

**INTRODUCTION TO EXCEL®**  
**FOR**  
**DATA PROCESSING AND STATISTICS**

Lin Hammill, Jim Gunson and Jan Verster



---

## Table of Contents

---

1. Introduction .....	1
1.1 Starting <i>Excel</i> .....	1
1.2 Saving a File .....	2
1.3 Opening a File .....	3
1.4 Exiting <i>Excel</i> .....	3
1.5 Selecting a Region .....	3
1.6 Printing Work .....	4
1.7 Copying to a Word Processor .....	4
1.8 References and Help .....	5
2. Entering and Naming Data .....	6
2.1 Entering Data .....	6
2.2 Entering Formulas .....	6
2.3 Editing Cells .....	7
2.4 Naming Cells .....	7
2.5 Inserting and Deleting Rows/Columns .....	7
2.6 Formatting Entries .....	8
3. Creating Histograms and Other Charts .....	9
3.1 Histograms .....	9
3.2 Pie Charts .....	13
3.2 Bar Charts .....	14
4. Descriptive Statistics .....	16
4.1 Using built-in functions .....	16
4.2 Using a Template .....	16
4.3 Using <i>Excel</i> 's Built-in Package .....	17
5. Confidence Intervals For Means .....	19
6. Hypothesis Testing of Mean of One Sample .....	21
7. Hypothesis Tests for Two Samples .....	23
7.1 Comparison of means, dependent samples .....	23
7.2 Comparison of Means, Independent Samples .....	25
7.3 Comparison of Variances .....	26
8. Confidence Intervals/Hypothesis Tests for Proportions .....	27
9. Scatter Plots, Correlation and Regression .....	29
9.1 Plotting a Scatter Plot .....	29
9.2 Finding the Correlation .....	30
9.3 Finding the Regression Line .....	32
9.4 Plotting the Regression Line .....	32
9.5 Plotting Residuals .....	33

10. Chi-Square Tests for Variances and Standard Deviations .....	34
11. Testing Data For Being Normally Distributed .....	35
12. Presentations .....	37
Appendix 1: Location of Files .....	37
Appendix 2: <i>Excel</i> Functions used in Statistics .....	37

---

# 1. Introduction

---

*Excel* is an electronic spreadsheet, which may be used to organize data, perform calculations, create charts and for many other purposes. Graphics, tables and charts created by *Excel* can be pasted into word processors such as *Word* or *WordPerfect*, and thus be included in essays and reports.

*Excel* is a valuable tool in statistics, science and business. Statisticians use it to create histograms and scatter plots, calculate basic statistics, construct confidence intervals and carry out hypothesis testing and regression analysis. In science and business, it is used to manipulate data and plot data, and to find best-fit approximations.

**Note: to perform the examples and exercises in the manual at the college, you will need a floppy disk on which to store your work**

## 1.1 Starting *Excel*

First run *Windows*. How you then open *Excel* depends on your system. Here are some possibilities:

- Look for an *Excel* shortcut (see below), possibly on a taskbar, at the top of the screen.



Figure 1

- If you find it, click on it.
- Otherwise click on *Start, Programs* and look for *Excel*. Click on it.

Note that the object (file) created by *Excel* is a **workbook**, which is organized into **worksheets**, as a book is made up of pages. You can view the different sheets by clicking on the tabs (*Sheet 1, Sheet 2*, etc.) above the status bar, at the bottom of the screen.

A sample worksheet is shown on the next page. The sheet contains several bars at the top. The first one is the *title bar*, which reads "Microsoft Excel". Below that is the *menu bar*, followed by the *standard toolbar*, and the *formula bar*. At the bottom is the *status bar*, which should read "Ready". Between these is the *workbook window* with a new *workbook*.

A worksheet consists of cells; each one referred to by a column letter and a row number. The currently selected cell (A1) is highlighted by a black outline and its reference appears in the *name box* at the left of the *formula bar*. The contents of this cell appear in the *formula box* to the right of the *name box*.

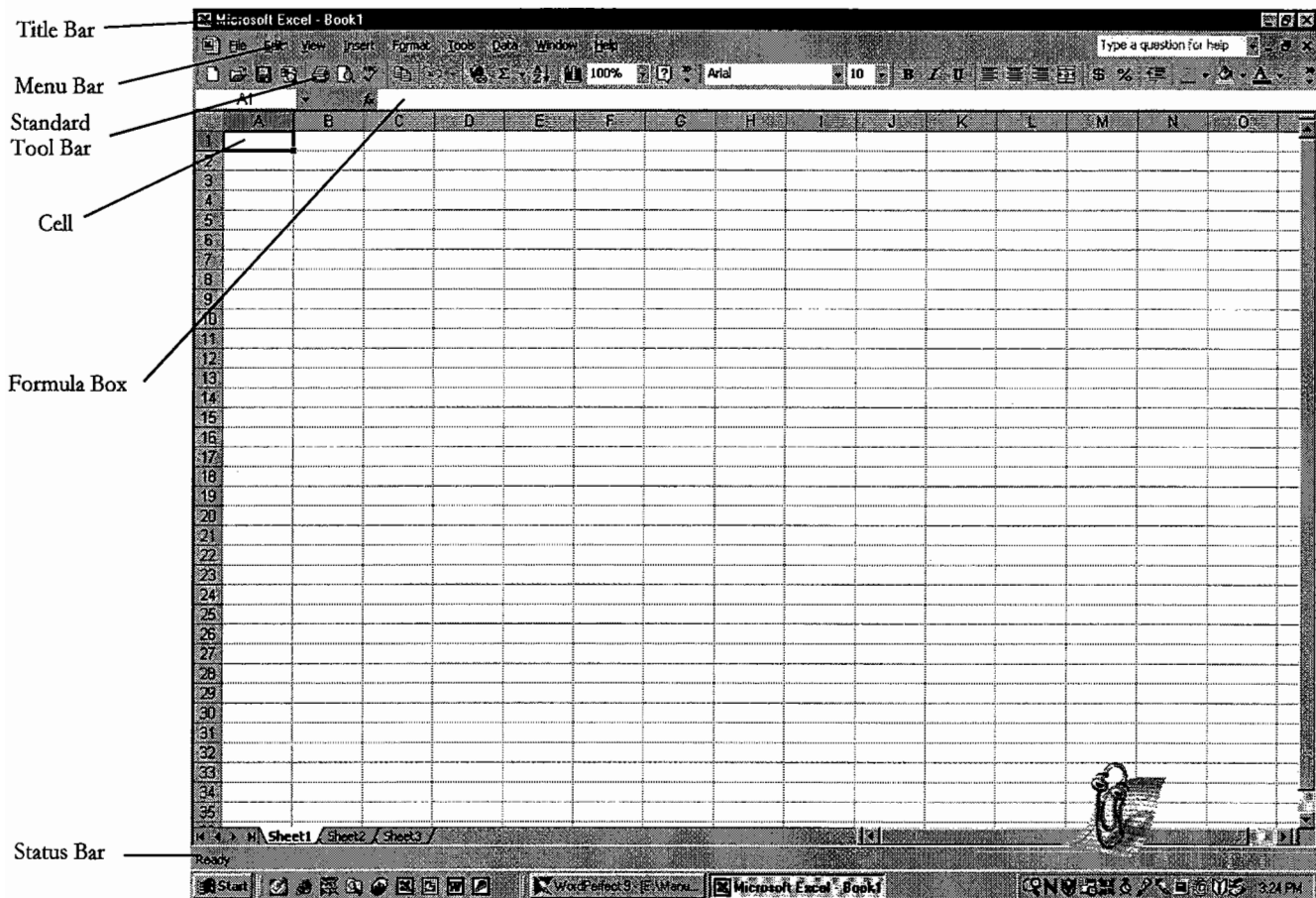


Figure 2: an Excel Worksheet

## 1.2 Saving a File

As you may work on a *workbook* over a period of time, you will want to be able to save it as a file on a floppy, and later re-open it for further work. Should you be working on your own computer, you may wish to ignore the references to floppies and the *A:* drive and save files on the *C:* hard drive.

Files may be saved using *Save* or *Save as*, which may be found by clicking on *File*, in the *Menu bar*. If a file already has a name, having been opened from a floppy in the *A* drive, *Save* will save it under that name on that drive, overwriting the original version. Thus once a file has been created and saved on a floppy, you may *Open* it from the floppy to work on it and then use *Save* to update the floppy. *Save as* may be used if you wish to save a file without destroying an earlier version.

### To *Save*:

- Pull down the *File* menu and select *Save*, or click on the *Save* icon on the *standard toolbar*. If the file has been loaded from a floppy, no further action is required. If the file is new, the *Save as* dialog box will open and the steps are as for *Save as* below.

To *Save as*:

- Pull down the *File* menu and select *Save as*. The *Save as* dialogue box will open.
- Be sure drive A is chosen and that you have a formatted disk in drive A.
- Delete what is in the *File name* box, and replace by a name of your choosing, such as *myproject*.
- In the *Save as type* box be sure that *Microsoft Excel Notebook* is chosen. *Excel* will automatically add the extension *.xls* to the file name so that it will be saved as *myproject.xls*, for example.
- Now click the *Save* button.

### 1.3 Opening a File

Opening a file loads a worksheet from your floppy and creates an *Excel workbook*.

