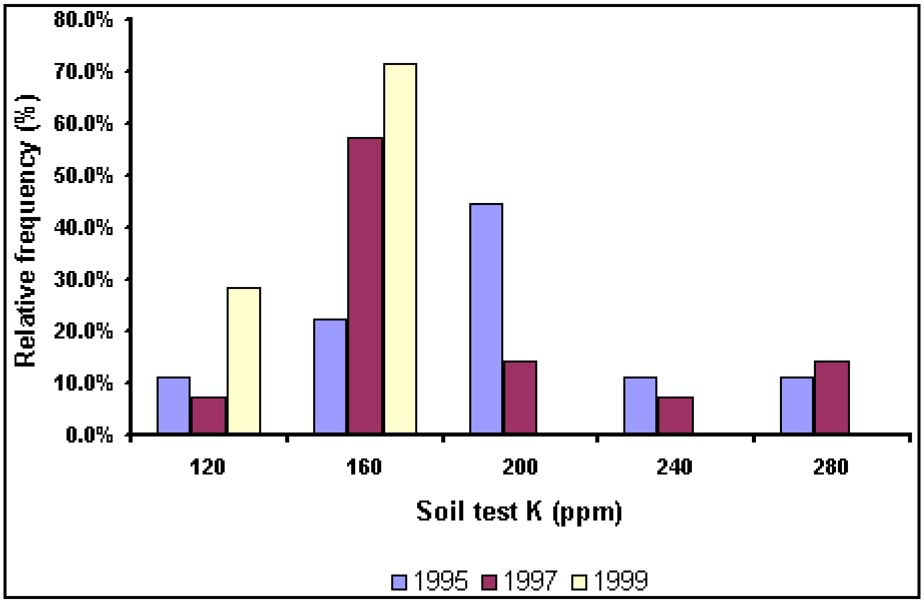
Finding Trends in Soil Test Data



Scott Murrell and Harold Reetz
Potash & Phosphate Institute

Lance Murrell
The Andersons, Inc.

Quentin Rund
PAQ Interactive, Inc.

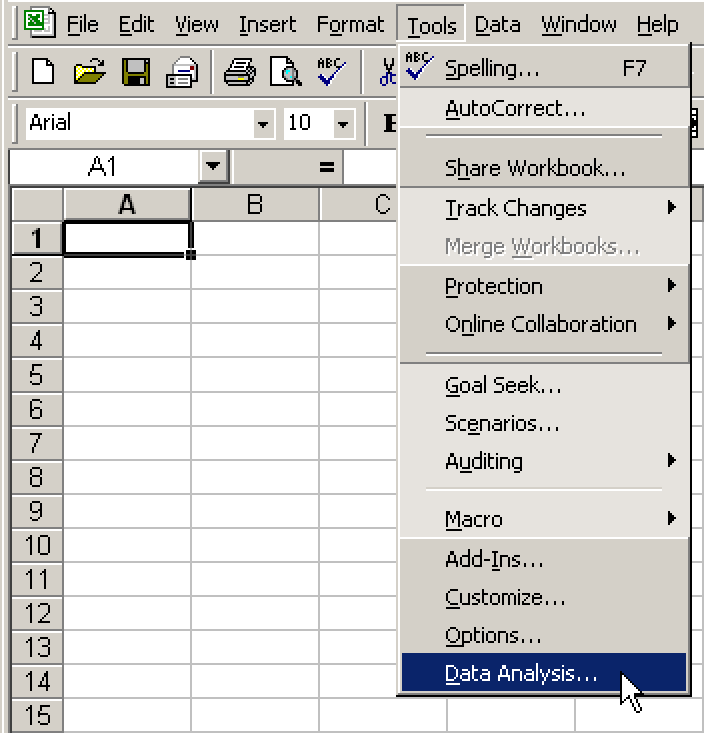
CHAPTER 1

Installing the Analysis ToolPak

Installing Analysis ToolPak

Microsoft Excel comes with many available statistical functions. We will depend heavily upon those contained in the Analysis ToolPak. This add-in contains many ready-to-use procedures. For each procedure, you only need to specify a few parameters and Excel does the rest. Although the Analysis ToolPak automates many tasks, Excel also contains an extensive list of functions that allow you to tailor analyses to suit your individual needs. Occasionally, we will use functions to generate analyses that seek to answer specific questions we will ask of our data.

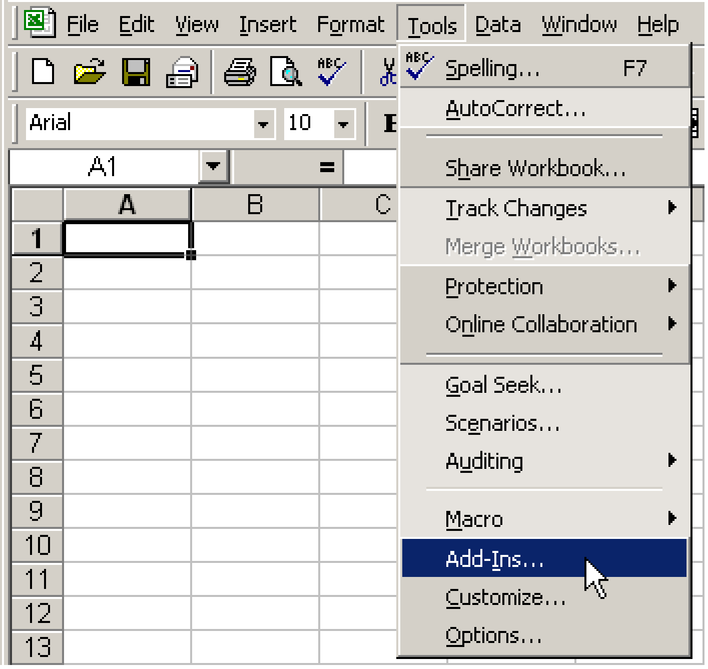
The Analysis ToolPak is an add-in that must be installed. If you are using Excel on a computer other than your own, someone else may have already installed the utility. To see if Analysis ToolPak is installed, click Tools in the main menu. If the drop-down list contains the Data Analysis option, then it is already installed.



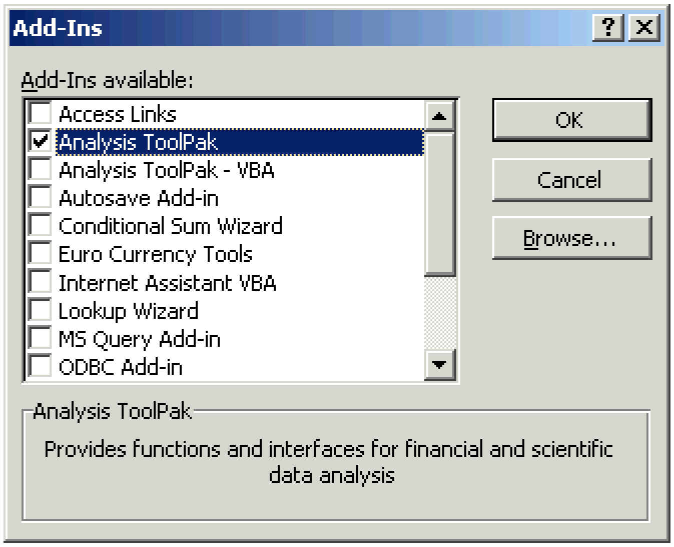
If not, you will need to follow the steps below.

