rectangular range of cells, type in the cell at left upper corner, a “:”, and the cell at the lower right corner, for example, **A3 : C5** is the rectangular range of cells with upper left corner at A3, and lower right at C5. Of course, **A2 : A7** will give the second through seventh row of the “A” column, and **B3 : F3** will give the B through F column in the third row.

## ***Cell Contents***

You can type in words, numbers, or formulas into any cell, and what you type appears in the toolbar. If you need to edit what you have typed, click what you have typed in the toolbar, and edit away. By default, if you type in a number, Excel assumes it is data rather than text, and if what you type starts with an =, Excel assumes it is a formula.

Excel allows you to do calculations on a cell or a range of cells. For example, the following spreadsheet shows the grades of five students in a course where the two midterms are worth 25% each, and the final exam is worth 50%.



	A	B	C	D	E	F
1	Name	midterm 1	midterm 2	final exam	term grade	
2	Student 1	75	60	66		
3	Student 2	60	65	62		
4	Student 3	83	90	75		
5	Student 4	40	35	35		
6	Student 5	25	10	6		
7						
8						

To find the course grade of Student 1, we want  $\frac{1}{4}$  of midterm 1 and midterm 2, and  $\frac{1}{2}$  of the final exam. In the E2 cell, we type **=b2/4+c2/4+d2/2**, or, if we like decimals,

$=0.25*b2+0.25*c2+0.5*d2$ . The “=” tells Excel that what follows is a formula. The arithmetic operations are what you expect: “\*” means multiply, “+” means add, “-” means subtract, and “/” means divide. We could type in the relevant formula for the remaining students, but we can also “copy down.” Click on the E2 cell, and drag the mouse down to the E6, while holding the mouse button down. Release the mouse button, and either press “Ctrl-d”, or, in the “Edit” menu, select “Fill”, and the “Down.” Excel changes the row numbers in the cell references, so that the cell E3 contains the value obtained from  $=b3/4+c3/4+d3/2$ . Click on E4, and E5 to see how the formula has changes.

Excel also has built-in functions. If you want the average of the five students on the first midterm, type in **=average (b2 : b6)** . If you have typed this in to the B8 cell, you can copy across rows to get the averages on the other midterms. Select the B8 to D8 cells by clicking on the b8 cell, dragging across to the d8 cell while holding the mouse button down, and then releasing the mouse button. Type “Ctrl-r” (press both at the same time), or select “Fill right” under “Fill” in the “Edit” menu. If you include the word “average” in the A8 cell, your spread sheet should look like:

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F
1	Name	midterm 1	midterm 2	final exam	term grade	
2	Student 1	75	60	66	66.75	
3	Student 2	60	65	62	62.25	
4	Student 3	83	90	75	80.75	
5	Student 4	40	35	35	36.25	
6	Student 5	25	10	6	11.75	
7						
8	average	56.6	52	48.8	51.55	
9						
10						
11						

Useful Excel functions are:

average( <i>range</i> )	Finds the average of a range of cells. If a cell has no entry, it is ignored
sum( <i>range</i> )	Finds the sum of a range of cells.
countif( <i>range,criterion</i> )	Counts the number of cells within the range that satisfy the criterion. The criterion can be of the form 10, “>5”, “<=3” (include the “!”), “Ford” for example, or B9. For the first one, Excel counts the number of cells that contain the number 10, for the second, the number of cells containing a number larger than 5, for the third, less than or equal to 3, for the fourth, containing the word “Ford,” and for the last, having the same contents as the cell B9.
stdev( <i>range</i> )	Finds the sample standard deviation of a range of cells
stdevp( <i>range</i> )	Finds the population standard deviation of a range of cells
mode( <i>range</i> )	Finds the mode of a range of cells

<code>median(range)</code>	Finds the median of a range of cells
<code>quartile(range,number)</code>	Finds the 1 <sup>st</sup> , 2 <sup>nd</sup> , or 3 <sup>rd</sup> quartile of a range of cells, depending on whether the number is 1, 2, or 3.
<code>correl(range)</code>	Finds the correlation of a range of cells. The range should consist of two columns or two rows.