- Click on *File*, then *Open*, or click on the *Open* icon on the *standard toolbar*. Specify the name of the file in the dialog box and then click on *Open*. You may need to change the directory in the *Look in* box to find your file. If the file is on a floppy, place it drive A. Select drive A at the top of the dialogue box and then click on the file you want.

### 1.4 Exiting *Excel*

To quit using *Excel*:

- Click on *Exit* under the *File* menu, or on the cross at the top right corner of the *Excel* window. If you have not saved the *workbook* since you last made changes, *Excel* will prompt you to save it.

### 1.5 Selecting a Region

In order to copy-and-paste, and perhaps print, you must be able to select from the worksheet.

To select a cell:

- Click on it.

To select a row or column:

- Click on the number or letter that labels it. These are located on the row above the worksheet or the column to the left of it.

To select a rectangular region:

- Place the cursor in one corner, press and hold down [*Shift*]. Click in the opposite corner and release *Shift*. The region should be highlighted.

Alternatively:

- Place the cursor in one corner of the region. Press the left mouse button and hold it down.
- Drag the cursor to the opposite corner and release.

To select a plot:

- Click on the plot.

## 1.6 Printing Work

You may never wish to print a worksheet, in whole or in part, as it is usual to copy selections into reports that are produced by a word processor. If, however, you do wish to do so, here are the steps:

- Open the *File* menu and select *Print Preview* to see how your file will appear when printed. (This will not work if your worksheet is blank.)
- If you are happy with its appearance, choose *Print* from the *File* menu. Check the details in the dialog box, edit these, if necessary, and then click *Print*. (If you are working in a lab, you may have to ask the technician for help in determining which computer to use and where the document will be printed.)
- If you are not happy with the appearance of your document you can open the *Page Set Up* dialog box from the *File* menu and make changes in the appearances of your document. You may wish to check the Excel reference for help with this.

## 1.7 Copying to a Word Processor

You can copy charts, graphs, tables or other parts of your workbook to a word processor such as *Word* or *WordPerfect*. To do so you must run the two applications, *Excel* and the word processor, in separate windows.

- Open one of the applications as usual and open the appropriate file.
- Open the second application and open the appropriate file.
- You should now see the names of the applications at the bottom of the screen. Clicking on one will allow you to work on it. Thus, at a given time, you may choose which application you wish to use.

To copy from *Excel*:

- In the *workbook*, click on the worksheet containing the chart, table or graph you wish to copy.
- Select the region or plot.
- Copy either by choosing *Copy* from the *Edit* menu or by clicking on the *Copy* icon on the *standard toolbar*.
- Switch to the word processor. Move the text cursor to the point where you want the chart, table or graph to be placed.
- Paste either by choosing *Paste* from the *Edit* menu, or the paste icon from the *standard toolbar*.

Pasted images may be edited for size and position using the usual word processor tools.

## 1.8 References and Help

Excel has several useful avenues for getting help:

### **Help menu:**

- Choose Contents and Index to search the Help files, or click on What's this? from the Help Menu.

### **Tool Tips:**

- When the mouse is placed over a button on the tool bar and left there for a short time, its name appears. This helps in using the Help Menu.

### **Screen Tips:**

- Click on the question mark sign that appears in the top-right corner of most dialog boxes.

You should look for Excel references in two places, the computer labs and on reserve in the library. The lab technician or tutor may be able to help with some problems. You may also ask your instructor for help.



---

## 2. Entering and Naming Data

---

The following example shows how to enter and label data.

### Example 1

A college advisor wishes to study the relationship between incoming students' high school averages (HSA's) and their first semester college grade point averages (GPA's). A sample of 10 students is obtained and their HSA's and GPA's are recorded, in the following table.

HSA	70	75	78	80	85	82	72	65	68	75
GPA	2.4	2.8	3.3	3.1	3.7	3.0	2.5	2.3	2.8	3.1

For convenience, this table has been printed horizontally. In *Excel*, however, it is usual to enter data vertically, so that the data corresponding to a variable appear as a column, rather than as a row.

### 2.1 Entering Data

To enter data you must select a cell and then type in the number. For the above data:

- Open a new workbook.
- Type HSA in cell A1. (*Excel* may well correct your spelling to HAS. You will need to turn off the automatic spell checker. Under the *Tools* menu, pick *AutoCorrect* and remove the checkmark beside *Replace text as you type*.) Note that as you type, the letters appear in both the cell and the formula box on the formula bar. Press [*Enter*] and the highlight moves to cell A2.
- Type 70. Press [*Enter*].
- Continue down that column, entering 75, 78, etc.
- When you get to the bottom use the arrow keys or the mouse to move the cursor to B1. First enter the name GPA and then moving down the column enter the data 2.4, 2.8, etc.

*Save this workbook for later use.*

### 2.2 Entering Formulas

You may specify the value of a cell by a formula that depends on the values in other cells. You can, for example, define a cell to be the mean (average) of a range of cells. To enter a formula:

- Type the equals sign (=) followed by the formula.

For example, to find the average of the first four HSA's and put the result in cell D2, select cell D2 and type  $= (A2+A3+A4+A5)/4$  and press [Enter]. The cell should now display 75.75.

Excel has many built in functions which can make it easier to enter complicated formulas. For example, if you type  $=AVERAGE(A2:A5)$  into cell D3, you will get the same result as in D2.

There is a glossary of commonly used statistics functions in appendix 2.

## 2.3 Editing Cells

To replace the contents of a cell:

- Move the cursor to the cell you want to replace and click. If you type anything now, the current contents will be erased and the new typing will replace it.

For example, if you want D2 to show the number of data in the second column, select D2 and type  $=COUNT(B2:B11)$  After you press [Enter] you should see the result 10.

To edit the contents of a cell:

- Select the cell and press the key *F2*. This will display the formula which you can now edit.

For example, if you want D3 to be the mean of all the HSA's, select the cell, press *F2*, and change the 5 in the formula to 11 and press [Enter]. The result should be 75.

## 2.4 Naming Cells

Naming cells or a group of cells makes it easier to refer to them in formulas and dialogue boxes. To make the label HSA at the top of the first column a name for the data in that column:

- Select the cells from A1 to A11
- From the *Insert* menu, click on *Name*, then *Define*. If the first box under *Names in Workbook* contains the name you want, click *OK*. If it is blank, the name already exists, and you can either use a different name, or edit the list of existing names. When done, press *OK*.

### ***Do this for both HSA and GPA***

You can now refer to each column by name. For example to find the standard deviation of the HSA data, select cell D4 and type  $=STDEV(HSA)$  and press [Enter]. You should see 6.377042.

## 2.5 Inserting and Deleting Rows/Columns

To insert a row between existing rows:

- Place the cursor in the row below where the new row should go. Choose *Insert* from the menu bar and then *Rows*.

Named cells or ranges are adjusted to allow for the new row.

To insert a column between existing columns:

- Place the cursor in the column to the right of where the new column should go. Choose *Insert* from the menu bar and then *Columns*.

To delete a row or column:

- Place the cursor in any cell of the row or column. Choose *Edit* from the menu bar, then *Delete* and in the dialogue box check *Entire Row* or *Entire Column* and click *Ok*.

## 2.6 Formatting Entries

If the entry in a cell does not completely display because it is too long to fit, you may wish to make the column wider or to use a smaller font.

To change column widths:

- Move the cursor to the margin at the top of the worksheet that contains the letters identifying the columns. Position it over a break between two columns. It will change its appearance to a double arrow with a vertical line. Hold down the left mouse button and drag the symbol to make the column wider or narrower as you wish. If you double click, it will make the column wide enough for all entries in that column.

To change the font or size:

- Select the cell or region you wish to format.
- Under the *Format* menu, choose *Cells* and click on the appropriate tabs at the top of the dialogue box that opens. Change the formatting as you wish.



---

## 3. Creating Histograms and Other Charts

---

### 3.1 Histograms

#### Example 2

Create a histogram for the high school grades from example 1.

*If necessary, open the file from your floppy disk*

#### 3.1.1 Deciding on Classes

As the data is integer, the classes 65-69, 70-74, 75-79, 80-84 and 85-89 are suitable, using a class width of 5. The first class has lower class boundary at 64.5 and upper class boundary at 69.5. We will create a list of class boundaries, in column D, as follows:

- Put the first two class boundaries, 64.5 and 69.5, in D2 and D3
- Rather than type the others (up to 89.5) we exploit *Excel's* ability to perform linear extrapolation to take a list of two numbers and continue it so that the numbers are equally spaced. This saves time for large data sets and avoids arithmetic errors.

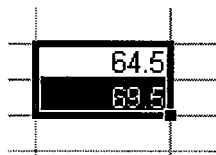


Figure 3

- Select these two cells. Put the cursor on the little square you see at the lower right corner of the highlighted cells.
- Drag downward with the mouse to D7 and let go. (Notice that as you drag a small yellow box appears beside the cursor, giving the latest number.) Now D2 to D7 should contain the class boundaries we want.

Since these boundaries are upper class marks, type UCM into D1. (You may also wish to use this to name the column.)

#### 3.1.2 Creating a Histogram from Raw Data

- Open the *Tools* menu. If you see *Data Analysis* at the bottom, click on it.
- If you do not see *Data Analysis*, click on *Add-ins* and put a check beside *Analysis Toolpak* and click *OK*.