Installing Analysis ToolPak

1. Click **Tools** on the Standard Toolbar and click **Add-Ins**.



2. In the Add-Ins dialog box, click the **check box** to the left of **Analysis ToolPak** and click **OK**.



Add-Ins available:

□ Access Links
☑ Analysis ToolPak
□ Analysis ToolPak - VBA
□ Autosave Add-in
□ Conditional Sum Wizard
□ Euro Currency Tools
□ Internet Assistant VBA
□ Lookup Wizard
□ MS Query Add-in
□ ODBC Add-in

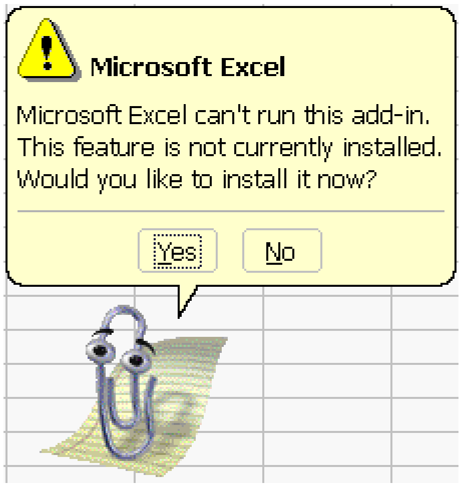
Analysis ToolPak

Provides functions and interfaces for financial and scientific data analysis

This will load the Analysis ToolPak into Excel. After installation, the dialog boxes will disappear, and Excel will be ready to use with the Analysis ToolPak add-in.

In some cases, you may get an error message like the one below. This occurs if the option for the add-in was not selected during the original installation of Microsoft Excel. Don't worry, because it can be installed from the original Office 2000 disk now, by following the next few steps.

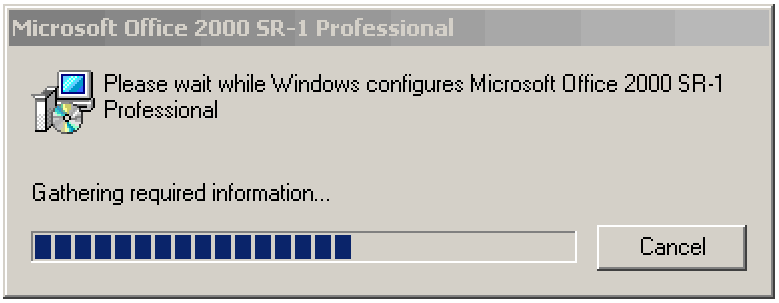
3. At the error notification, click **Yes** to install the add-in.



Microsoft Excel

Microsoft Excel can't run this add-in. This feature is not currently installed. Would you like to install it now?

4. A dialog box will appear with a message similar to the one below. Just allow the configuration to take place. Do not click Cancel.

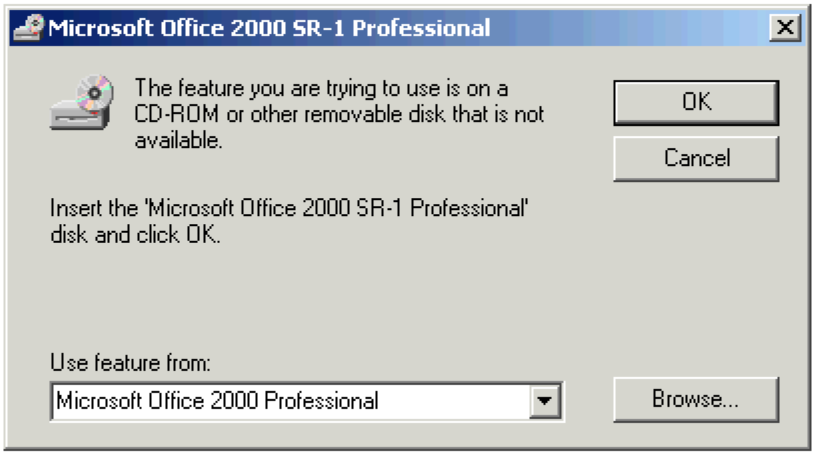


Microsoft Office 2000 SR-1 Professional

Please wait while Windows configures Microsoft Office 2000 SR-1 Professional

Gathering required information...

5. During the configuration process, if you have not already inserted your Microsoft Office 2000 disk into your CD ROM drive, you will get a message like that below. **Insert** your original **Microsoft Office 2000 CD** into the CD ROM drive and click **OK**.



Microsoft Office 2000 SR-1 Professional

The feature you are trying to use is on a CD-ROM or other removable disk that is not available.

Insert the 'Microsoft Office 2000 SR-1 Professional' disk and click OK.

Use feature from: Microsoft Office 2000 Professional

Hint: If the installation utility cannot find the needed files, you may need to click on Browse in the dialog box and select the drive letter of your CD ROM.

The configuration will continue until finished. You are now ready to use the features of Analysis ToolPak.

Notes

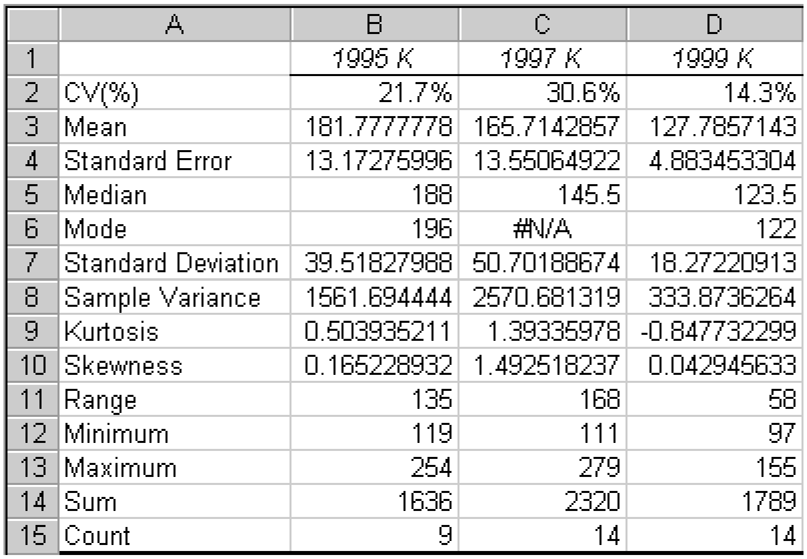
CHAPTER 2

Descriptive Statistics

Generating descriptive statistics

Meaning of descriptive statistics

In this chapter, we will learn how to create the following table of descriptive statistics.

