

**Part 1. Descriptive Statistics**

**Introduction.**

In statistical analysis, the term "descriptive statistics" refers to the methods of describing a set of data using statistical measures, charts, and graphs. In this first part, you will learn how to use Excel to compute some common statistical measures.

**Part 1 Outline.**

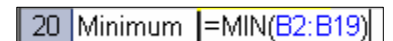
1. Enter data into a spreadsheet.
2. Compute the MIN and MAX, and use these to compute the range of data values.
3. Compute COUNT, SUM, AVERAGE (also called the mean in statistics), and STANDARD DEVIATION of the set of data.
4. Compute the MEDIAN of the data values.

1. Start with a new spreadsheet. Enter your name in cell A1. In cells B2 through B16, denoted B2:B16 in Excel, enter these data:  
 4, 9, 15.3, 7, 25, 43, 2, 36.7, 8, 3, 40, 57, 29, 8, 19, 9, 19.9, 22.4  
 noting that some of these have decimal point values which are used below to illustrate the problem with bin values.

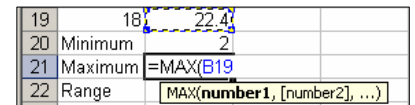
	A	B
1	Your name here	
2	1	4
3	2	9
4	3	15.3
5	4	7
6	5	25
7	6	43
8	7	2
9	8	36.7
10	9	8
11	10	3
12	11	40
13	12	57
14	13	29
15	14	8
16	15	19
17	16	9
18	17	19.9
19	18	22.4

2. In cells A2:A19, enter the numbers from 1 to 18.
3. In cell A20, enter the label "Minimum:", and in cell A21, enter "Maximum:" and in cell A22, enter "Range:".

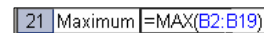
4. In cell B20, enter the function =MIN(B2:B19). You must begin any formula with an equal sign, "=". This is how Excel knows that the rest of what you are typing is a formula  
 Note the value that is displayed. What is it? \_\_\_\_\_



5. In cell B21, start entering the function =MAX( ending with the left hand parenthesis. Then click on cell B19. It will be highlighted with a twinkling border, and the cell address B19 will be entered after your left parenthesis in the =MAX(B19 in cell B21, as illustrated in the picture here.



Now move the mouse cursor on top of the small square box in the upper right corner of cell B19. The mouse cursor turns into a small double headed diagonal arrow. Left click and drag the two headed arrow up to cell B2. Note that this highlights the border of this range of cells and also prints this range after the left parenthesis in your =MAX(B19 entry in cell B18. You now finish this function by typing the right parenthesis, so now the cell B21 should contain the function =MAX(B2..B19). Press Enter to store this formula in cell B21  
 Cell B21 now displays the maximum value in the range B2:B19.  
 What is the maximum value displayed? \_\_\_\_\_



	A	B
1	Your name here	
2	1	4
3	2	9
4	3	15.3
5	4	7
6	5	25
7	6	43
8	7	2
9	8	36.7
10	9	8
11	10	3
12	11	40
13	12	57
14	13	29
15	14	8
16	15	19
17	16	9
18	17	19.9
19	18	22.4
20	Minimum	2
21	Maximum	=MAX(B2:B19)

6. In cell B22, enter the formula for the difference between the cells B21 and B20, that is, type the formula **=B21-B20**.

Your spreadsheet should look like this in cells A20:B23.

Press Enter to store this formula in cell B22. What is its value? \_\_\_\_\_

20	Minimum	2
21	Maximum	57
22	Range	=B21-B20

7. In row 23, column A, enter the label "Median". Then in column B, enter the formula for the median of the values in B2 through B19 using the builtin function **MEDIAN**. What do you actually enter in this cell? Write it below:

formula in cell B23: \_\_\_\_\_

value in cell B23: \_\_\_\_\_

8. Why is this the median for these 18 values? To see why, copy the values in B2 through B19 to cells C2:C19. Now sort the data values in cells C2:C19 in ascending order (highlight the cells C2:C19, then click on the **Data** main menu item, then click **Sort**, click "Continue with current selection", if it asks, then sort on column C ascending).

The point to sorting the data values is that the "median" of the data values is defined to be the "middle-most" value, but since there are an even number of data values, 18 of them, then there is no middle value, and so the median is computed to be the average of the two middle values, in this case the 9th and 10th values, 15.3 and 19. This makes the median equal to  $(15.3+19)/2 = 17.15$ . Your spreadsheet should look like this:

	A	B	C
1	Your name here		
2	1	4	2
3	2	9	3
4	3	15.3	4
5	4	7	7
6	5	25	8
7	6	43	8
8	7	2	9
9	8	36.7	9
10	9	8	15.3
11	10	3	19
12	11	40	19.9
13	12	57	22.4
14	13	29	25
15	14	8	29
16	15	19	36.7
17	16	9	40
18	17	19.9	43
19	18	22.4	57
20	Minimum	2	
21	Maximum	57	
22	Range	55	
23	Median	17.15	

9. In cells A24, A25, A26 and A27, type the labels "Count", "Sum", "Average" and "Std Dev". Then in cells B24, B25, B26 and B27, enter these formulas for the count, sum and average of the values in cells B2:B19