- Now *Data Analysis* will appear on the *Tools* menu. Click on it and then click on *Histogram* and *OK*. A dialogue box will open.

(Note that dialogue boxes can be moved to a more convenient spot, by clicking on them and then dragging.)

- Fill in the boxes as follows:

Input range: A1:A11 (This is the HSA data.)

Bin Range: D1:D7 (*Excel* uses *Bin Range* to mean upper class boundaries.)

Output Range: A13 (This is a convenient location: *Excel* will place the output to the right and below A13.)

Note that these ranges can be entered by first clicking on a box in the dialogue box and then selecting the data using the mouse.

- Put a check beside *Labels* to indicate that the labels are included in the columns, otherwise *Excel* will treat the labels as data and report an error.
- Put a check beside *Chart Output*

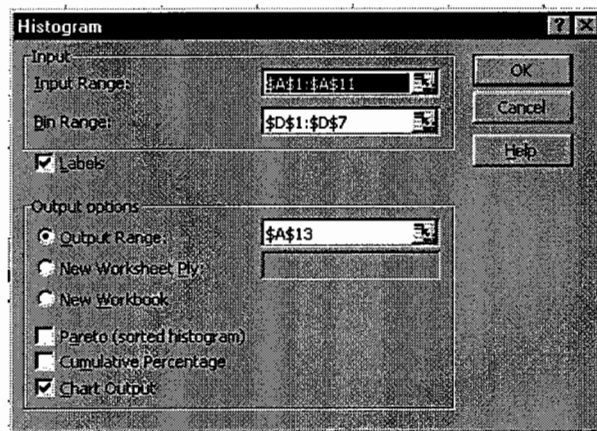


Figure 4

- Click *OK*. This gives figure 5:

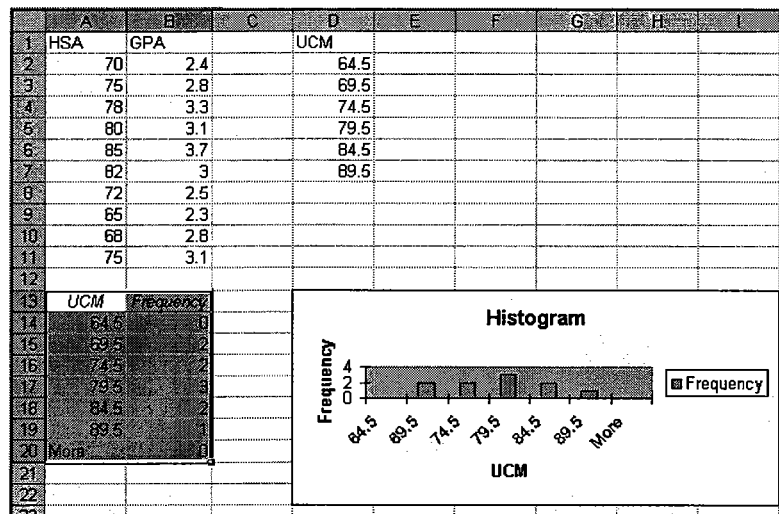


Figure 5

Note that the Output Table (high-lighted in figure 5) gives the frequency distribution. The first number in each row is the upper class limit and the second is the number of data in that class. Thus the first row gives the number of data less than or equal 64.5; the second row gives the data greater than 64.5 and less than or equal 69.5 and so on. Since there is no data less than or equal 64.5, this class mark could have been omitted from the list. However, it is a good idea to include it, as it acts as a safeguard in the event that you misread the table and mistake the lowest number. These classes can be removed, see below.

### 3.1.3 Modifying the Chart

It is usually necessary to edit a chart to improve its appearance. It is important to know that the Histogram is linked to the output table, but not to the original data. If you edit the data, nothing will happen unless you repeat the above procedure. If you edit the Output Table, it will immediately change the Histogram. We can exploit this by removing the first and last class, and by changing the remaining upper class marks to class marks, so that the chart is correctly labelled.

### 3.1.4 Labelling the Classes

To change the labels to the class marks, we need to replace the upper class marks in the Output Table by the class marks: 67, 72, ..., 87

- Type 67 in A15 and 72 in A16, and then select and drag as we did before when creating the class boundaries.
- Delete the entries in the top and bottom rows of the Output Table, which have no data. (Don't delete the rows.) This simplifies the plot.

Your Histogram should now look like figure 6.

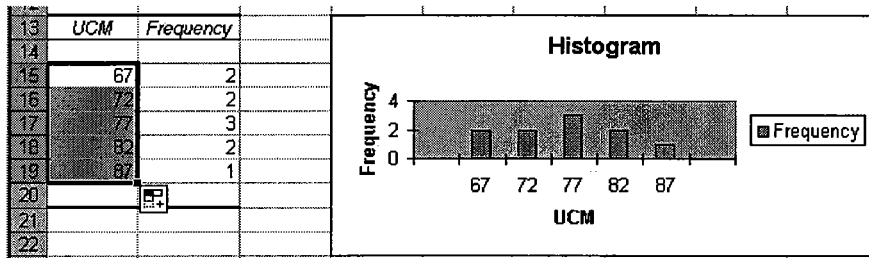


Figure 6

### 3.1.5 Editing the Chart

The chart's appearance should be further improved. To edit any part of the plot, point to it with the cursor. Its name will appear beside the cursor. Right clicking will then display a dialogue box that lets you edit that object. Make the following changes.

- The word *Frequency* in the box at right is unnecessary. To remove it, click on it and press the [Delete] key.
- To make the plot taller, move the cursor to the actual plot (so that you see the name *Plot Area* and click on it. Then drag on the sizing handles that now surround the plot. If you are going to print in black and white, you should remove the grey colour by right clicking on the *Plot Area*, choose *Format Plot Area*, and under *Area* choose *None*.
- In a Histogram, the bars should touch. To change the gap width, point the cursor at a bar, and right click. From the menu that appears, choose *Format Data Series*. Choose the *Options* tab and set the *Gap Width* to 0. (If you want to change the colour or the fill pattern of the bars, choose the *Patterns* tab.) When done, click *OK*.
- Move the cursor to the horizontal axis so that the name *Category Axis* appears. Right click and choose *Format Axis*. Select the *Scale* tab and make sure that the box *Value (Y) axis crosses between categories* is not checked. Click *OK*.
- Move the cursor to the grey part (so you see *Plot Area*). Right click and choose *Format Plot Area*. For area, choose none. Click *OK*.

By now, your histogram should look like the following:

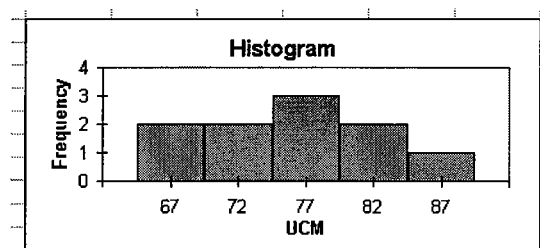


Figure 7

To finish it off, you will need to edit the titles.

- Click on the word *Histogram* so that a sizing box appears around it, and then click on it again. You should have a text cursor which allows you to backspace or edit as usual. Change the word *Histogram* to *High School Averages*.
- In the same way, change *UCM* to *Student Averages (%)* and *frequency* to *Number of Students*.

Your final histogram should look like the following. You can cut and paste it into your word processor.

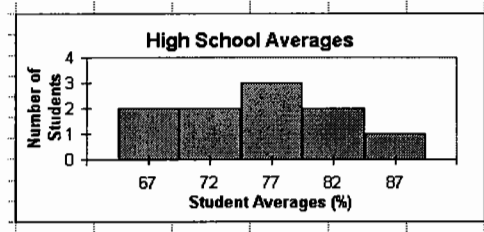


Figure 8

### 3.2 Pie Charts

#### Example 3.

A college has an annual budget of \$8 million. This is spent as follows.

Faculty salaries	Staff salaries	Administration	Facilities	Travel	Misc
\$4 000 000	\$2 000 000	\$700 000	\$500 000	\$300 000	\$500 000

To draw a pie chart

- Enter the data with the labels in column 1 and the amounts in column 2. Using units of \$1 million (so that \$400 000 is given by 0.4) leads to simpler labelling.

	A	B
1	Function	Amount
2	Faculty	4
3	Staff	2
4	Admin	0.7
5	Facilities	0.5
6	Travel	0.3
7	Misc	0.5

Figure 9

- Select both columns, including the top row.

- Click on the Chart Wizard (Figure 10) on the *Standard Toolbar* and select *Pie*.

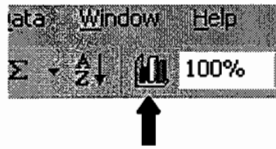


Figure 10

- Select the type you want from the examples. Then click on *Next>*.
- The next dialogue box confirms your choice of data. Click *Next>*.
- Next you get the *Chart Options* dialogue box.
  - Click on the *Title* tab, and type the title for the chart in the *Chart Title* box.
  - Click on the *Legend* tab, and remove the check mark beside *Show Legend*.
  - Click on the *Data Labels* tab and put a check mark beside *Category Name*.
 Then click *Next>*.
- Next you get the *Chart Location* dialogue box. If it is what you want, click on *Finish*.