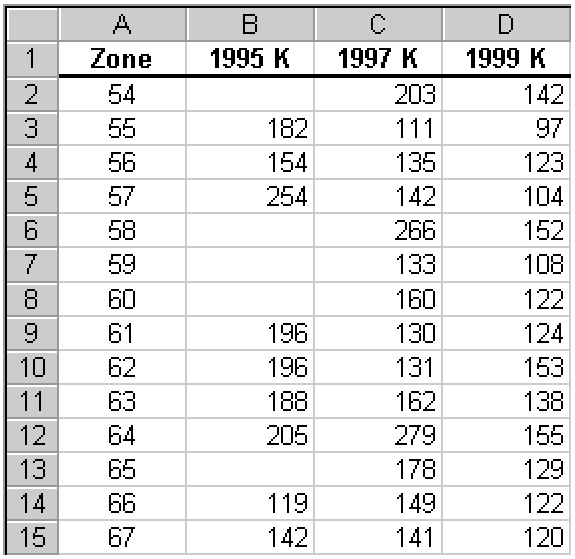


	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	21.7%	30.6%	14.3%
3	Mean	181.7777778	165.7142857	127.7857143
4	Standard Error	13.17275996	13.55064922	4.883453304
5	Median	188	145.5	123.5
6	Mode	196	#N/A	122
7	Standard Deviation	39.51827988	50.70188674	18.27220913
8	Sample Variance	1561.694444	2570.681319	333.8736264
9	Kurtosis	0.503935211	1.39335978	-0.847732299
10	Skewness	0.165228932	1.492518237	0.042945633
11	Range	135	168	58
12	Minimum	119	111	97
13	Maximum	254	279	155
14	Sum	1636	2320	1789
15	Count	9	14	14

For those interested, a discussion of the more commonly used statistics is also included in the second half of the chapter.

Generating descriptive statistics

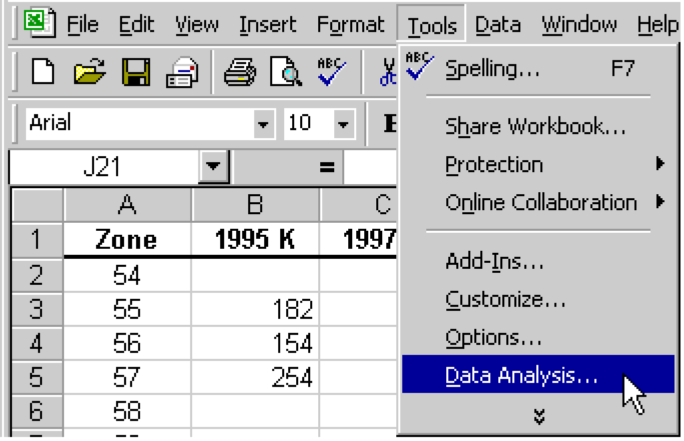
1. Start Microsoft Excel. On the Standard toolbar, Click File | Open and navigate to C:\PPIStat. Open file **Ex01**.



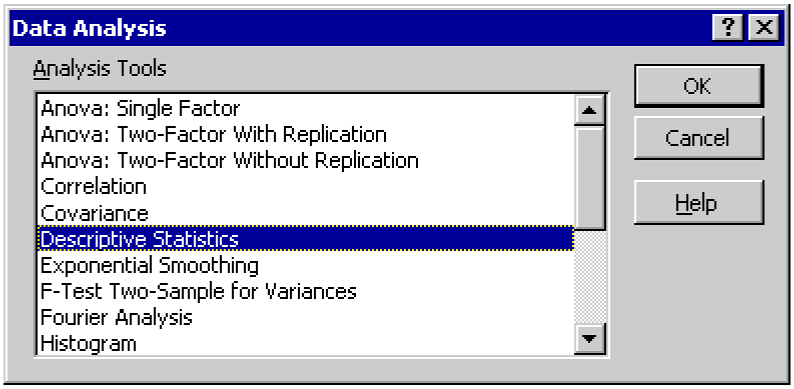
	A	B	C	D
1	**Zone**	**1995 K**	**1997 K**	**1999 K**
2	54		203	142
3	55	182	111	97
4	56	154	135	123
5	57	254	142	104
6	58		266	152
7	59		133	108
8	60		160	122
9	61	196	130	124
10	62	196	131	153
11	63	188	162	138
12	64	205	279	155
13	65		178	129
14	66	119	149	122
15	67	142	141	120

This file contains 3 years of soil test potassium (K) data from one field. The field has been subdivided into zones. Zone designations are in column A. The K tests corresponding to each zone are in column B, C, and D for years 1995, 1997, and 1999, respectively.

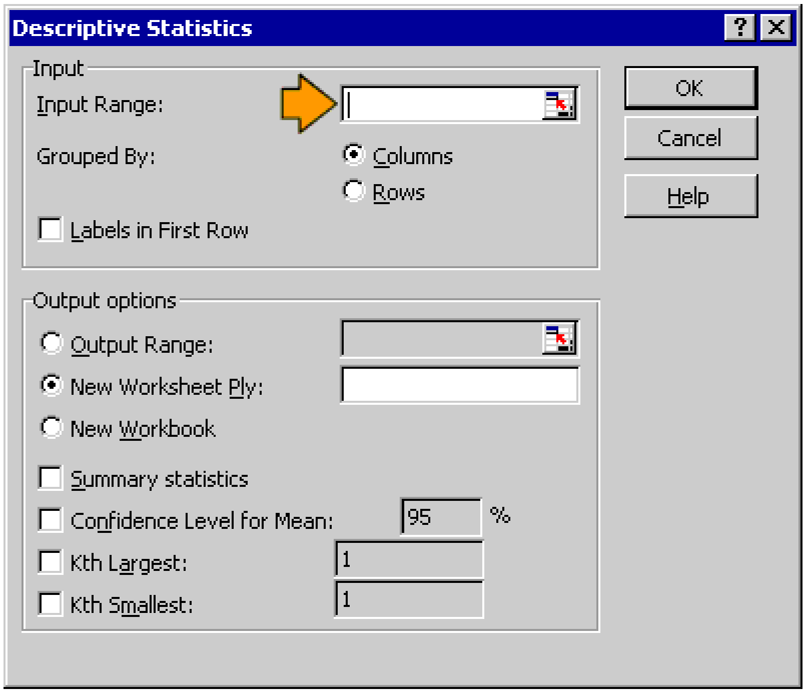
2. On the Standard toolbar, click **Tools | Data Analysis**. This brings up the Data Analysis dialog box.



3. In the Data Analysis dialog box, choose **Descriptive Statistics** in the Analysis Tools pane and click **OK**.



4. In the Descriptive Statistics dialog box, click on the blank text box to the right of the Input Range label.



Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Summary statistics

□ Confidence Level for Mean: 95 %

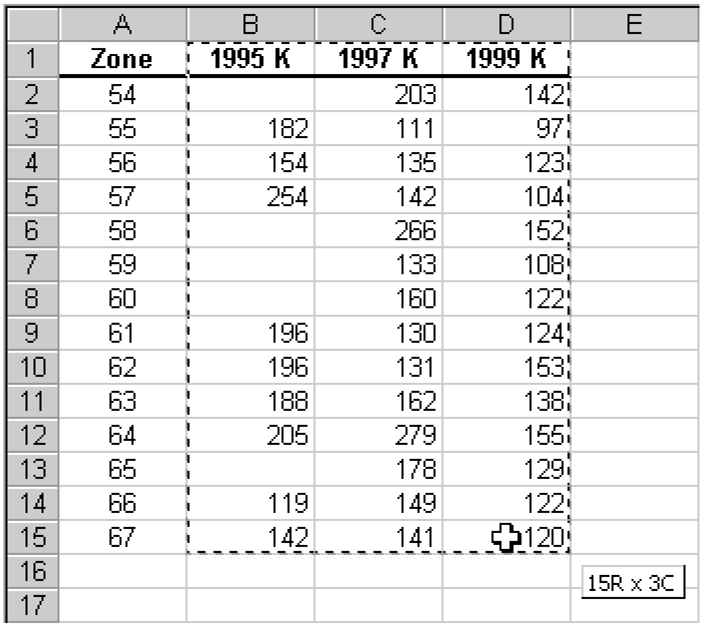
□ Kth Largest: 1

□ Kth Smallest: 1

Hint: A **range** of cells is denoted as first cell:last cell. For instance, the block of cells from B1 to D15 is denoted B1:D15. The first cell indicates the upper left hand corner of the cell block and the last cell indicates the lower right hand corner.