In B24, enter: **=COUNT(B2:B19)**

In B25, enter: **=SUM(B2:B19)**

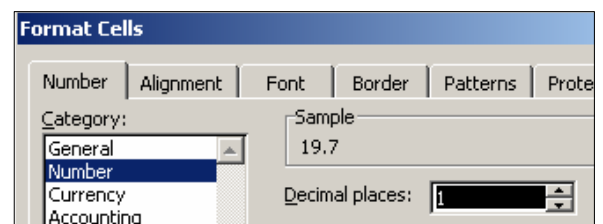
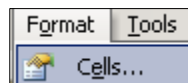
In B26, enter: **=AVERAGE(B2:B19)**

In B27, enter: **=STDEV(B2:B19)**

These statistical functions will compute the count of the values in cells B2 to B19, their sum, their average (or mean), and the standard deviation of the values. The standard deviation is a measure of how spread out the values are over their range. The result should look like this:

	A	B
20	Minimum	2
21	Maximum	57
22	Range	55
23	Count	18
24	Sum	357.3
25	Average	19.73889
26	Std Dev	15.86588

If you would prefer to have the number of decimal points limited to one for the average and standard deviation, then highlight the cells B25 and B26 and click on the **Format** main menu item, then click on **Cells**, and in the **Format Cells** window, click on **Number** under Category, and click or type 1 in the Decimal places box. Then click the **OK** button. The average and standard deviation values should now be displayed with just one decimal point of accuracy, rounded.



	A	B
25	Average	19.7
26	Std Dev	15.9

## Exercise 1

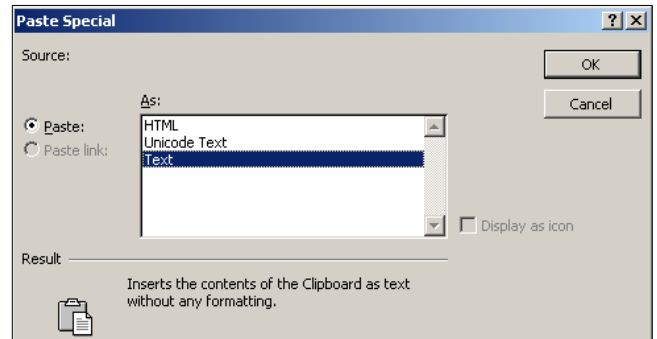
Start Internet Explorer or another web browser, and type in the internet address (URL)

<http://cba.winthrop.edu/fosterk/>

Click on the **NFL.dat** item on that web page. (These are statistics from the [www.NFL.com](http://www.NFL.com) web site for all teams in the National football League for the regular 2005 season.) The data file will be opened as a text file. Highlight all of the text in the file and click Edit | Copy (or click Edit | Select All) to copy this text into the clipboard.

```
Team, G, Plys, Yds/G, Y/P, FD/G, 3rd Md, 3rd Att, 3rd %, 4th Md,
Seattle, 1, 66, 413.0, 6.3, 24.0, 5, 12, 41.7, 0, 2, 0, 9, 61, 28
San Diego, 1, 77, 408.0, 5.3, 24.0, 6, 15, 40.0, 0, 0, 0, 9, 75,
Indianapolis, 2, 116, 403.0, 6.9, 22.5, 13, 22, 59.1, 0, 0, 0, 8
```

Next, in Excel, open a new spreadsheet, and since the cursor is by default, in cell A1, now click **Edit | Paste Special** and click the **Text** radio button, then click **OK**. The NFL data should now be inserted into your spreadsheet with the data values in columns.

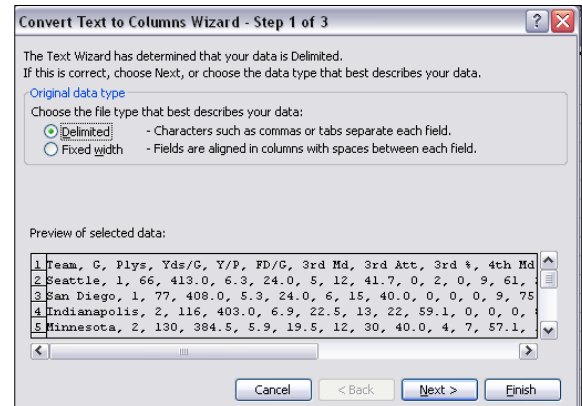
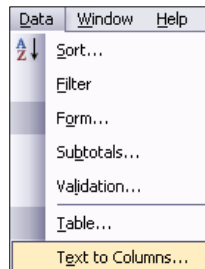


If the **Paste Special** separates the data into individual columns, then skip the next part, and go to the next page.

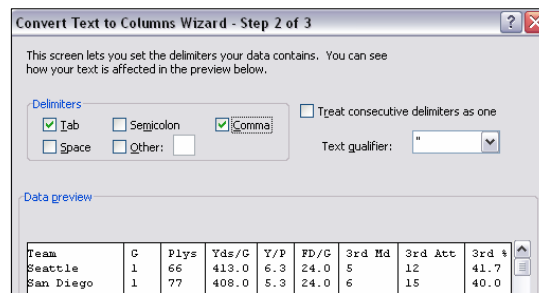
If the **Paste Special** option does not put the individual columns of the NFL data into individual columns

	A	B	C	D
1	Team, G, Plys, Yds/G, Y/P, FD/G, 3rd M			
2	Seattle, 1, 66, 413.0, 6.3, 24.0, 5, 12, 41			
3	San Diego, 1, 77, 408.0, 5.3, 24.0, 6, 15,			
4	Indianapolis, 2, 116, 403.0, 6.9, 22.5, 13			

in the spreadsheet (all of the data will be in only column A), then you can convert the data in column A to separate columns by clicking on the **Data** main menu item, then clicking on the **Text to Columns** option. You then get the **Convert Text to Columns – Step 1 of 3** window. Click the **Delimited** radio button if it is not already selected, then click next.