Your final chart should look something like the one below:

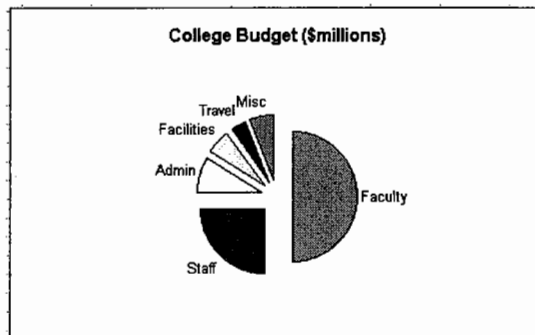


Figure 11

### 3.2.1 Editing a Chart

A chart may not appear exactly as you want it. It can be edited in a similar manner to the way we edited histograms.

### 3.2 Bar Charts

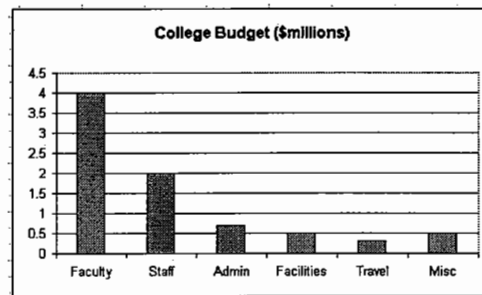
Bar charts can be constructed from frequency distributions tables.

#### Example 4

Construct a bar chart for the college budget in the previous example.

- As in the previous example, select both columns of data, and click on the Chart Wizard, but this time select *Column* instead of *Pie*. Select the type you want - likely the top left one. Click *Next>*.
- The next dialogue box confirms your choice of data. Click *Next>*.
- Next you get the *Chart Options* dialogue box.
  - Click on the *Title* tab, and type the title for the chart in the *Chart Title* box.
  - Click on the *Legend* tab, and remove the check mark beside *Show Legend*.
  - Click on the *Data Labels* tab and remove any checkmarks..
 Then click *Next>*.
- Next you get the *Chart Location* dialogue box. If it is what you want, click on *Finish*.

You can then edit as we did Histograms. The only difference is that we do not want the bars to touch. Your final chart should look something like the following:





## 4. Descriptive Statistics

Once you have entered your data into a worksheet, there are several ways to calculate appropriate descriptive statistics. One way is to enter the required formulas directly, another is to paste in the formulas using a template, and another way is to use a built in package supplied by *Excel*.

**Example 6.** Use each way to find descriptive statistics for the HSA data of example 1.

### 4.1 Using built-in functions

- Open the worksheet with the data from example 1
- In a convenient empty cell type the formula =AVERAGE(HSA) and press *Enter*. The mean should be displayed in the cell. It is a good idea to type “Mean” in the cell to the left.
- In a similar way you can enter other functions such as COUNT(HSA) (Sample Size); STDEV(HSA) (Standard Deviation), MEDIAN(HSA) (Median), MIN(HSA) (minimum), MAX(HSA) (Maximum), QUARTILE(HSA,1) (first quartile).

### 4.2 Using a Template

A template is a section of worksheet that contains the formulas needed for a particular calculation. In the worksheet *stattemplates.xls* (on the M drive in the computer labs) are a number of statistical templates. Each template is in a box with a title at the top. After copying it into your worksheet you need to edit the formulas in some of the cells to point to your data.

- Open the worksheet with the data from example 1
- Open the worksheet *stattemplate.xls*.
- Click on the tab *Miscellaneous*. Select the entire box titled Basic Descriptive Statistics and copy it into your worksheet. Most of the formulas will display error messages.

	A	B	C	D	E
1	HSA	GPA			
2	70	2.4		Basic Descriptive Statistics	
3	75	2.8			
4	78	3.3		Data Summary	
5	80	3.1		n	0
6	85	3.7		mean	#DIV/0!
7	82	3		std deviation	#DIV/0!
8	72	2.5		Five Number Summary	
9	65	2.3		minimum	0
10	68	2.8		Q <sub>1</sub>	#NUM!
11	75	3.1		median	#NUM!
12				Q <sub>3</sub>	#NUM!
13				maximum	0
14					
15					

- Press *Ctrl-`* (i.e. press the *Ctrl* key and the *`* (grave accent key at the top left of the keyboard) at the same time. This will make the formulas visible.

Basic Descriptive Statist	
Data Summary	
n	=COUNT(datad)
mean	=AVERAGE(datad)
std deviation	=STDEV(datad)
Five Number Summary	
minimum	=MIN(datad)
Q <sub>1</sub>	=QUARTILE(datad,1)
median	=MEDIAN(datad)
Q <sub>3</sub>	=QUARTILE(datad,3)
maximum	=MAX(datad)

- Replace each instance of *datad* with *HSA* (your name for the HSA data). This can be done by clicking on the cell, pressing *F2* and editing as you edit any text. When done editing all cells, press *Ctrl-`* again and you will see the results. You will need to adjust the widths of the columns for easy readability.

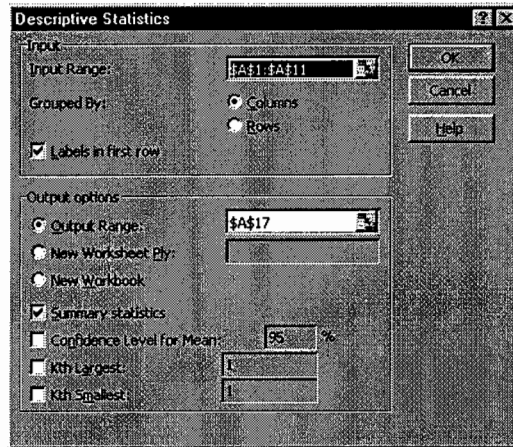
	A	B	C	D	E
1	HSA	GPA			
2	70	2.4		Basic Descriptive Statistics	
3	75	2.8			
4	78	3.3		Data Summary	
5	80	3.1		n	10
6	85	3.7		mean	75
7	82	3		std deviation	6.377042157
8	72	2.5			
9	65	2.3		Five Number Summary	
10	68	2.8		minimum	65
11	75	3.1		Q <sub>1</sub>	70.5
12				median	75
13				Q <sub>3</sub>	79.5
14				maximum	85
15					

### 4.3 Using *Excel's* Built-in Package

Starting from your worksheet with the HSA data:

- From the *Tools* menu, choose *Data Analysis*. (If it does not appear on the *Tools* menu, click on *Add-ins* and put a check beside *Analysis Toolpak* and click *OK*.) Then choose *Descriptive Statistics*.
- Enter the name *HSA* or *A1:A11* in the input range box.
- Put a check beside *Labels in First Row*.
- To insert the table into the current worksheet, click on the circle beside *Output Range* and enter a suitable cell in the box to the right. (You will need to allow for a table two columns wide and 15 rows deep.)

- Put a check beside *Summary Statistics*. The dialogue box will look like the figure below:



- Now click *OK*.

You will probably have to adjust the column widths. Your table should look like the following:

16		
17	<i>HSA</i>	
18		
19	Mean	75
20	Standard Error	2.016598
21	Median	75
22	Mode	75
23	Standard Deviation	6.377042
24	Sample Variance	40.66667
25	Kurtosis	-0.87975
26	Skewness	-1.2E-16
27	Range	20
28	Minimum	65
29	Maximum	85
30	Sum	750
31	Count	10

Do you know what all the entries mean? Look them up in your statistics textbook if necessary.

While this method is straight forward, it does not give you the quartiles. To find these you will need to use one of the first two methods.



---

## 5. Confidence Intervals For Means

---

*Excel* itself has no built-in method of finding confidence intervals. To do this you need to enter formulae as necessary, but this is time consuming and error prone. It is better to use templates, which are sections of a worksheet containing the necessary formulae that can be copied into your worksheet. Linking these with your data will give the required confidence intervals.

### Example 7

Construct a 90% confidence interval for the GPA's in example 1

As the population standard deviation is not known, we must use the *t* distribution template contained in the *statstemplates.xls* workbook.

- Open the workbook containing the data
- Open *statstemplates.xls*, and find the *Confidence Interval for Mean using t Distribution* template. (It will be on the worksheet titled *Conf - 1 pop.*) Select the template using the mouse.
- Select *Copy* from the *Edit* menu.
- Now switch to the workbook with the data. Choose a convenient place for the template and click on the cell in the upper left corner.
- Choose *Paste* from the *Edit* menu. Click and the template will appear in your workbook.

<b>Confidence Interval for Mean using t Distribution</b>	
<b>Data Summary</b>	
n	0
Mean	#DIV/0!
StdDev	#DIV/0!
<b>User Input</b>	
Conf Level	
<b>Confidence Interval</b>	
Lower Limit	#DIV/0!
Upper Limit	#DIV/0!

The first column contains names, while the second column contains the corresponding formulas. For example the formula beside *n* is =COUNT(datacm). When you first copy the template, these cells will display error messages. "DIV/0!" means an illegal division by 0. This happens because datacm is not the name of the data in your worksheet.

To view and edit the formulas in the worksheet:

- Press *Ctrl-`* (i.e. press the *Ctrl* key and the *`* (grave accent key at the top left of the keyboard) at the same time.