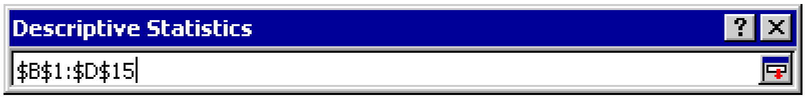
Hint: In the next step we will select cells. To **select** a range of cells, move the cursor to the upper left corner of the block of cells you want to select, click and hold down the left mouse button while moving the cursor to the bottom right corner of the block of cells. Release the left mouse button. The cells will be highlighted to indicate they are selected. The process of moving the cursor while depressing the left mouse button is termed **dragging**.

5. In the spreadsheet, select the soil test K data in cells **B1:D15**.



	A	B	C	D	E
1	Zone	1995 K	1997 K	1999 K	
2	54		203	142	
3	55	182	111	97	
4	56	154	135	123	
5	57	254	142	104	
6	58		266	152	
7	59		133	108	
8	60		160	122	
9	61	196	130	124	
10	62	196	131	153	
11	63	188	162	138	
12	64	205	279	155	
13	65		178	129	
14	66	119	149	122	
15	67	142	141	120	
16					15R x 3C
17					

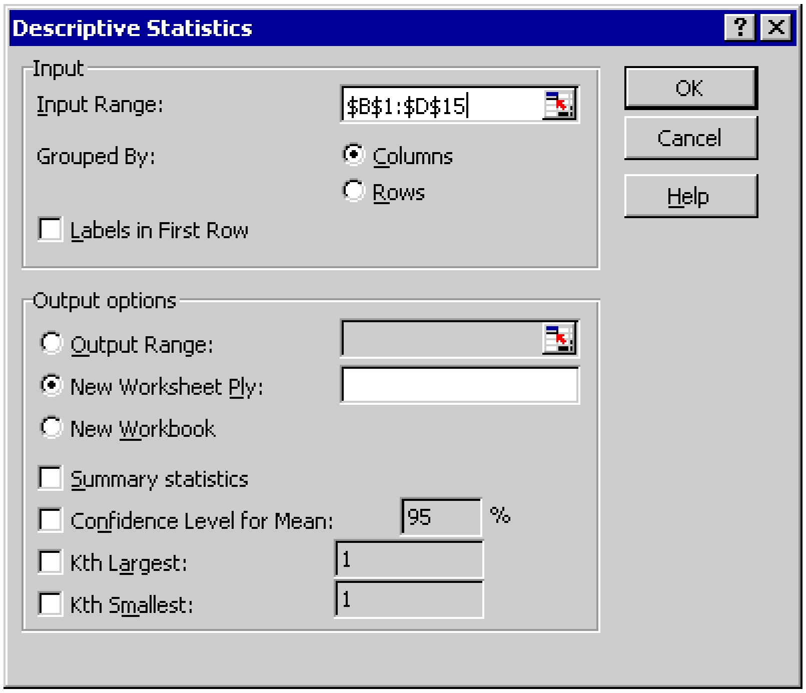
As you begin to select cells, you will notice that the Descriptive Statistics dialog box shrinks in size to display only the cell references of the highlighted cells.



Descriptive Statistics

$B$1:$D$15

After selecting the cells, the Descriptive Statistics dialog box returns to its original size and the cell references are displayed in the text box to the right of the Input Range label.



Descriptive Statistics

Input

Input Range: $B$1:$D$15

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

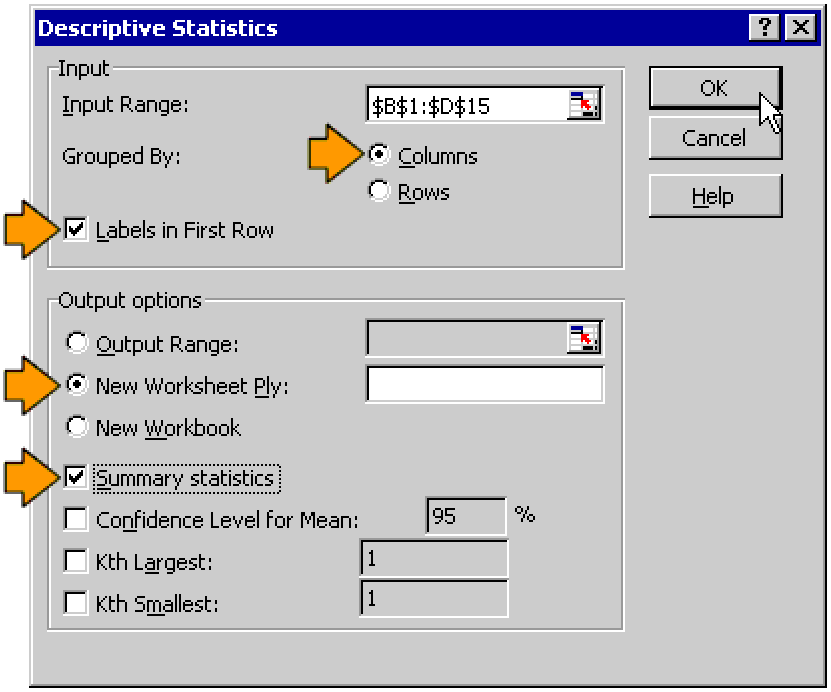
□ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

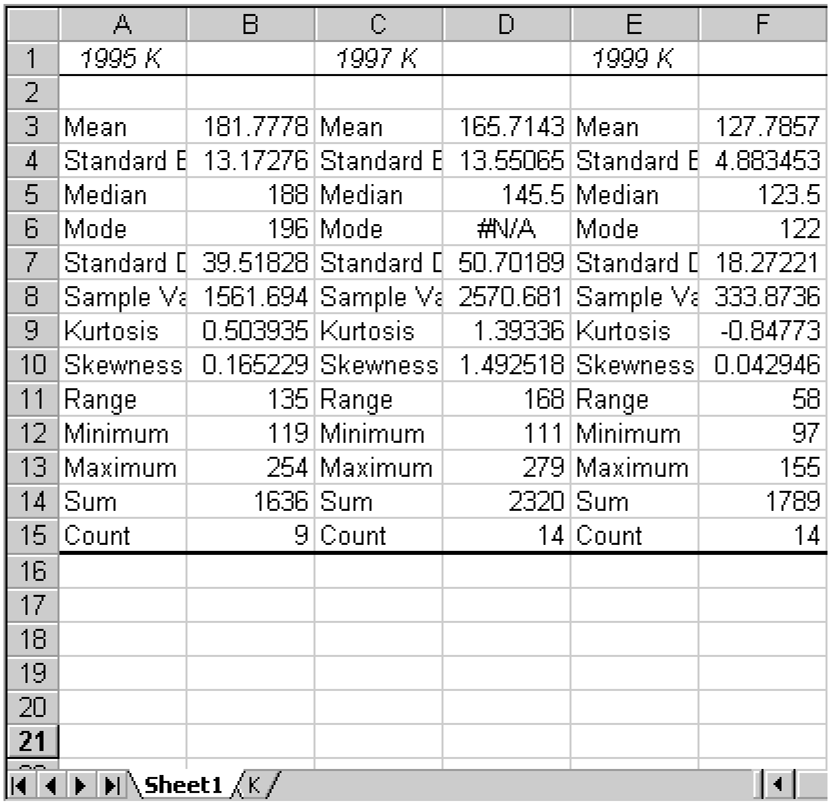
6. In the Input pane of the Descriptive Statistics dialog box, choose **Columns** for **Grouped By**, and check **Labels in First Row**. In the Output options pane, select **New Worksheet Ply** and check **Summary statistics**. Click **OK**.