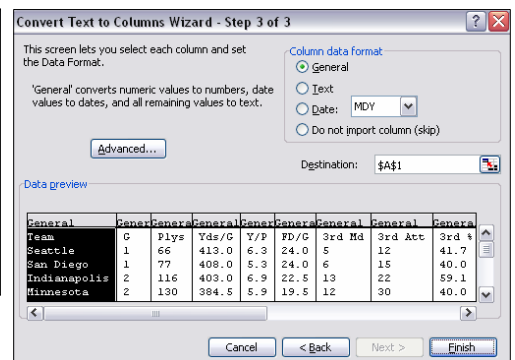


In the **Step 2 of 3** window, click the **Commas** check box in the **Delimiters** box. The data in column A should be displayed in separate columns as illustrated here:



Click **Next**.

The **Step 3 of 3** window allows you to select a data type for each column, or to not include (import) any column. For this data set, you should not need to do anything, so just click **Finish**. The NFL data should now be separated into columns as illustrated here:



	A	B	C	D	E	F
1	Team	G	Plys	Yds/G	Y/P	FD/G
2	Seattle	1	66	413	6.3	24
3	San Diego	1	77	408	5.3	24
4	Indianapol	2	116	403	6.9	22.5

## Exercise 1 (continued)

After you have the NFL data divided into columns, in the sequence of cells below the **Yds/G** (yards per game) column, enter the formulas for the following statistics:

- The count of the teams
- The sum of the yards for all of the teams
- The average of the yards per game for all of the teams
- The standard deviation of the yards per game for the teams
- The maximum yards per game for the teams
- The minimum yards per game for the teams
- The range of the yards per game for the teams

## Exercise 2

Start Internet Explorer or another web browser, and type in the internet address (URL)

<http://cba.winthrop.edu/fosterk/>

Click on the **Employment.xls** item on that web page. (These are statistics from the Bureau of Labor Statistics web site).

In the row of cells for the year 2005, below this table of data, compute these statistics.

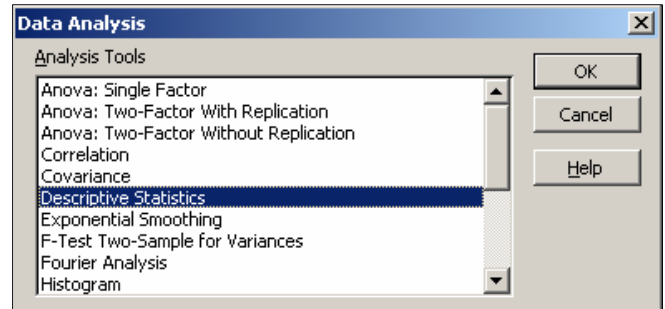
- The average of the employment values
- The standard deviation of the values
- The maximum of the values
- The minimum of the values
- The range of the values

## Using the Data Analysis Option to Compute the Descriptive Statistics

An alternative to entering each of the statistical functions in individual cells is to use the Univariate option in the Data Analysis menu. Begin by clicking on the **Tools** main menu item, and then, probably way at the bottom of the Tools menu, you will find the Data Analysis option. Click on it. If the Data Analysis is not in the Tools drop down menu, then add it by clicking Tools | Add-Ins, then check Analysis Toolpak item, and then click OK. The Data Analysis window is displayed.

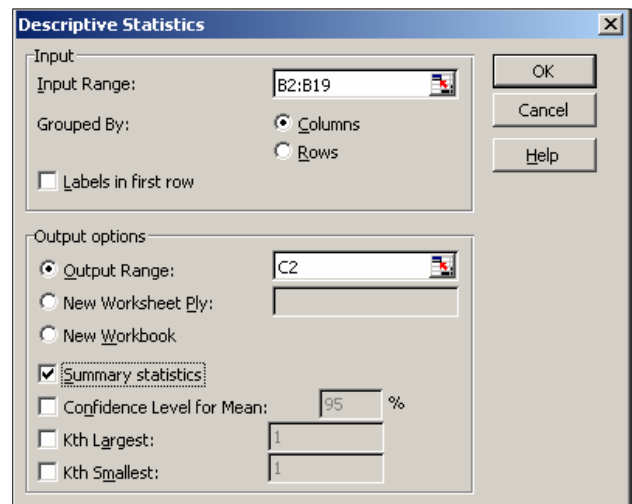
Data Analysis...

Find and click on the **Descriptive Statistics** item.



The **Descriptive Statistics** window will be displayed.

In the **Input Range:** box, enter the cell range **B2:B19**, the cells that contain the 18 data values. Click on the **Grouped by:** Columns radio button. Then in the **Output range:** box, type **C2**.



The columns C and D will be used to display the descriptive statistics for the data values in cells B2 to B19. These statistics include, in this order, the Mean (or average), Standard Error (the label is cut off since the column width is not sufficiently large), the median, the mode, the standard deviation, and so on. More than we want to know, probably.