C	D
<b>Confidence Interval for Mean using t Distribution</b>	
<b>Data Summary</b>	
	n=COUNT(datacm)
	Mean=AVERAGE(datacm)
	StdDev=STDEV(datacm)
<b>User Input</b>	
	Conf Level

- Replace each instance of *datacm* with *GPA* (your name for the GPA data). This can be done by clicking on the cell, pressing *F2* and editing as you edit any text. Enter the confidence level (0.90 for 90%) into the cell beside *Conf Level*.

C	D
<b>Confidence Interval for Mean using t Distribution</b>	
<b>Data Summary</b>	
	n=COUNT(GPA)
	Mean=AVERAGE(GPA)
	StdDev=STDEV(GPA)
<b>User Input</b>	
	Conf Level 0.9

- When done, press *Ctrl-`* again and you will see the results. You will need to adjust the widths of the columns for easy readability.

C	D
<b>Confidence Interval for Mean using t Distribution</b>	
<b>Data Summary</b>	
	n 10
	Mean 2.9
	StdDev 0.43204938
<b>User Input</b>	
	Conf Level 0.9
<b>Confidence Interval</b>	
	Lower Limit 2.649548968
	Upper Limit 3.150451032

The 90% confidence interval for the

mean GPA is from 2.64 to 3.16.

## 6. Hypothesis Testing of Mean of One Sample

The general procedure is:

- Choose the appropriate test and copy the corresponding template to your workbook
- If you have not already named the data, highlight it and choose *Insert, Name, and Define*.
- Edit the formulas in the template to change *datahm* into the name of your data.
- Input your null hypothesis and  $\alpha$ .

### Example 8.

Consider the GPA data from example 1. Test (at the 5% significance level) the claim that the mean GPA in the population is greater than 2.5.

In this example, if  $\mu$  is the mean GPA in the population, we are testing the hypotheses:  $H_0: \mu \leq 2.5$  and  $H_1: \mu > 2.5$ .

- Paste the appropriate template into your workbook.

Hypothesis Test for Mean	
<b>Data Summary</b>	
n	=COUNT(datahm)
Mean	=AVERAGE(datahm)
StdDev	=STDEV(datahm)
<b>User Input</b>	
H <sub>0</sub> Mean	
Alpha	

If you examine the formulas you will find that there are three that refer to the data, and two to the hypotheses. Once these are fixed, the template will perform the calculations.

- If necessary name the data GPA. Then click on the formula for  $n$ , press  $F2$ , and edit *datahm* to read GPA. Do this also for the formulas for mean and standard deviation.

Hypothesis Test for Mean	
<b>Data Summary</b>	
n	10
Mean	2.9
StdDev	=STDEV(datahm)
	STDEV(number1, [number2], ...)
<b>User Input</b>	
H <sub>0</sub> Mean	
Alpha	

Now choose  $H_0$  Mean and  $Alpha$

- Enter 2.5 for  $H_0$  Mean and 0.05 for  $Alpha$

The template computes the statistics for left, right and two-tailed tests. From  $H_1$ , our test is right-tailed. Thus the critical value is 1.833114, and the p value is 0.0084. As p is less than alpha, we reject the null hypothesis, and can conclude that the mean GPA is greater than 2.5. (Alternatively, because the  $t$  for the test data is 2.9277 which is to the right of 1.833114, we reject the null hypothesis.)

Note that you can make this more readable, by deleting the parts of the template (here the Left-tail and Two-Tail tests) that you don't need.

Hypothesis Test for Mean	
<b>Data Summary</b>	
n	10
Mean	2.9
StdDev	0.43204938
<b>User Input</b>	
$H_0$ Mean	2.5
Alpha	0.05
<b>Computed Values</b>	
df	9
t	2.927700219
<b>Left-Tail Test</b>	
Left Critical t	-1.833113856
p-value	0.991590098
<b>Right-tail Test</b>	
Right Critical t	1.833113856
p-value	0.008409902
<b>Two-tail Test</b>	
Critical t	2.262158887
p-value	0.016819804

## 7. Hypothesis Tests for Two Samples

Several two sample hypothesis tests are built into *Excel*: comparison of means (dependent and independent samples) and comparison of variances. Unfortunately the ones for means are only-right tailed tests or two-tailed tests, so you need to pick the variables carefully to avoid a left-tailed test. Also, the one for independent samples assuming unequal variances miscalculates the number of degrees of freedom. The templates in *statstemplates.xls* avoid these problems.

### 7.1 Comparison of means, dependent samples

#### Example 9

The Central City Insurance Company is concerned about high estimates of clean up time received from Royal Restorers. A random sample of 15 clean up operations is taken from their records and the estimates given by Royal and by Continental Cleanup, a rival company, are recorded. The company wishes to test whether Royal's estimates are higher than those of Continental at the 5% significance level.

Job	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R	7.6	10.2	9.5	1.3	3.0	6.3	5.3	6.2	2.2	4.8	11.3	12.1	6.9	7.6	8.4
C	7.3	9.1	8.4	1.5	2.7	5.8	4.9	5.3	2.0	4.2	11.0	11.0	6.2	6.7	7.5

If  $\mu_R$  and  $\mu_C$  denote the population mean estimates for Royal and Continental, the hypotheses are  $H_0: \mu_R \leq \mu_C$  and  $H_1: \mu_R > \mu_C$ . These may be re-written as  $H_0: \mu_R - \mu_C \leq 0$  and  $H_1: \mu_R - \mu_C > 0$

#### 7.1.1 Using Templates

In the dependent case, you do a hypothesis test on the difference between the values. Proceed as follows:

- Enter the data and use labels like Royal and Continental in the first row to name the data.
- In another column enter the label Difference at the top. In the second cell, enter the formula =Royal-Continental and fill down. Choose *Insert, Name, Define* to name the differences.
- Now follow the steps from Chapter 6: Paste the template titled *Hypothesis Test for Mean* into the worksheet. Edit the first three formulas to refer to the name Difference. Enter 0 for  $H_0$  Mean and 0.05 for  $\alpha$ .

	A	B	C	D
1	Royal	Continental	Difference	
2	7.6	7.3	0.3	
3	10.2	9.1	1.1	
4	9.5	8.4	1.1	
5	1.3	1.5	-0.2	
6	3	2.7	0.3	
7	6.3	5.8	0.5	
8	5.3	4.9	0.4	
9	6.2	5.3	0.9	
10	2.2	2	0.2	
11	4.8	4.2	0.6	
12	11.3	11	0.3	
13	12.1	11	1.1	
14	6.9	6.2	0.7	
15	7.6	6.7	0.9	
16	8.4	7.5	0.9	

The template should now look like the one at the right.

Since the p value is  $1.63 \times 10^{-5}$ , which is less than 0.05, we reject the null hypothesis and conclude that Royal does have longer clean up times. (Alternately, as the test statistic is 5.997 which is bigger than the right critical value of 1.761, so we would reject the null hypothesis.)

### 7.1.2 Using Excel's built in tool

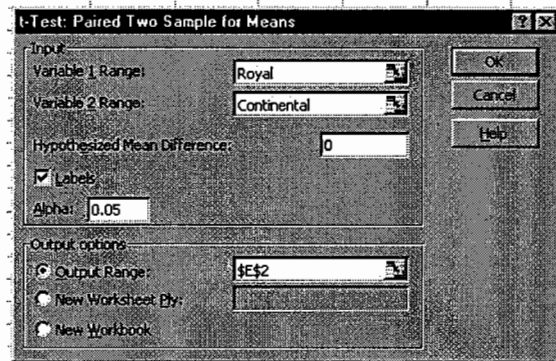
- Enter the data and use labels like Royal and Continental in the first row to name the data.
- Select the *Tools* menu and select *Data Analysis*. (If it does not appear on the *Tools* menu, click on *Add-ins* and put a check beside *Analysis Toolpak* and click OK.)
- Select *t-Test: Paired Two Sample For Means* and click OK.

Before you can use this feature, you need to decide which data will be variable 1 and which will be variable 2. As the alternate hypothesis is  $H_1: \mu_R - \mu_C > 0$ , and Excel will test  $H_1: \mu_1 - \mu_2 > 0$ , the Royal data will be variable 1 and the Continental data variable 2.

- Enter *Royal* or A1:A16 into the box beside *Variable 1 Range* and *Continental* or B1:B16 into the box below it. Enter 0 for *Hypothesized Mean Difference*: and 0.05 for *Alpha*. Put a check in the *Labels* box. Finally enter a suitable location for the output range, and click OK.

From the output below, we come to the same conclusion as when using the template.

Hypothesis Test for Mean	
<b>Data Summary</b>	
n	15
Mean	0.606666667
StdDev	0.391821145
<b>User Input</b>	
H <sub>0</sub> Mean	0
Alpha	0.05
<b>Computed Values</b>	
df	14
t	5.996638843
<b>Left-Tail Test</b>	
Left Critical t	-1.76130925
p-value	0.999983632
<b>Right-tail Test</b>	
Right Critical t	1.76130925
p-value	1.63682E-05
<b>Two-tail Test</b>	
Critical t	2.144788596
p-value	3.27364E-05



t-Test: Paired Two Sample for Means		
	Royal	Continental
Mean	6.846666667	6.24
Variance	10.26552381	8.649714286
Observations	15	15
Pearson Correlation	0.995522509	
Hypothesized Mean Difference	0	
df	14	
t Stat	5.996638843	
P(T<=t) one-tail	1.63682E-05	
t Critical one-tail	1.76130925	
P(T<=t) two-tail	3.27364E-05	
t Critical two-tail	2.144788596	

## 7.2 Comparison of Means, Independent Samples

There are two possible cases, depending on whether the variances are equal or unequal. Both can be handled by the templates in *statstemplates.xls*. They can also be handled by the *Data Analysis* package in *Excel*, **but the number of degrees of freedom in the unequal variance case is wrong**.