Countif can be used to create bar charts. For example, you may want to summarize the following data:


	A	B
1	Car bought	
2	Ford	
3	GM	
4	Toyota	
5	Toyota	
6	GM	
7	Ford	
8	Ford	
9	Ford	
10	GM	
11	Toyota	
12	GM	
13	GM	
14		
15		

Being lazy, you want Excel to count for you. You can type in Ford into the C2, GM in the C3, and Toyota in the C4 cells, and in the D2 cell, type in **=countif(A\$2:A\$13,C2)**. You can then copy the formula down to the other two cells. Notice the \$! Without it, the D3 cell would contain **=countif(A3:A14,C3)**, and the D4 cell, **=countif(A4:A15,C4)**. With this table, you can make your bar chart.

## Charts

Excel is able to make charts such as bar charts, and pie charts. Once you have summarized your data just described, highlight your table of data,

	A	B	C	D	E
1	Car bought		Cars	Frequency	
2	Ford		Ford	4	
3	GM		GM	5	
4	Toyota		Toyota	3	
5	Toyota				
6	GM				
7	Ford				
8	Ford				
9	Ford				
10	GM				
11	Toyota				
12	GM				
13	GM				
14					

click on the “Chart” icon () in the Excel toolbar, and follow the instructions. Excel tries to guess whether your data (called “Series” for Excel) is in columns or rows, and usually guesses correctly. Fill in the screens as they arise, making whatever changes you want. By the way, it is good practice to give your charts titles, and to label the axes. You can eliminate the legend either using the “Legend” tab on the second window of the “Histogram” command, or by deleting it from the finished chart. In fact, most editing can be done on the final chart.

### **Data Analysis Toolpak**

Most of the stats functions we want are included in the Data Analysis package. To see if the package has been installed, click on “Tools.” If “Data Analysis” is in the menu, it has been installed. Otherwise, click on “Add-ins” in the “Tools” menu, select the “Analysis Toolpak,” and click “OK.” The most useful commands in the Data Analysis are “Descriptive Statistics,” “Histogram,” and “Regression.” These all assume that your data is a column or row of numbers.

How to use the tools in Data Analysis is largely self-evident. Most will ask you to select a range. You can type this in yourself, or, click on the button on the right-hand side of the entry field. You will then see your spread sheet, with an entry field over it. Select the range of cells you want, and click on the button to the right of the entry field.

For histograms and “Descriptive Statistics,” your data will be a single column of numbers, as shown below:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	120	180	360	240	120	180	120	240	170	150	120	180	180	150	120	180	200	150

To get the descriptive statistics, choose “Descriptive Statistics” from the “Data Analysis” menu under “Tools.” Select the range of cells (A1:R1) for the above example, identify the left upper corner for the output, select what type of output you want (Summary Statistics, for example), and see the results. Notice that the summary statistics do not include the quartiles. These you have to find separately, using the quartile command.

For histograms, you want to set your classes before drawing the histogram. In an empty column or row, type in the right-hand endpoints of your classes. Using the data above, for example, if the classes are 0-100, 101-120, 121-140, 141-160, 161-180, and 181-200, we can type in:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	120	180	360	240	120	180	120	240	170	150	120	180	180	150	120	180	200	150
2																		
3	100	120	140	160	180	200												
4																		

Excel knows this row or column of right-hand endpoints as a “Bin Range.” Choose “Histogram” from the “Data Analysis” menu under “Tools”, identify the data range (A1:R1), and the bin range (A3:F3), identify the upper left corner of the table Excel will produce, and click “OK.” You should get:

	A	B	C	D	E	F	G	H	I	J
1	120	180	360	240	120	180	120	240	170	150
2										
3	100	120	140	160	180	200				
4										
5	Bin	frequency								
6	100	0								
7	120	5								
8	140	0								
9	160	3								
10	180	6								
11	200	1								
12	More	3								
13										
14										