Input
Input Range: $B$1:$D$15
Grouped By: ● Columns ○ Rows
☑ Labels in First Row

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook
☑ Summary statistics
☐ Confidence Level for Mean: 95 %
☐ Kth Largest: 1
☐ Kth Smallest: 1

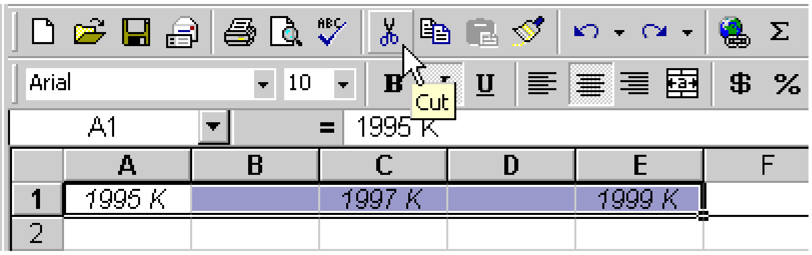
The Descriptive Statistics procedure generates a new sheet and navigates to it.



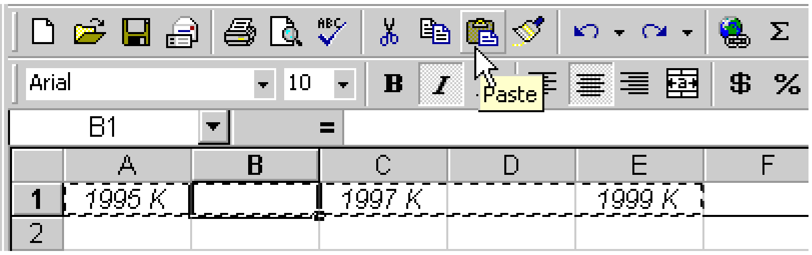
	A	B	C	D	E	F
1	*1995 K*		*1997 K*		*1999 K*	
2						
3	Mean	181.7778	Mean	165.7143	Mean	127.7857
4	Standard E	13.17276	Standard E	13.55065	Standard E	4.883453
5	Median	188	Median	145.5	Median	123.5
6	Mode	196	Mode	#N/A	Mode	122
7	Standard D	39.51828	Standard D	50.70189	Standard D	18.27221
8	Sample Va	1561.694	Sample Va	2570.681	Sample Va	333.8736
9	Kurtosis	0.503935	Kurtosis	1.39336	Kurtosis	-0.84773
10	Skewness	0.165229	Skewness	1.492518	Skewness	0.042946
11	Range	135	Range	168	Range	58
12	Minimum	119	Minimum	111	Minimum	97
13	Maximum	254	Maximum	279	Maximum	155
14	Sum	1636	Sum	2320	Sum	1789
15	Count	9	Count	14	Count	14
16						
17						
18						
19						
20						
21						

We need to move the labels of the columns so that the years are over the numbers, rather than the names of the statistics.

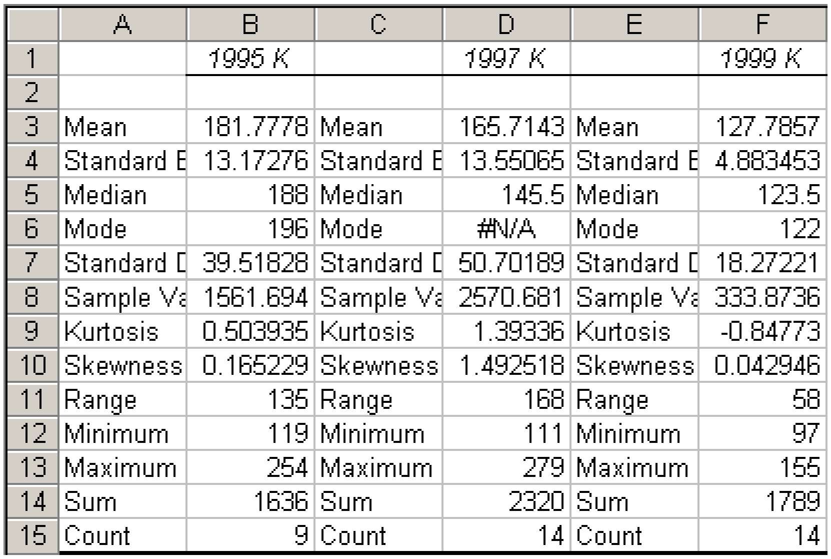
7. Select cells **A1:E1** and press **Cut** on the Standard Toolbar.



8. Select cell **B1** and click **Paste** on the Standard Toolbar.



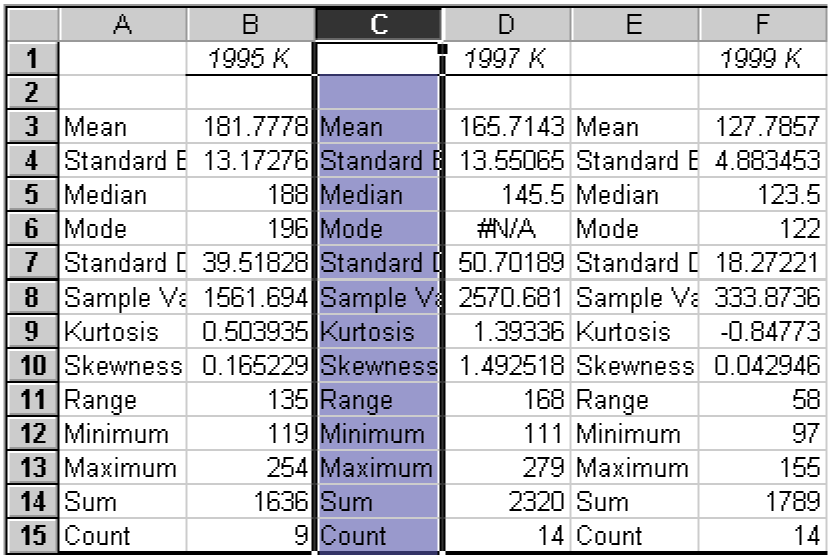
Now the labels appear over the values of the statistics.