In this section we look at the equal variance test. The unequal variance case is similar.

### Example 10.

A new treatment for a type of leaf eating caterpillar that attacks miniature decorative maple trees is tested. A sample of 14 trees is obtained; half of these are treated, and the other half are left untreated. One week later the trees are inspected and the number of caterpillars on each tree is found. The results are shown below:

Treated	18	43	28	50	16	32	13
Untreated	40	54	26	63	21	37	39

Can we say that the treatment works (fewer caterpillars) at the 5% significance level?

If  $\mu_T$  and  $\mu_U$  denote the population mean number of caterpillars for treated and untreated trees, the hypotheses are:  $H_0: \mu_T \geq \mu_U$  and  $H_1: \mu_T < \mu_U$ . These may be re-written as  $H_0: \mu_T - \mu_U \geq 0$  and  $H_1: \mu_T - \mu_U < 0$

Assuming that the populations are normally distributed and that the variances are equal (check first using descriptive statistics), proceed as follows:

- Enter and name the data as usual.
- Paste the template from *statstemplates.xls* titled *Two independent Samples assuming equal variances* into your workbook. In this case there are 6 formulas that need to be edited to refer to the data, plus 0 for  $H_0$  mean difference and 0.05 for *Alpha*.

Two independent Samples assuming equal variances		
<b>Data Summary</b>	Sample 1	Sample 2
n	=COUNT(Treated)	=COUNT(Untreated)
mean	=AVERAGE(Treated)	=AVERAGE(Untreated)
std deviation	=STDEV(Treated)	=STDEV(Untreated)
<b>User Input</b>		
$H_0$ mean difference	0	
Alpha	0.05	

The final output should look like the following:

Since we are doing a left tail test and the p value is 0.0815 which is bigger than 0.05, we would fail to reject the null hypothesis, and conclude that there is not enough evidence to show that the treated trees have fewer caterpillars.

Two independent Samples assuming equal variances		
<b>Data Summary</b>	Sample 1	Sample 2
n	7	7
mean	28.571429	40
std deviation	14.093227	14.67424
<b>User Input</b>		
H <sub>0</sub> mean difference	0	
Alpha	0.05	
<b>Computed Values</b>		
df	12	
t	-1.486161	
<b>Left-Tail Test</b>		
Left Critical t	-1.782287	
p-value	0.0815151	
<b>Right-tail Test</b>		
Right Critical t	1.7822867	
p-value	0.9184849	
<b>Two-tail Test</b>		
Critical t	2.1788128	
p-value	0.1630303	

### 7.3 Comparison of Variances

To do a hypothesis test comparing two variances, you need to do an F test. In the last example we assumed that the variances were equal. Was this a reasonable assumption?

#### Example 11.

Test whether it is reasonable to assume that the variances in example 10 are equal.

We will take the following hypotheses:  $H_0: \sigma_T = \sigma_U$  and  $H_1: \sigma_T \neq \sigma_U$  and use  $\alpha = 0.05$ .

- Paste the template titled *Comparison of Two Variances* into the same worksheet as the previous example.

Comparison of Two Variances		
<b>Data Summary</b>	Sample 1	Sample 2
n	=COUNT(Treated)	=COUNT(Untreated)
variance	=VAR(Treated)	=VAR(Untreated)
<b>User Input</b>		
Alpha	0.05	

- Edit the four formulas above, and input *Alpha*. The output is shown at right.

The test statistic is 0.9224, and the p value is 0.9244, so fail to reject the null hypothesis. There is not enough evidence to show that the variances are different, so we can assume that they are equal.

Warning: The assumption that the data is normally distributed is crucial for this test, and should be checked first.

Warning. If you use *Excel's* built in package, it will only do a right-tailed test, so you will need to change  $\alpha$  and be sure that variable 1 is the data with greater variance.

Comparison of Two Variances		
<b>Data Summary</b>	Sample 1	Sample 2
n	7	7
variance	198.619	215.3333
<b>User Input</b>		
Alpha	0.05	
<b>Computed Value</b>		
F	0.922379	
<b>Left-Tail Test</b>		
Left Critical F	0.233435	
p-value	0.462187	
<b>Right-tail Test</b>		
Right Critical F	4.283862	
p-value	0.537813	
<b>Two-tail Test</b>		
Left Critical F	0.171829	
Right Critical F	5.819743	
p-value	0.924375	

## 8. Confidence Intervals/Hypothesis Tests for Proportions

There are templates in the template workbook which will calculate confidence intervals and do hypothesis tests for proportions. As with means, you can calculate in terms of a data set (generally preferred), or enter the summary values. Note that you first need to check that the samples and proportions are large enough to allow approximation using the normal distribution.

### Example 12

When 40 people were sampled about whether they support a ban on smoking in restaurants the following results were obtained. Find a 95% confidence interval for the proportion in the population that supports such a ban.

Yes, Yes, No, No, Yes, No, No, Yes, No, Yes, Yes, No, Yes, No, Yes, No, Yes, No, Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, No

- Put the data into a column, together with a Label in the first row, and use this to name the data.
- Paste the template into the worksheet.

Confidence Interval for Proportion	
Data Summary	
	n =COUNTA(SmokeBan)-1
	x =COUNTIF(SmokeBan,"Yes")
	p-hat =D8/D7

- Edit the first two formulas to refer to your data and enter the confidence level.

	A	B	C	D
1	SmokeBan			
2	Yes			
3	Yes			
4	No		Confidence Interval for Proportion	
5	No			
6	Yes		Data Summary	
7	No		n	40
8	No		x	24
9	Yes		p-hat	0.6
10	No			
11	Yes		User Input	
12	Yes		Conf level	0.9
13	No			
14	Yes		Confidence Interval	
15	No		Lower Limit	0.472590198
16	Yes		Upper Limit	0.727409802
17	No			

The 90% confidence interval for the population proportion is from 47.2% to 72.8%.

### Example 13

In a sample of 254 male college students, 55 were taking business, while a 68 of a sample of 261 female college students were taking business. Test the hypothesis that the proportion of male college students who take business is the same as the proportion of female college students. Use a significance level of 0.05.

- In this case we have:  $H_0: p_1 = p_2$   
and  $H_a: p_1 \neq p_2$
- Copy the template into your workbook.
- Enter the summary statistics in the correct cells.

As the  $p$  value is 0.242, we fail to reject the null hypothesis, and conclude that there is not enough evidence to show that the proportion of male students taking business is different from the proportion of female students.

Two Proportions	
Data Summary	
n 1	254
x 1	55
p_hat_1	0.216535433
n 2	261
x 2	68
p_hat_2	0.260536396
User Input	
H0Mean Diff	0
Alpha	0.05
Computed Value	
z	-1.170866397
Left-tail Test	
Left Critical z	-1.644853476
p-value	0.120826299
Right-tail Test	
Right Critical z	1.644853476
p-value	0.879173701
Two-tail Test	
Critical value	1.959962787
p-value	0.241652598

---

## 9. Scatter Plots, Correlation and Regression

---

### 9.1 Plotting a Scatter Plot

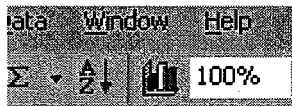
#### Example 1 (Revisited)

A college advisor wishes to study the relationship between incoming students' high school averages (HSA's) and their first semester college grade point averages (GPA's). A sample of 10 students is obtained and their HSA's and GPA's are recorded, in the following table.

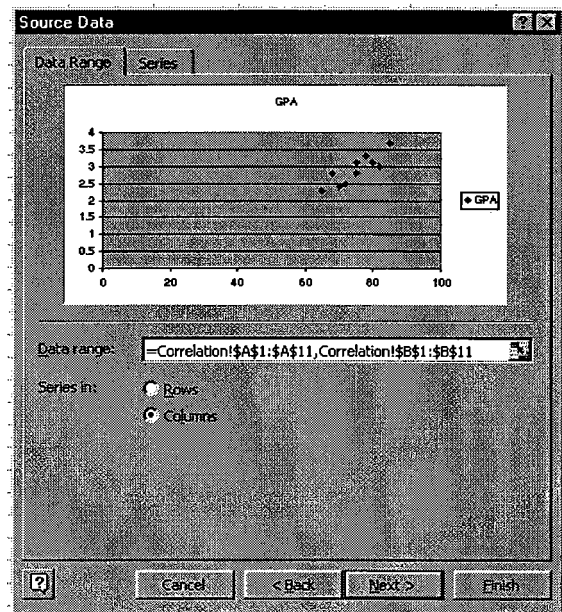
HSA	70	75	78	80	85	82	72	65	68	75
GPA	2.4	2.8	3.3	3.1	3.7	3.0	2.5	2.3	2.8	3.1

To visually see the relationship, you need to draw a scatter plot. This is done as follows:

- Open the workbook with the data. (You may want to copy the data to a clean worksheet.)
- Click on the *Chart Wizard* on the standard toolbar.