C	D
	<i>Column1</i>
Mean	19.85
Standard E	3.739624
Median	17.15
Mode	9
Standard D	15.86588
Sample V	251.7262
Kurtosis	0.071977
Skewness	0.911495
Range	55
Minimum	2
Maximum	57
Sum	357.3
Count	18

### Exercise 1 (continued)

Using the NFL data that you got from Exercise 1, use the **Descriptive Statistics** option of the **Data Analysis** option on the **Tools** main menu item to compute the descriptive statistics for the **Yds/G** (yards per game) column below that column.

### Exercise 2 (continued)

Using Employment data you got in Exercise 2, use the **Descriptive Statistics** option of the **Data Analysis** option on the **Tools** main menu item to compute the descriptive statistics for the employment values for the year 2005 row. Locate these statistics in a column below the table of data values.

## Part 2. Frequency Distributions

A "frequency distribution" is a way of summarizing the many values in the list of a data set into a relatively short list of intervals of values for the data set, together with the count of the values in the data set in each interval. For example, for the data set with the grade values **67, 71, 79, 90, 91**, suppose the four intervals considered are **0 to 69, 70 to 79, 80 to 89, and 90 to 100**. The frequency distribution, the cumulative frequency distribution, the relative frequency distribution in which the frequency values of the frequency distribution are divided by the total number (5 in this case), and the relative cumulative frequency distribution of this data set for these intervals is summarized by the list:

Interval	Values in Interval	Frequency	Cumulative Frequency	Relative Frequency	Rel. Cum. Frequency
0 - 69	67	1	1	0.2	0.2
70 - 79	71, 79	2	3	0.4	0.6
80 - 89	none	0	3	0.0	0.6
90 - 100	90, 91	2	5	0.4	1.0

In this part, you will develop these frequency distributions for example data sets.

### Part 2 Outline.

0. Enter the data values.
  1. Enter the right hand end points of the intervals, called the bin values.
  2. Create the frequency distribution using bin values.
  3. Create the cumulative frequency distribution of percentages.
  4. Create the a chart of the frequency distribution and the cumulative frequency percentage distribution.

0. Click on the File main menu item, and then click on New to create a new spreadsheet. Alternatively, you can click on the icon on the far left of the tool bar that looks like a blank sheet. Enter your name in cell A1. Enter these values in cells A2..A6:

67, 71, 80, 90, 91

	A
1	Your name
2	67
3	71
4	79
5	90
6	91

1. Now select a block of cells in which you will create the list of bin values. Excel will create the list of frequencies for the specified range of cells for the data set based on the interval endpoints specified by the bin values. Here, the term "bin values" means the right hand endpoints of the intervals for the distribution. We will put the bins block to the right of the data cells, but you could also put the block of cells below, or anywhere.

In the four cells **B2:B5**, enter these bin values for the right end points of the intervals: **69, 79, 89, 100**.

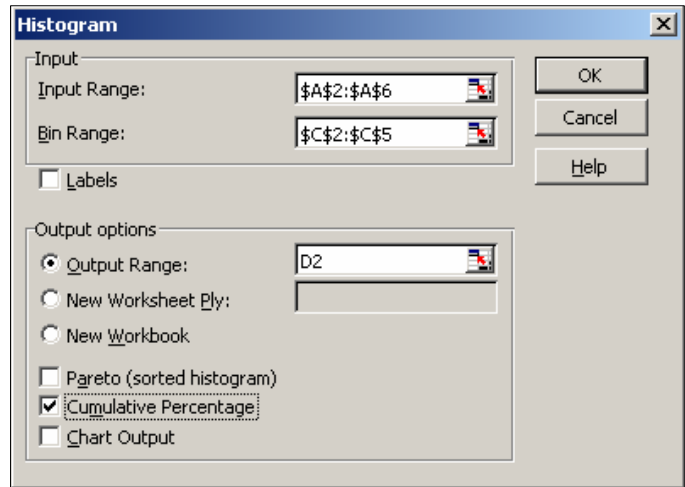
	A	B	C
1	Your name		
2	67		69
3	71		79
4	79		89
5	90		100
6	91		



Next, to create the frequency distribution values for this data set, click the **Tools** menu item, then **Data Analysis**, then choose the **Histogram** option.

In the **Histogram** window, you fill in these items:

- First, enter the data values cell range A2 to A6 in the **Input Range:** box. You can type it as **A2:A6** but Excel will change it to **\$A\$2:\$A\$6**.
- Next, enter the bin range, **C2:C5** in the **Bin Range:** box.
- Now, enter a cell address in the **Output Range:** box. Here, cell D2 was entered, and the cells below and to the right of D2 will be used for the output.
- If you want the cumulative percentages, also called the relative cumulative frequency in statistics, then click the **Cumulative Percentage** check box.
- Click OK.



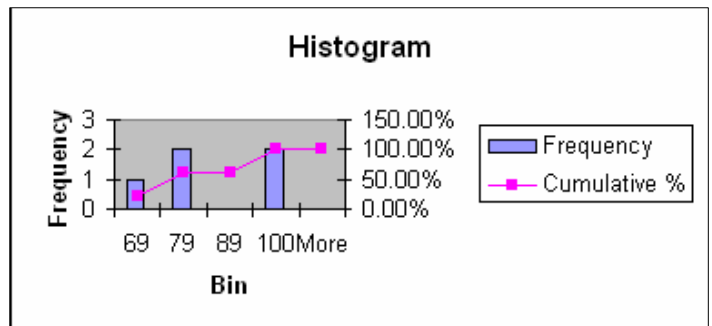
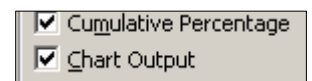
The output is located starting in cell D2. In column D, the bin values are listed. Notice that the bin values are repeated in the output range, and there is no attempt by Excel to line them up with your bin values in column C. That is, the bin values in column C are not part of the output – those values were used only by Excel to create the frequency distribution in the output range.