Now you can make a bar chart with the table. Unlike a bar chart, however, you want the bars next to each other. Once the chart is complete, right-click on a bar, select “Format Data Series”, select “Options,” and change the “Gap width” to 0. The resulting chart has two defects: the labeling of the bars is somewhat confusing, and, if you have classes of different widths, the bars still have the same width. The second defect cannot be changed, so when making a chart with Excel, *always* uses classes of equal width. For the first defect, change the numbers in the “Bin” column to “0-100,” “101-200,” etc. Unfortunately, Excel has a nasty tendency to replace things like “1-4”, by “01 Apr,” which can be prevented by selecting the range of cells that will be affected, right clicking on them, and, under the “Number” tab, selecting “Text.”

For linear regression, once you have done a scatter plot of the data, you can add the Trendline (that is, the line of best fit), display its equation, and also  $r^2$  by right-clicking on a data point, selecting “Add Trendline,” and checking the relevant boxes. If you want more detailed linear regression analysis, such as residuals, use the “Regression” command in the “Data Analysis” Toolpak.

## Confidence Intervals and Hypothesis Testing for Single Populations

Here is where Excel's shortcomings are most evident: instead of being able to find confidence intervals, for example, using a dialogue box, everything must be done "by hand." An understanding of how to find confidence intervals and do hypothesis testing without Excel is therefore necessary to understanding this introduction. Most of the commands for statistical distributions behave similarly to those for the chi-square distribution, but the normal and t-distributions are exceptions.

Distribution	Commands	Behaviour
Normal	normdist( <i>x,mean,standard deviation,true</i> ) norminv( <i>p,mean,standard deviation</i> )	Both deal with the cumulative distribution, that is they deal with $P(X \leq x) = p$ , where X has a normal distribution with the given mean and standard deviation. For normdist, you specify x, and obtain p, and for norminv, you specify p, and obtain x. For normdist, "true" tells Excel to use the cumulative distribution, "false" tells Excel to find the value of the probability density function at x.
T	tdist( <i>x,degrees of freedom,tails</i> ) tinv( <i>p,degrees of freedom</i> )	Both deal with the tails of the t distribution with the given degrees of freedom. The statistic you use is $\frac{X - \mu}{std\ dev}$ . For tdist, x must be positive. If tails is 1, it assumes you want the area of the right-hand tail, and if tail is 2, it assumes you want the area to the right of x, and to the left of -x. tinv, assumes the probability is split equally between the two tails, and finds only the positive x. If you want a single tailed distribution, double your probability, and remember that the x you get corresponds to the right-hand tail only.
Chi-square	chidist( <i>x,degrees of freedom</i> ) chiinv( <i>p,degrees of freedom</i> )	Both deal with $P(X > x) = p$ . The distribution is standardized, and your statistic is $\frac{(n-1)s^2}{\sigma^2}$ .

### Confidence Intervals

#### Chi-Square

We deal with the chi-square distribution first because the commands for the distributions other than normal or t are similar to it. If you want a confidence interval of the

population standard deviation, you are solving the equation  $P\left(\frac{(n-1)s^2}{\sigma^2} < x\right) = p$ , for  $\sigma$ ,

where  $p$  is given, and  $x$  comes from  $\text{chiinv}(p, n-1)$ . Solving for  $\sigma$  involves taking reciprocals, so the direction of the inequality changes. If you want the value of  $x$  for

which  $P(\sigma^2 > x) = \alpha$ , you really want  $P\left(\frac{(n-1)s^2}{\sigma^2} < \frac{(n-1)s^2}{x}\right) = \alpha$ . Because of the

change in the direction of the inequality, you must replace  $\alpha$  by  $1-\alpha$ .  $\text{chiinv}(1-\alpha, n-1)$  will find the value for  $\frac{(n-1)s^2}{x}$ , and you must solve for  $x$ :  $x = \frac{(n-1)s^2}{\text{chiinv}(1-\alpha, n-1)}$ . You will,

of course, use the Excel to do the calculations, so make sure to identify what each cell contains (include some text in an adjacent cell). If you want the value of  $x$  for which  $P(\sigma^2 < x) = \alpha$ , you will be able to do your calculations with the  $\alpha$  given. If you want a two-tailed distribution, remember that the chi-square distribution is not symmetric. You will need both  $\text{chiinv}(\alpha, n-1)$  and  $\text{chiinv}(1-\alpha, n-1)$ .