	A	B	C	D	E	F
1		1995 K		1997 K		1999 K
2						
3	Mean	181.7778	Mean	165.7143	Mean	127.7857
4	Standard E	13.17276	Standard E	13.55065	Standard E	4.883453
5	Median	188	Median	145.5	Median	123.5
6	Mode	196	Mode	#N/A	Mode	122
7	Standard D	39.51828	Standard D	50.70189	Standard D	18.27221
8	Sample Va	1561.694	Sample Va	2570.681	Sample Va	333.8736
9	Kurtosis	0.503935	Kurtosis	1.39336	Kurtosis	-0.84773
10	Skewness	0.165229	Skewness	1.492518	Skewness	0.042946
11	Range	135	Range	168	Range	58
12	Minimum	119	Minimum	111	Minimum	97
13	Maximum	254	Maximum	279	Maximum	155
14	Sum	1636	Sum	2320	Sum	1789
15	Count	9	Count	14	Count	14

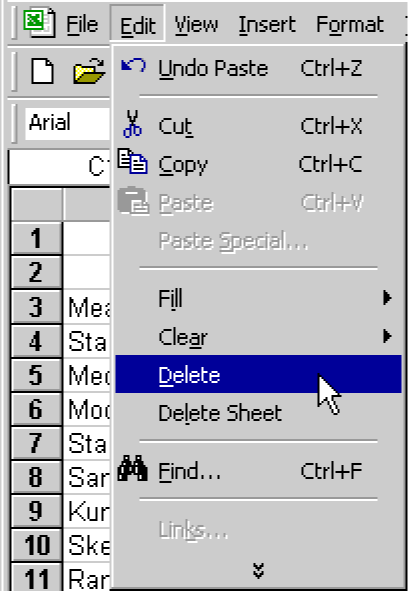
For our purposes, we need only the first column of statistics names.

9. **Click** on column label **C** to select the entire column.

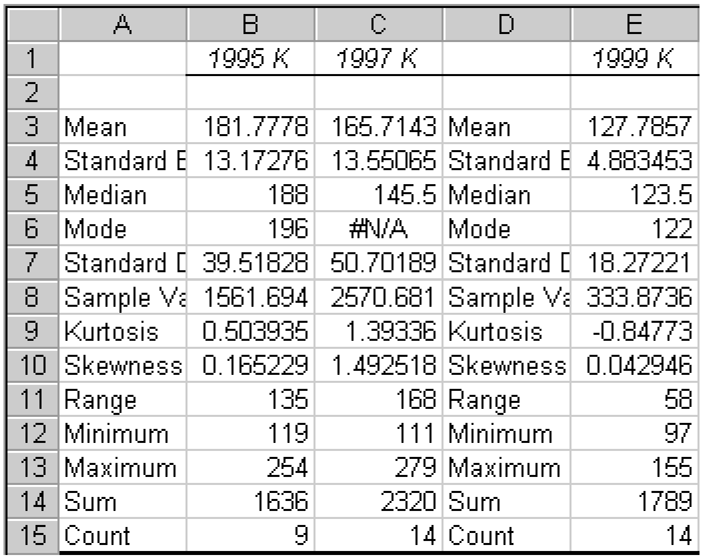


	A	B	C	D	E	F
1		*1995 K*		*1997 K*		*1999 K*
2						
3	Mean	181.7778	Mean	165.7143	Mean	127.7857
4	Standard E	13.17276	Standard E	13.55065	Standard E	4.883453
5	Median	188	Median	145.5	Median	123.5
6	Mode	196	Mode	#N/A	Mode	122
7	Standard D	39.51828	Standard D	50.70189	Standard D	18.27221
8	Sample Va	1561.694	Sample Va	2570.681	Sample Va	333.8736
9	Kurtosis	0.503935	Kurtosis	1.39336	Kurtosis	-0.84773
10	Skewness	0.165229	Skewness	1.492518	Skewness	0.042946
11	Range	135	Range	168	Range	58
12	Minimum	119	Minimum	111	Minimum	97
13	Maximum	254	Maximum	279	Maximum	155
14	Sum	1636	Sum	2320	Sum	1789
15	Count	9	Count	14	Count	14

10. On the Standard Toolbar, click **Edit | Delete**.

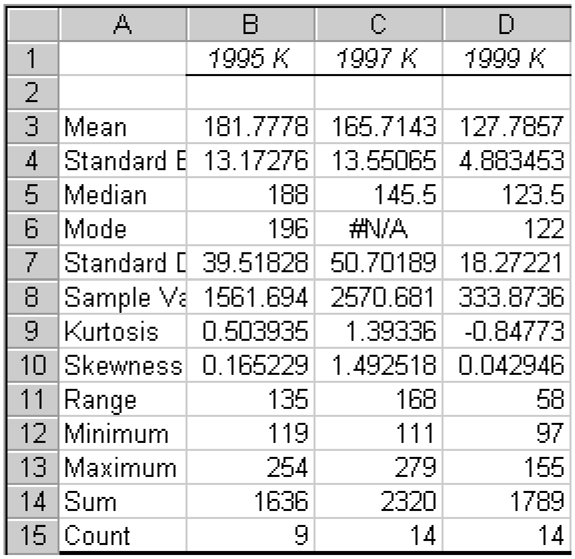


The result is the deletion of the statistics names that used to be in column C. Data in columns D, E, and F have now been shifted to the left and occupy columns C, D, and E. We now want to get rid of the remaining statistics names in column D.



	A	B	C	D	E
1		*1995 K*	*1997 K*		*1999 K*
2					
3	Mean	181.7778	165.7143	Mean	127.7857
4	Standard E	13.17276	13.55065	Standard E	4.883453
5	Median	188	145.5	Median	123.5
6	Mode	196	#N/A	Mode	122
7	Standard D	39.51828	50.70189	Standard D	18.27221
8	Sample Va	1561.694	2570.681	Sample Va	333.8736
9	Kurtosis	0.503935	1.39336	Kurtosis	-0.84773
10	Skewness	0.165229	1.492518	Skewness	0.042946
11	Range	135	168	Range	58
12	Minimum	119	111	Minimum	97
13	Maximum	254	279	Maximum	155
14	Sum	1636	2320	Sum	1789
15	Count	9	14	Count	14

11. **Select column D** and click **Edit | Delete** on the Standard Toolbar.

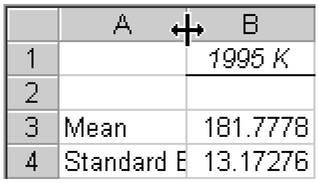


	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2				
3	Mean	181.7778	165.7143	127.7857
4	Standard E	13.17276	13.55065	4.883453
5	Median	188	145.5	123.5
6	Mode	196	#N/A	122
7	Standard D	39.51828	50.70189	18.27221
8	Sample Va	1561.694	2570.681	333.8736
9	Kurtosis	0.503935	1.39336	-0.84773
10	Skewness	0.165229	1.492518	0.042946
11	Range	135	168	58
12	Minimum	119	111	97
13	Maximum	254	279	155
14	Sum	1636	2320	1789
15	Count	9	14	14

We now have one column of statistics names and 3 columns of statistical values.

Currently, we cannot see all of the contents in some of the cells. The width of the columns is too narrow. We will let Excel resize the column widths automatically to ensure that all of the cell contents are visible.

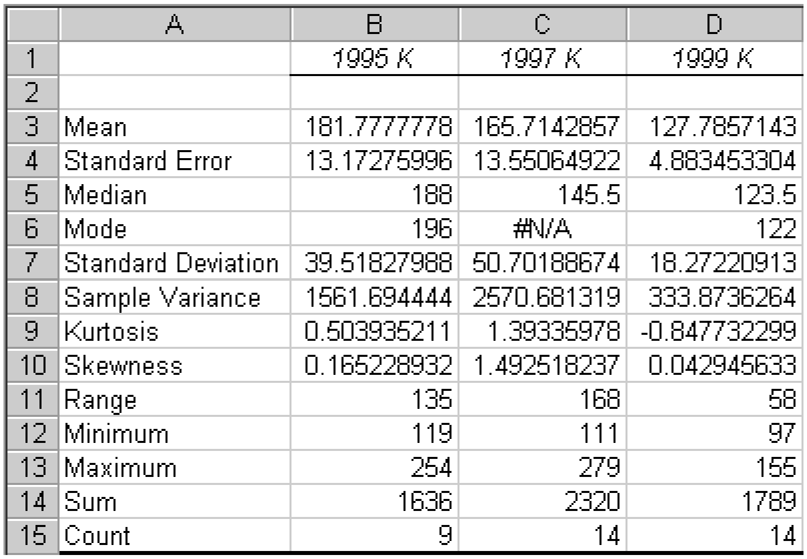
12. Move the cursor between column labels **A** and **B**. The cursor will change to a double headed arrow. **Double click** to resize the width of column A.