Bin	Frequency	Cumulative %
69	1	20.00%
79	2	60.00%
89	0	60.00%
100	2	100.00%
More	0	100.00%

In the next column, E, the frequency values are listed. Here, a “1” next to the bin value “69” means that one of the data values in the data cells (A2:A6) is less than or equal to 69, that is, the bin value of 69 is the high value of that interval. Then the 2 next to the bin value of 79 means that there are two values greater than 69 and less than or equal to 79, the values 71 and 79. There are no value above 79 and less than to equal to 89, and two valued above 89 and less than or equal to 100. There are no values above 100.

In column F (the width of column F was expanded to allow the title to fit), the cumulative frequency percentages are listed. Here, 20% of the data values are up to and include 69, and 60% of the data values are up to 79 (three of the five values are up to 79, namely 67, 71 and 79). Then still 60% of the values (same 3 out of the 5) are up to 89, and 100% of the data values are up to 100.

Let’s look at one more option in the Histogram window – the graph option. Again, click Tools, Data Analysis, and Histogram. All of the Histogram window options that you chose before are still selected. In addition, click the **Chart Output** check box, and then **OK**. In addition to the output table with the bin, frequency and cumulative percentage columns, a graph of a chart of the histogram is output to the right of the table. This chart includes two graphs, a “histogram” or bar chart of the data values frequencies in the bin intervals, and a line graph of the cumulative percentages for these intervals. The vertical axis on the left represents the range of the data values, and the vertical axis on the right represents the range of percentages.



### EXERCISE 3

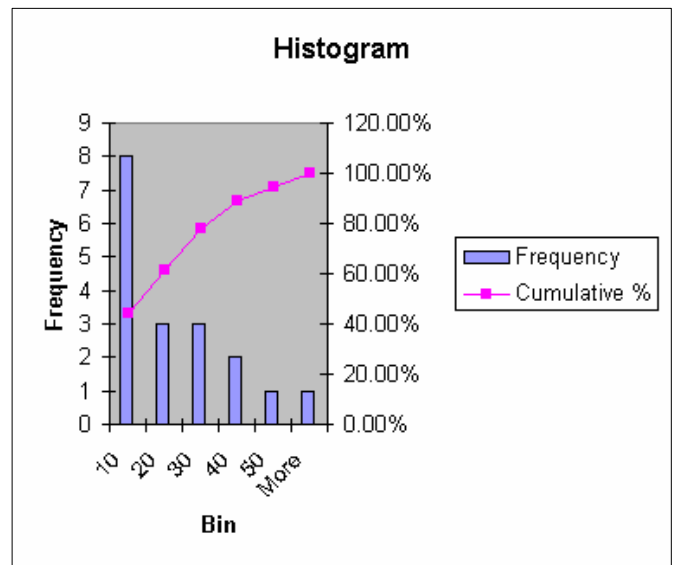
Start a new spreadsheet, entering your name in cell A1, and entering the 18 values from part 1 in cells A2 to A19. Then move your mouse pointer to cell C2. Create a range of bin values that will represent intervals 0-10, 10.1-20, 20.1-30, and so on up to 40.1-50. What is the first bin value? Answer: 10.

What is the last bin value? Answer: \_\_\_\_\_

Now create the frequency distribution for the data values in the range A2 to A19 using the bin values in the range column C. If you do all of this successfully, then your Histogram output should look something like this:

Bin	Frequency	Cumulative %
10	8	44.44%
20	3	61.11%
30	3	77.78%
40	2	88.89%
50	1	94.44%
More	1	100.00%

Your graph will look too squatty to be meaningful. Enlarge it by clicking on the graph, and then put the cursor on the small black box on the bottom border of the graph and drag it downwards. You should get a graph that looks like this:



### Exercise 1 (continued)

Using the NFL data from Exercise 1, choose five intervals of equal lengths for a frequency distribution for the Yds/G yards per game column. Enter the upper endpoint values of your intervals as bin values in the column below the yards per game values. Then create a frequency distribution and cumulative percentage distribution and a chart of these distributions.

### Exercise 2 (continued)

Using the Employment data from Exercise 1, choose five intervals of equal lengths for a frequency distribution for the year 2005 row. Enter the upper endpoint values of your intervals as bin values in column A below the data values. Then create a frequency distribution and cumulative percentage distribution and a chart of these distributions.

## Plotting Related Data with Scatter Plot

In some situations, we want to determine if one type of variable is related to the values of another type of variable. An initial exploration of the data can be done by plotting pairs of data value to see if there is a visually identifiable relationship. In Excel, you can plot pairs of data using the scatter plotting option in the Chart Wizard. To illustrate, we will use the data from the Nation Assessment of Education Progress (NAEP) available from the text book *Statistics for Business and Economics, 9ed*, by Anderson, Sweeney, and Williams. You can download the Excel file of this data from <http://cba.winthrop.edu/fosterk/>.