### T-Distribution

$\text{tinv}$  assumes an interval with both tails missing, splits  $p$  between the two missing tails, and returns the positive  $x$  for which  $P\left(-x < \frac{X - \mu}{\text{std dev}} < x\right) = 1 - \alpha$ . Because your statistic is

$\frac{X - \mu}{\text{std dev}}$ , solving for  $\mu$  will give the interval

$(X - (\text{std dev})\text{tinv}(\alpha, n-1), X + (\text{std dev})\text{tinv}(\alpha, n-1))$ . Notice that symmetry allows us to be sloppy: technically, the inequalities change directions because of the  $-\mu$ . As with the chi-square distribution, make sure to identify the meaning of this number in an adjacent cell. If you want an interval with the right-hand tail missing, double  $\alpha$ . Doing some algebra, and using symmetry will produce the interval  $(-\infty, X + (\text{std dev})\text{tinv}(2\alpha, n-1))$ . Again, after finding the right-hand endpoint of this interval, make sure to explain what the number represents! A similar technique will produce the interval with left-hand tail missing.

### Normal Distribution

Excel's  $\text{confidence}(\alpha, \text{std dev}, \text{size})$  will find half the length of the two-tailed confidence interval where  $\alpha$  is the area of the combined tails. The upper bound will therefore be  $X + \text{confidence}(\alpha, \text{std dev}, n)$ , and the lower bound will be  $X - \text{confidence}(\alpha, \text{std dev}, n)$ . Determining the intervals with one tail missing are left to the reader.

## One Population Hypothesis Testing and P-Values

For the critical value approach, the critical values are, of course, calculated as for confidence intervals. For p-values, you want to use  $\text{chidist}$ ,  $\text{tdist}$ , or  $\text{normdist}$ .

### Chi-Square

If you want the size of the right-hand tail, that is, if  $S^2$  is the random variable associated to the sample variance, and  $s^2$  is the sample variance you measured, you want

$$P(S^2 > s^2) = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)s^2}{\sigma^2}\right), \text{ which will be } \text{chidist}\left(\frac{(n-1)s^2}{\sigma^2}, n-1\right).$$

### **T-Distribution**

For `tdist`, you can specify whether  $\alpha$  is split evenly between the two tails, or is only the area of the right-hand tail. If you want the left-hand tail, use symmetry to get the p-value. If  $X$  is the random variable,  $x$  is the measured value, and you are interested in the right-

hand tail, you want  $P(X > x) = P\left(\frac{X - \mu}{std\ dev} > \frac{x - \mu}{std\ dev}\right)$ , so

$\text{tdist}\left(\frac{x - \mu}{std\ dev}, \text{degrees freedom}, 1\right)$  will give the value. For the two-tailed distribution,  $x - \mu$

must be positive, and the value returned is  $P\left(\left|\frac{X - \mu}{std\ dev}\right| > \frac{x - \mu}{std\ dev}\right)$ . If your  $x - \mu$  is negative,

you will need to use symmetry.

### **Normal Distribution**

`normdist` returns the area under the normal distribution excluding the right-hand tail. If  $X$  is the random variable,  $x$  the value measured, and you want  $P(X > x) = 1 - P(X < x)$ , then  $1 - \text{normdist}(x, \mu, \sigma, \text{true})$  gives the desired result. If you want  $P(X < x)$ , omit the "1 -," and use symmetry to find the p-value for the two-sided case.