	A	B
1		*1995 K*
2		
3	Mean	181.7778
4	Standard E	13.17276

Hint: To change the width of a column, move the cursor to the right border of the column label. Double clicking automatically changes the width of the column so that the contents of all of the cells in the column are displayed. You can also change the column width manually by clicking and dragging the cursor toward the right or left to narrow or widen the column, respectively.

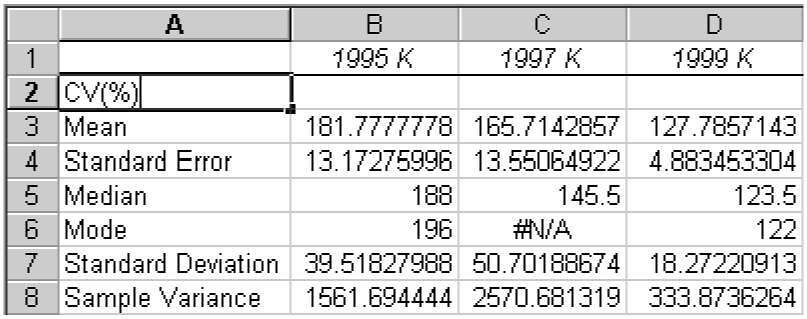
13. **Resize** the widths of the remaining columns (**B**, **C**, and **D**) to display all cell contents.



	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2				
3	Mean	181.7777778	165.7142857	127.7857143
4	Standard Error	13.17275996	13.55064922	4.883453304
5	Median	188	145.5	123.5
6	Mode	196	#N/A	122
7	Standard Deviation	39.51827988	50.70188674	18.27220913
8	Sample Variance	1561.694444	2570.681319	333.8736264
9	Kurtosis	0.503935211	1.39335978	-0.847732299
10	Skewness	0.165228932	1.492518237	0.042945633
11	Range	135	168	58
12	Minimum	119	111	97
13	Maximum	254	279	155
14	Sum	1636	2320	1789
15	Count	9	14	14

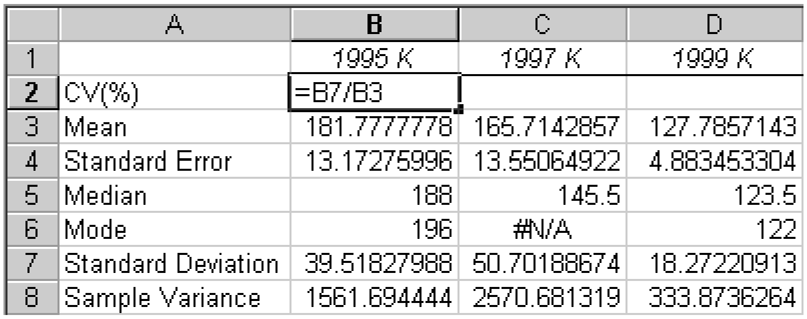
There is one important statistic that has not been calculated for us. It is the coefficient of variation (CV). The CV is the standard deviation divided by the mean, expressed as a percentage.

14. Type **CV(%)** in cell **A2**.



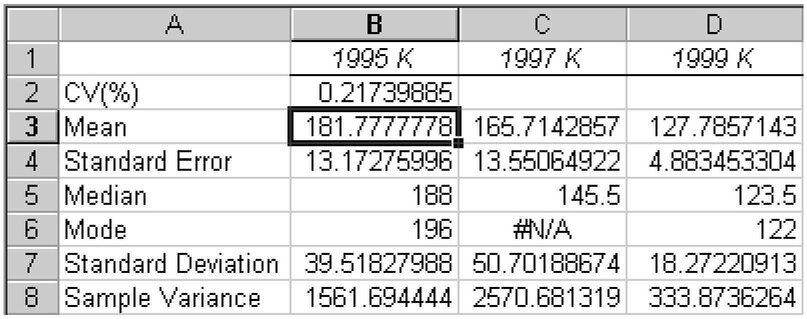
	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)			
3	Mean	181.7777778	165.7142857	127.7857143
4	Standard Error	13.17275996	13.55064922	4.883453304
5	Median	188	145.5	123.5
6	Mode	196	#N/A	122
7	Standard Deviation	39.51827988	50.70188674	18.27220913
8	Sample Variance	1561.694444	2570.681319	333.8736264

15. In cell **B2**, type **=B7/B3** and press **Enter**.



	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	=B7/B3		
3	Mean	181.7777778	165.7142857	127.7857143
4	Standard Error	13.17275996	13.55064922	4.883453304
5	Median	188	145.5	123.5
6	Mode	196	#N/A	122
7	Standard Deviation	39.51827988	50.70188674	18.27220913
8	Sample Variance	1561.694444	2570.681319	333.8736264

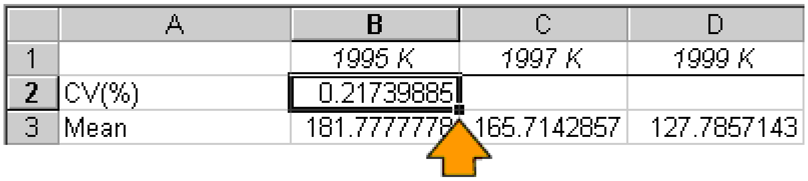
The = sign tells Excel that a formula follows. The formula we entered divides (denoted by the / symbol) the standard deviation (cell B7) by the mean (cell B3). The result is a decimal fraction in cell B2.



	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	0.21739885		
3	Mean	181.7777778	165.7142857	127.7857143
4	Standard Error	13.17275996	13.55064922	4.883453304
5	Median	188	145.5	123.5
6	Mode	196	#N/A	122
7	Standard Deviation	39.51827988	50.70188674	18.27220913
8	Sample Variance	1561.694444	2570.681319	333.8736264

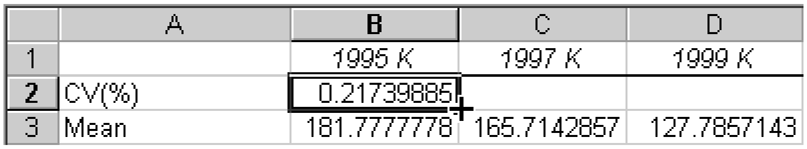
We want to repeat the calculation we just performed for the 1997 and 1999 K data. To do this, we will use the **fill handle**, located on the bottom right hand corner of the cell highlight, as shown below.