	A	B	C
1	State	Spending per Pupil (\$)	Composite Score
2	Alabama	3,777	604
3	Arizona	4,041	618
4	Arkansas	4,060	615
5	California	4,917	580
6	Colorado	4,772	644
7	Connecticut	7,629	657
8	Delaware	6,208	615

The data is ordered by state, and if we want to investigate the relationship between the **Spending per Pupil** variable and the **Composite Score** variable, begin by re-sorting the data in ascending order by the **Spending** column, so that your sorted spreadsheet looks like this.

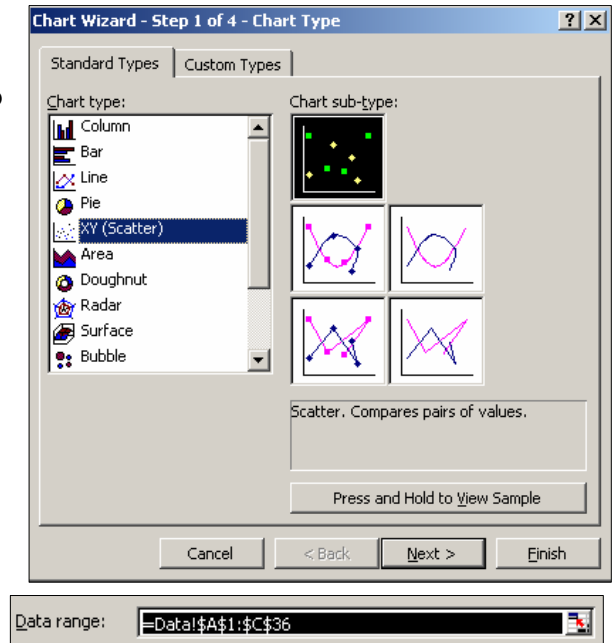
	A	B	C
1	State	Spending per Pupil (\$)	Composite Score
2	Utah	3,280	650
3	Mississippi	3,423	582
4	Alabama	3,777	604
5	Tennessee	3,800	618
6	Arizona	4,041	618
7	Louisiana	4,049	581

Click the **Chart Wizard** icon or click **Insert | Chart**.

You get the Chart Wizard displayed. Click on the **XY (Scatter)** chart type. The default option of the five options is to plot the pairs of points, which is described in the text box as "Scatter. Compares pairs of values."

Click the **Next** button.

By default, the **Data Range** selected by Excel is the whole spreadsheet.



We want to make the data range the two columns for the Spending variable and the Score variable.

You can type in the range in the **Data range:** box, but an easier way to specify it is to click on the

data range button at the right end of the **Data range:** box, then highlight the

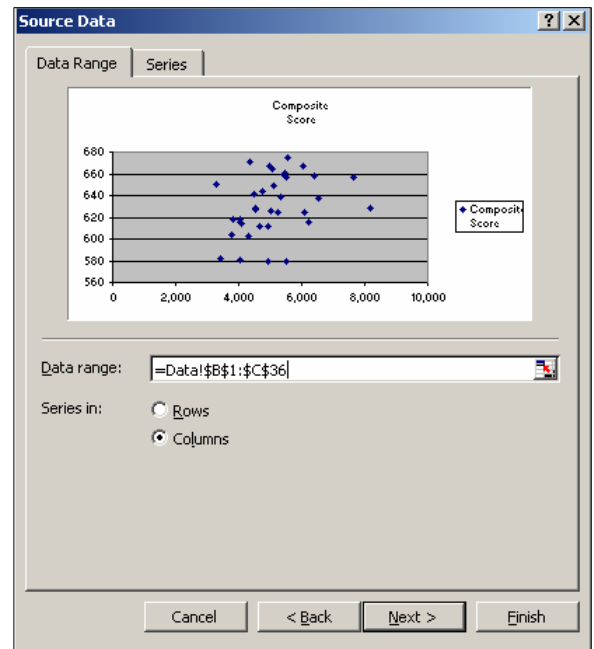
	A	B	C	D	E	F	G	H	I
1	State	Spending per Pupil (\$)	Composite Score						
2	Utah	3,280	650						
3	Mississippi	3,423	582						

Spending and the Score columns, including the column headers, so that these two columns now have a twinkling border, and Excel fills in the data range addresses for you in the **Data range:** box. To finish selecting this data range, click the data range button at the right end of the box again

The **Source Data** window is now displayed with the data range that you selected, and with a plot of the pairs of data value in this data range, and should look like this:

Also, notice that Excel has selected that the series of data is listed in columns by choosing the **Columns** radio button in the **Series in:** option. If the data had been listed along two rows, we would have selected the **Rows** option.

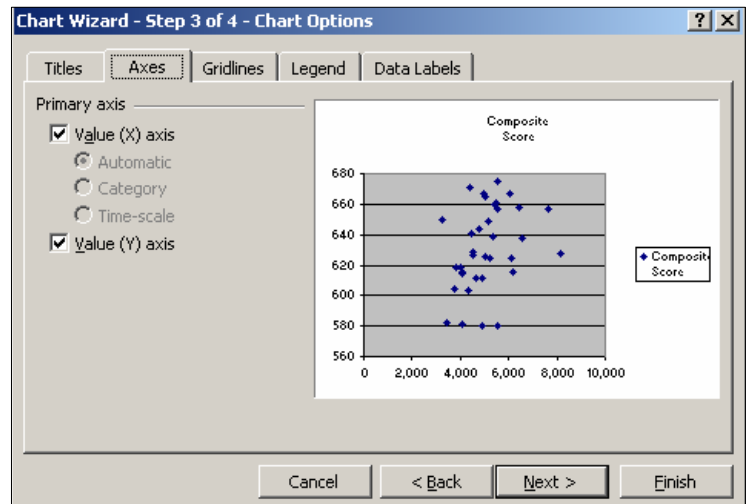
Click **Next**.