- Choose *XY (Scatter)*, and then click on *Next>*.
- In the *Chart Source Data* dialogue box, the cursor should be in the *Data Range* box. Use the mouse to select the column corresponding to the horizontal axis (in this case HSA), and then **while holding down the *Ctrl* key**, use the mouse to select the data corresponding to the vertical axis (in this case GPA). The box should look like the one at right:
- Click on *Next>*.

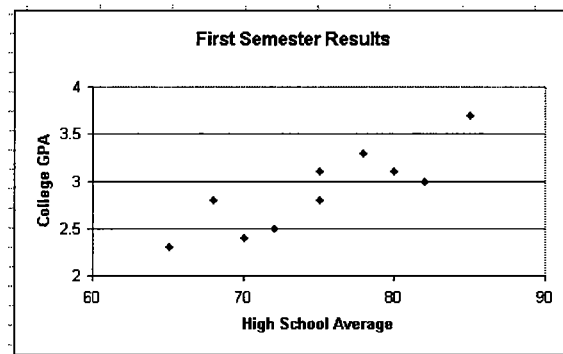


- You are now at the *Chart Options* dialogue box. Under the *Title* tab, enter a suitable title, and labels for the two axes. Under the *Legend* tab remove the check beside *Show Legend*. Click on *Finish*.

You will now need to edit the graph (as you did with histograms in Chapter 3) to make it useful and clear.. The biggest problem is that all the points are in one corner of the graph.

- Right click on the vertical axis and choose *Format Axis*.
- Choose the *Scale* tab and remove all the checks under *Auto*. Then set the minimum to 2 and the maximum to 4.
- Take similar steps to minimum and maximum values for the *X* axis.
- If necessary adjust the size of the plot by dragging the sizing handles to appropriate positions.

Other formatting options are possible. Your chart should look like the following:



## 9.2 Finding the Correlation

The scatter plot seems to indicate that there is a linear relationship between the high school average and the first semester gpa. We will find the correlation coefficient and, if it is significant, the linear regression equation.

- From the *Tools* menu, choose *Data Analysis*. (If it does not appear on the *Tools* menu, click on *Add-ins* and put a check beside *Analysis Toolpak* and click *OK*.) Then choose *Regression*. A dialogue box will open.
- Enter the *X* and *Y* ranges. We want High School Averages to be the independent (*X*) variable, and Grade Point Average to be the dependent (*Y*) variable. Enter either the names or ranges in the input boxes.
- Check *Labels*, if you have included the labels in the ranges.
- Enter the *Output Range*. (Note that the output will be quite large. Leave lots of room.)
- Choose a confidence level, say 95% ( $\alpha = 0.05$ )

### 9.3 Finding the Regression Line

The coefficients of the regression line are given in the output on the previous page. The intercept ( $b_0$  or  $\beta_0$ ) is in the third table immediately to the right of the word *Intercept*, while the slope is in the same table immediately to the right of the word *HSA*. Thus the regression equation is:

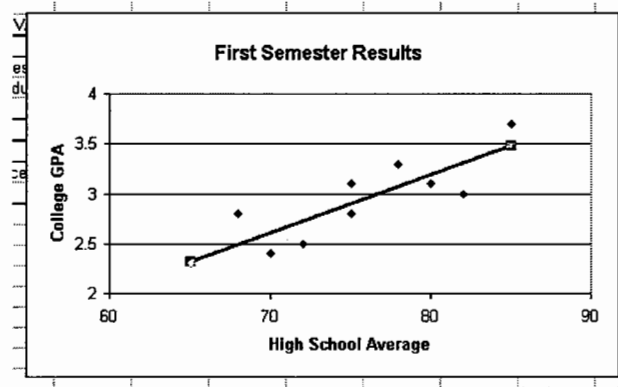
$$\text{predictedGPA} = -1.465 + 0.0582 \text{ HSA}$$

To understand the rest of the output, you may need to refer to your textbook.

### 9.4 Plotting the Regression Line

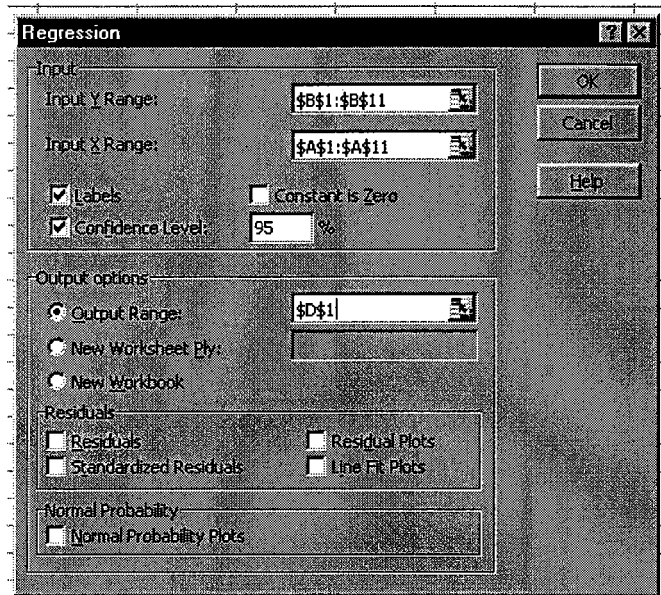
Now we want to add this line to the scatter plot we made in section 9.1. To do this:

- Click on the plot, and then point the cursor at any point on the graph. Right-click and choose *Add Trendline*. Make sure *Linear* is selected and click *OK*.



The line plotted is the regression line calculated above.

- Leave all other squares blank.



- Click OK.

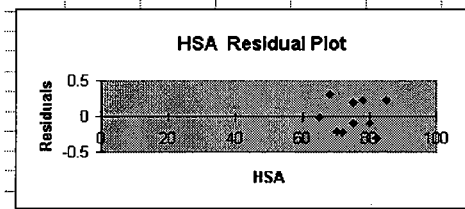
SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.858983							
R Square	0.737851							
Adjusted R	0.705083							
Standard Error	0.23463							
Observations	10							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	1.23959	1.23959	22.51703	0.001454			
Residual	8	0.44041	0.055051					
Total	9	1.68						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	-1.46475	0.922811	-1.58727	0.151112	-3.59276	0.663252	-3.59276	0.663252
HSA	0.058197	0.012264	4.745211	0.001454	0.029915	0.086478	0.029915	0.086478

This *Excel* tool allows for multiple regression, i.e. deciding whether a variable depends on several other variables. Thus, under *Regression Statistics* the correlation is labelled *Multiple R*. Thus the correlation is 0.8590. For a sample of size 11 is this significant? See the textbook.

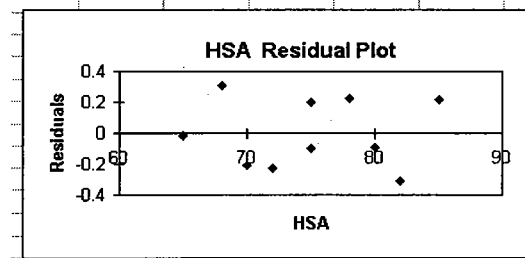
Warning: *Excel* gives the absolute value of the correlation coefficient. It is up to you to determine if it is positive or negative.

## 9.5 Plotting Residuals

To get a plot of the residuals, you just need to put a check mark beside *Residual Plots* in the dialogue box we used to calculate the correlation in section 9.2. The plot is as follows:



Of course the graph needs to be reformatted to look something like:





## 10. Chi-Square Tests for Variances and Standard Deviations

**Example 15.** Find the 95% confidence interval for variance and standard deviation of the HSA data from example 1.

As before:

- Open both the template file and the workbook with the data. Copy the *Confidence Interval for Variance and Standard Deviation* template from the template file into the worksheet with the data.
- Edit the two formulas in the template to change *datacc* into *HSA* (the name of the data).

<b>Confidence Interval for Variance and Standard Deviation</b>	
<b>Data Summary</b>	
	n=COUNT(datacc)
	Std Deviation=STDEV(datacc)
<b>User Input</b>	
	Conf Level

- Enter the confidence level (0.95).

The worksheet should now appear as below:

<b>Confidence Interval for Variance and Standard Deviation</b>	
<b>Data Summary</b>	
	n 10
	Std Deviation 6.377042157
<b>User Input</b>	
	Conf Level 0.95
<b>Confidence Interval for Variance</b>	
	Lower Limit 19.24009242
	Upper Limit 135.536042
<b>Confidence Interval for Std Deviation</b>	
	Lower Limit 4.386352974
	Upper Limit 11.64199476

Thus the 95% confidence interval for the variance is from 19.24 to 135.54 and the one for standard deviation is from 4.38 to 11.65.



## 11. Testing Data For Being Normally Distributed

To perform some calculations you need to know whether data is normally distributed.

### Example 16.

Determine whether the following data is normally distributed:

58, 71, 61, 73, 65, 77, 79, 66, 77, 82, 68, 54, 78, 76, 80, 83, 74, 75, 47

The data first needs to be sorted in ascending order.