The **Chart Options** window is displayed.

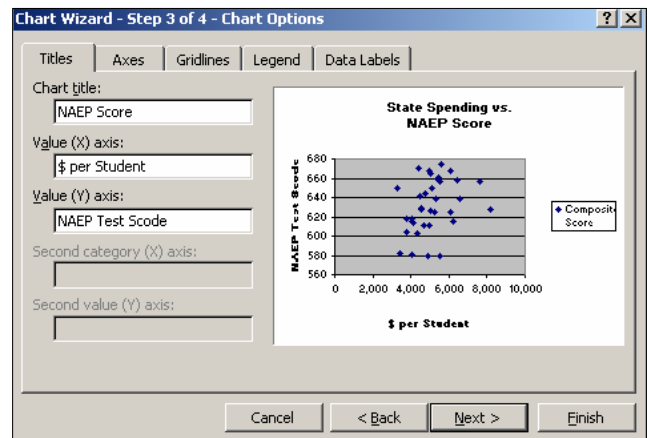
The **Axes** option tab is selected here with both check boxes selected for the **Value(X) axis** and the **Value(Y) axis**. Uncheck one of them and notice the change in the chart to the right. The values along that axis are no longer displayed. Check that box again and note the change to displaying the values along the axis.

Click the **Titles** tab.

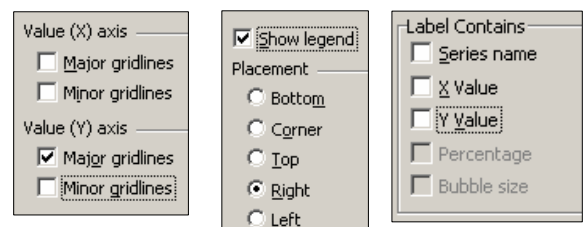


Fill in the three titles as illustrated.

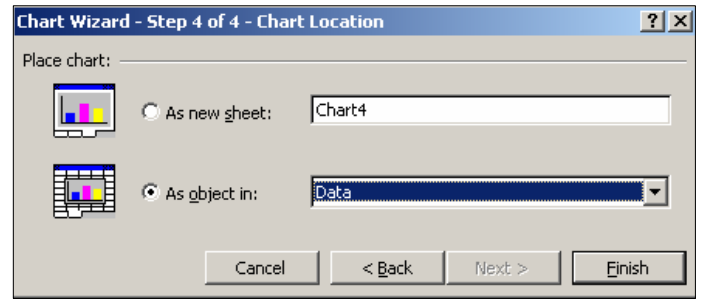
If the wizard window changes to the next window during this process, just click **Back** to get back to this **Chart Options** window.



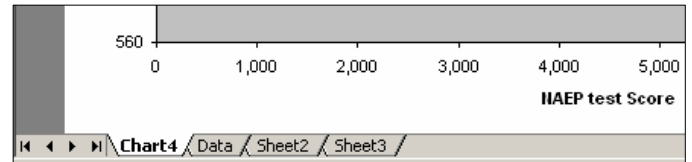
Now click the **Gridlines** tab. The **Major gridlines** option is the only one checked in this illustration. Try checking and unchecking the other options and observe the effect on the graph. Similarly, on the **Legend** tab, you can choose to show or not show the legend and choose its position. Finally, the **Data Labels** tab allows you to choose to have the data values displayed with the data points.



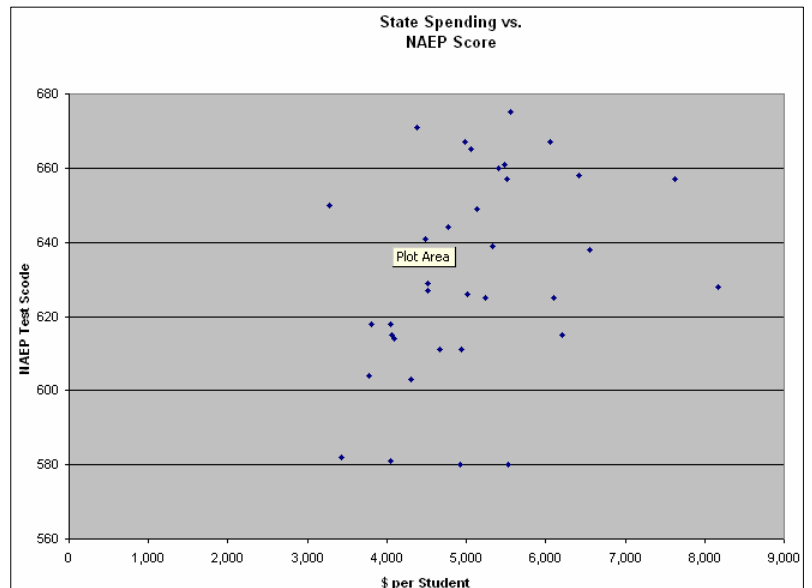
Click **Next** to get the **Chart Location** wizard window to choose whether you want the graph put in the current spreadsheet (as an object) or put into a separate spreadsheet in this workbook.



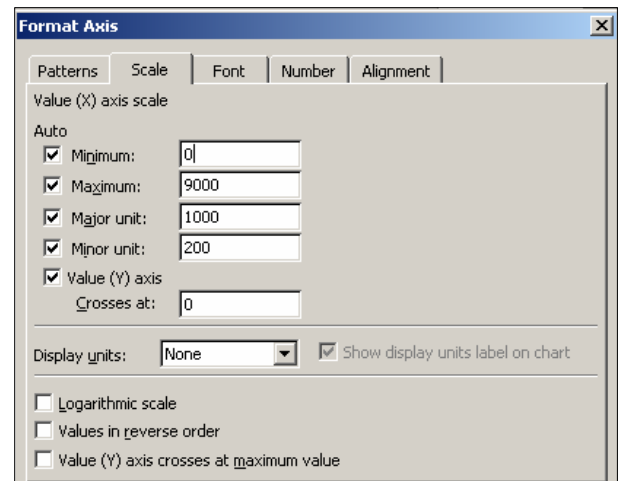
If you choose **As a new sheet**, then Excel create a new spreadsheet with a tab at the bottom of the workbook window as illustrated here with the sheet titled "Chart 4".



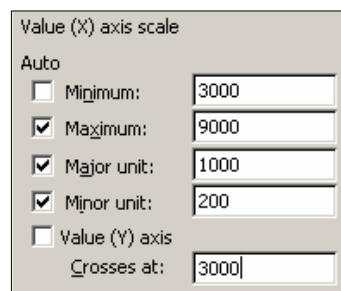
The resulting scatter plot is displayed here to illustrate a problem - Excel has selected the NAEP score axis values to start with 0 at the left. However, the first spending per student value is over \$3000, and so the area on the graph between 0 and 3000 is wasted..



To fix this, you want to "format the x-axis". Begin by putting the cursor on the any of the x-axis values and double clicking to get the **Format Axis** window displayed .



Change the **Minimum** value to 3000 and change the **Crosses at:** value to 3000. Click **OK**.

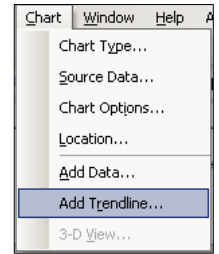


Your scatter plot is now spread out over an x-axis ranging from 3000 to 9000. The y-axis range could also be adjusted if desirable.

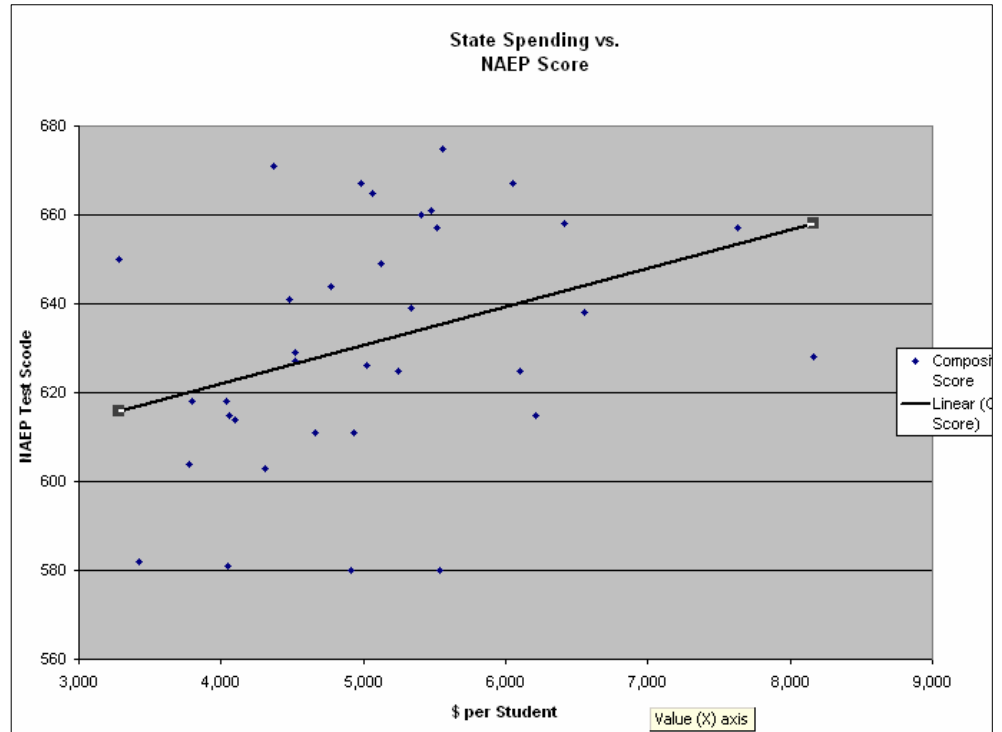
We will add one more feature to your chart.

Click the **Chart** main menu item, and then click the **Add Trendline** option.

This will add a trendline, also called the linear regression line, to your chart. This is a line that is a best fit of this data using a criterion called the least squares criterion. It is a topic in the QMTH 206 course.



Your graph should look something like this.



**Exercise 2 (continued)** For the NFL, data, create a scatter plot of the Plays variable and the Yards per Game variable. Include appropriate titles for your chart and the axes. Add a trendline. Does there appear to be a trend between plays and yards per game?