- Type the data into column C of a blank worksheet, with the name Data in C1. Make sure columns B and D are empty.
- Highlight the column of data, and Click on *Insert*, then *Name*, then *Define*. Then click *OK*.
- Click on *Data* (on the *Menu Bar*), then *Sort*. A dialogue box will appear.
- If the right data is highlighted, click on *OK*. If not, click cancel, highlight the data and repeat.

To create columns A and B:

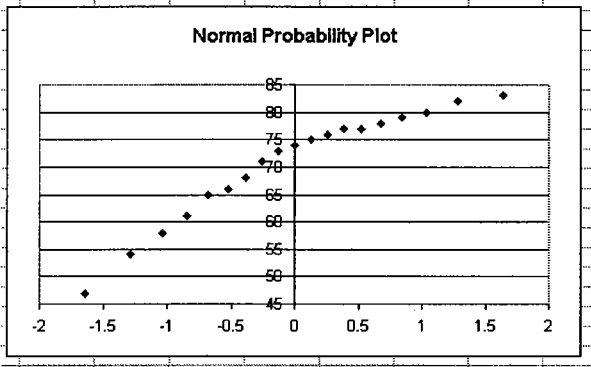
Type Rank, 1, and 2 into the first three cells of column A. Highlight the 1 and 2, and drag down the lower right corner square to complete the column.

- In B1, type Zvalue
- In B2 type the formula `=NORMSINV(A2/(COUNT(Data)+1))`
- Select B2 and drag down using the corner box to create the remaining entries in column 2.

Your worksheet should look like the one at the right..

	A	B	C
1	Rank	Zvalue	Data
2	1	-1.64485	47
3	2	-1.28155	54
4	3	-1.03643	58
5	4	-0.84162	61
6	5	-0.67449	65
7	6	-0.5244	66
8	7	-0.38532	68
9	8	-0.25335	71
10	9	-0.12566	73
11	10	5.47E-10	74
12	11	0.125661	75
13	12	0.253347	76
14	13	0.385321	77
15	14	0.5244	77
16	15	0.67449	78
17	16	0.841621	79
18	17	1.036433	80
19	18	1.281552	82
20	19	1.644853	83
21			

To check for normality, create a scatter plot of columns B and C. See section 9. If the data is normally distributed, the points would be on a straight line.



In this case, the graph is not linear, and thus we would conclude the data is not normally distributed. (See your text for further details.)

---

## 12. Presentations

---

In the previous sections we discussed how to use *Excel* to enter data, create graphics and carry out various statistical tests. The result is often a large worksheet that is too wide and long to print. In addition, the form of the spreadsheet is not suitable for inclusion in a document. For projects and reports it is necessary to use a word processor to create the text and then copy-and-paste relevant parts of the spreadsheet (charts, data, results of statistical tests) into this document. (See section 1.)

When writing a report, bear in mind the following points:

- Tables of data and plots may be re-sized after being pasted into a report, by clicking on the object and dragging the sizing handles. This permits these objects to be arranged side-by-side, or in groups, for easy comparison. The frequency tables, tables of statistics and plot for a data set are often presented side-by-side.
- When sets of data are to be compared, frequency tables and plots should use the same classes.

At times you may want to include the whole of the workbook as an appendix to your document. This requires you to pay careful attention to the layout of the workbook as you create it. You may need to provide captions or titles or to number charts and tables so that they can be referred to in the body of your document. If your workbook is wider than your printer can print, *Excel* will divide it horizontally and vertically which could be very confusing to your reader. To avoid this problem, move tables and charts around as you create them.

What is important is that your output is clear and easily read by the audience to whom it is directed.

---

## Appendix 1: Location of Files

---

The file containing the templates is available in the student computer labs, on the Richmond and Surrey campuses. Its name is *stattemplates.xls*. It will probably be on the M drive.

---

## Appendix 2: *Excel* Functions used in Statistics

---

### DESCRIPTIVE STATISTICS

AVERAGE(*data*) - mean (arithmetic) of the numbers in *data*  
MEDIAN(*data*) - median of the numbers in *data*  
MODE(*data*) - mode of the numbers in *data*  
STDEV(*data*) - sample standard deviation of the numbers in *data*

- STDEVP(*data*) - population standard deviation of the numbers in *data*
- MIN(*data*) - minimum value in *data*
- MAX(*data*) - maximum value in *data*
- QUARTILE(*data*,*n*) - *n*<sup>th</sup> quartile of the numbers in *data*. Warning: Excel® uses a different formula to calculate the first and third quartile than our textbook does.

## CONTINUOUS PROBABILITY DISTRIBUTIONS

- NORMSDIST(*z*) - cumulative standard normal distribution, i.e.  $P(Z < z)$
- NORMSINV(*prob*) - inverse of the standard normal distribution, i.e. *a* where  $P(Z < a) = prob$ .  
This is sometimes denoted  $z_{1-prob}$ . To find critical values use NORMSINV(1 - *alpha*)
- NORMDIST(*x*,*mu*,*sigma*,true) - cumulative nonstandard normal distribution
- NORMINV(*prob*,*mu*,*sigma*) - inverse of nonstandard distribution
- TDIST(*t*,*df*,1) - complement of cumulative *t* distribution with *df* degrees of freedom for  $t \geq 0$  only, i.e.  $P(t_{df} > t)$ .
- TINV(*prob*,*df*) - the critical value  $t^*$  which satisfies  $P(t > t^*) = prob/2$ , sometimes denoted  $t_{df, prob/2}$ .  
Note the divided by 2.
- CHIDIST(*x*,*df*) - complement of the cumulative  $\chi^2$  distribution with *df* degrees of freedom, i.e.  $P(\chi_{df}^2 > x)$
- CHIINV(*prob*,*df*) - the critical value  $\chi_{df, prob}^2$

### Examples:

- =NORMSDIST(1) - gives 0.84134474 =  $P(Z < 1)$
- =NORMSINV(0.95) - gives 1.644853 =  $z_{0.05}$
- =NORMSINV(1-0.1) - gives 1.281550794 =  $z_{0.1}$
- =NORMDIST(200,150,30,true)-NORMDIST(125,150,30,true)  
- gives 0.749881345 =  $P(125 < X < 200)$ , if  $X \sim N(150,30)$
- =2\*(1-NORMDIST(165,150,30/SQRT(15),true))  
- gives .052807373, the *P* value for a two tailed test when a sample of size 15 has a mean of 165, when taken from a population with mean 150 and standard deviation 30
- =TDIST(1.5,24,1) - gives 0.073327823 =  $P(t_{24} > 1.5)$
- =TINV(.1,24) - gives 1.710882316 =  $t_{24,0.05}$

=CHIDIST(15.7,10) - gives 0.108548343 =  $P(\chi_{10}^2 > 15.7)$

=CHIINV(0.025,14) - gives 26.11893491 =  $\chi_{14,0.025}^2$

---

## DISCRETE PROBABILITY DISTRIBUTIONS

BINOMDIST( $x,n,p$ ,false) - Binomial distribution ( $P(X = x)$ , where  $X$  has binomial distribution with  $n$  trials, and the probability of a success is  $p$ )

BINOMDIST( $x,n,p$ ,true) - Cumulative Binomial distribution ( $P(X \leq x)$ , where  $X$  has binomial distribution with  $n$  trials, and the probability of a success is  $p$ )

HYPGEOMDIST( $x,n,S,N$ ) - Hypergeometric distribution ( $P(X = x)$ , where  $X$  is the number of successes in a sample of size  $n$  drawn from a population of size  $N$  containing  $S$  successes.)

POISSON( $x,\mu$ ,false) - Poisson distribution ( $P(X = x)$ , where  $X$  has Poisson distribution with mean  $\mu$ .)

POISSON( $x,\mu$ ,true) - Cumulative Poisson distribution ( $P(X \leq x)$ , where  $X$  has Poisson distribution with mean  $\mu$ .)

### Examples:

Suppose that  $X$  has binomial distribution with  $n = 10$  and  $p = 0.2$

=BINOMDIST(3,10,0.2,false) - gives 0.201326592 =  $P(X = 3)$

=BINOMDIST(3,10,0.2,true) - gives 0.879126118 =  $P(X \leq 3)$

Suppose that  $X$  has Poisson distribution with  $\mu = 4.7$

=POISSON(3,4.7,false) - gives 0.15738316 =  $P(X = 3)$

=POISSON(3,4.7,true) - gives 0.30968357 =  $P(X \leq 3)$

---

## CORRELATION AND REGRESSION

CORREL( $ydata,xdata$ ) - Pearson Product Correlation coefficient between the sets of data in  $ydata$  and  $xdata$ .

SLOPE( $ydata,xdata$ ) - slope of the regression line (line of best fit) when using the data in  $xdata$  to predict the data in  $ydata$ .

INTERCEPT( $ydata,xdata$ ) - intercept of the regression line (line of best fit).

---

## MISCELLANEOUS FUNCTIONS

RAND() - gives a random number uniformly distributed between 0 and 1 This can be combined with other probability distributions for a random number with different distributions.

SUM( $data$ ) - gives the total of the numbers in  $data$ .

`SUMPRODUCT(xdata,ydata)` - multiplies each data value in *xdata* by the corresponding value in *ydata* and adds up the resulting products. This is useful if you want to find the expected value of a probability distribution, or when you want to find the mean of a frequency distribution.

#### Examples

`=NORMSINV(RAND())` - gives a random number with standard normal distribution.