16. Select cell **B2**.



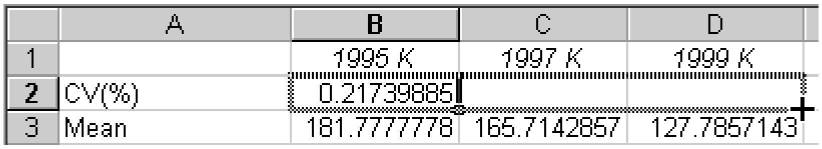
	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	0.21739885		
3	Mean	181.7777778	165.7142857	127.7857143

17. Move the cursor to the fill handle. The cursor changes to a cross.



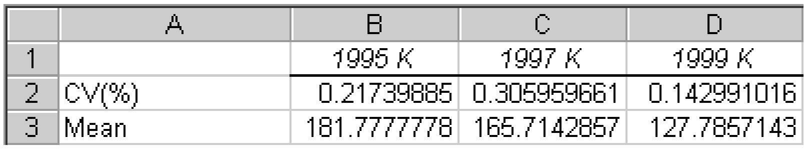
	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	0.21739885		
3	Mean	181.7777778	165.7142857	127.7857143

18. Click and drag the cursor from cell **B2** to the right side of cell **D2**.



	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	0.21739885		
3	Mean	181.7777778	165.7142857	127.7857143

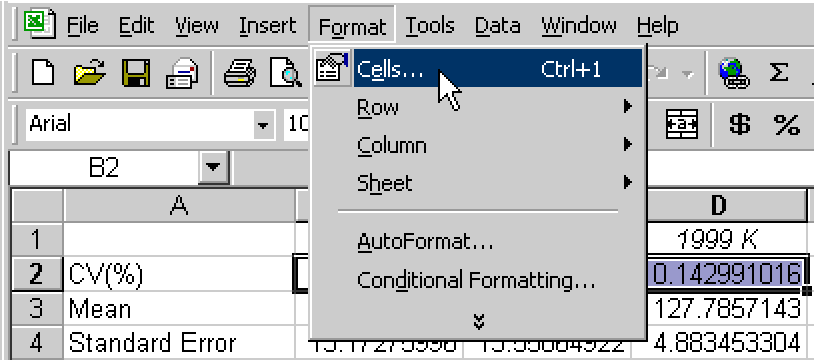
We now have CVs calculated for each of the 3 years of data.



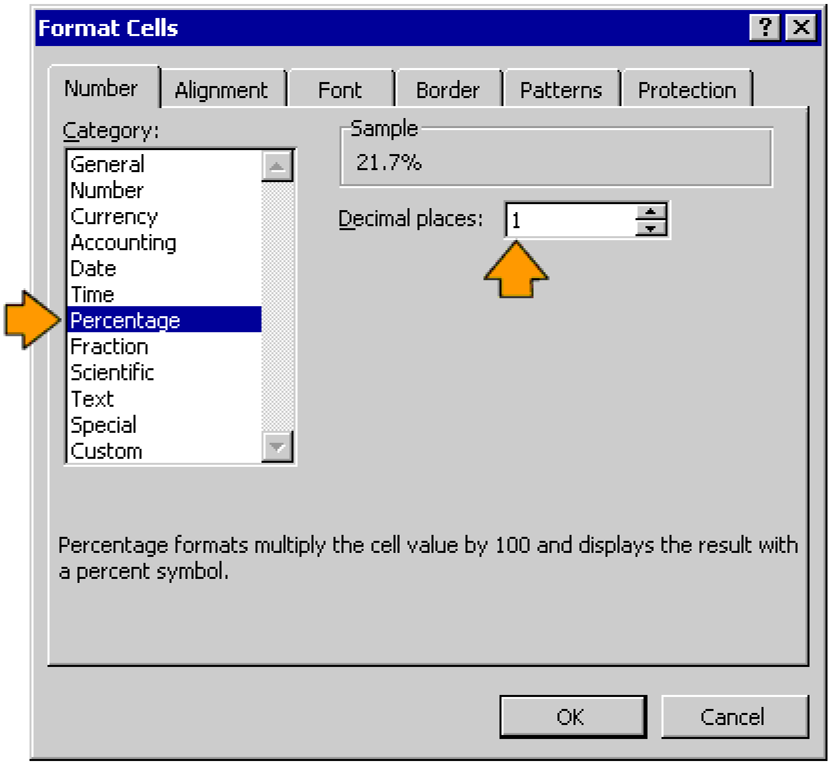
	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	0.21739885	0.305959661	0.142991016
3	Mean	181.7777778	165.7142857	127.7857143

Our next step is to express these decimal fractions as percentages. We will use cell formatting to do this.

19. Select cells **B2:D2**. On the Standard toolbar, click **Format | Cells**.



20. In the Format Cells dialog box, select **Percentage** in the **Category** pane. Select **1** for **Decimal places**.



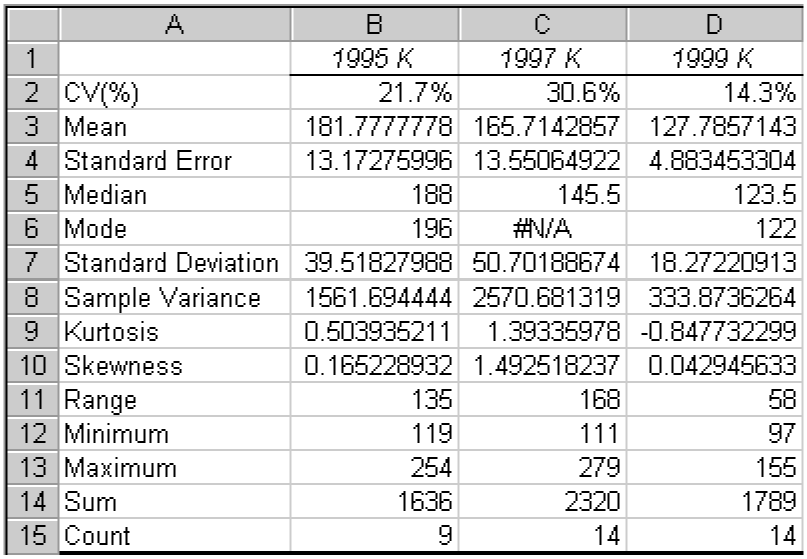
Category: Percentage

Sample
21.7%

Decimal places: 1

Percentage formats multiply the cell value by 100 and displays the result with a percent symbol.

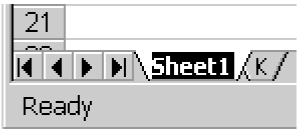
We now have a table complete with descriptive statistics for all 3 years of soil test data.



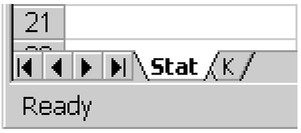
	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	21.7%	30.6%	14.3%
3	Mean	181.7777778	165.7142857	127.7857143
4	Standard Error	13.17275996	13.55064922	4.883453304
5	Median	188	145.5	123.5
6	Mode	196	#N/A	122
7	Standard Deviation	39.51827988	50.70188674	18.27220913
8	Sample Variance	1561.694444	2570.681319	333.8736264
9	Kurtosis	0.503935211	1.39335978	-0.847732299
10	Skewness	0.165228932	1.492518237	0.042945633
11	Range	135	168	58
12	Minimum	119	111	97
13	Maximum	254	279	155
14	Sum	1636	2320	1789
15	Count	9	14	14

Next, we change the sheet label to be more descriptive of the statistical contents.

21. Double click the tab label to select it.



22. Type **Stat** to remember that this sheet contains the descriptive statistics, then press **Enter**.



This completes the task of creating descriptive statistics. For those interested, the next section explains the meaning of the most commonly used descriptive statistics.

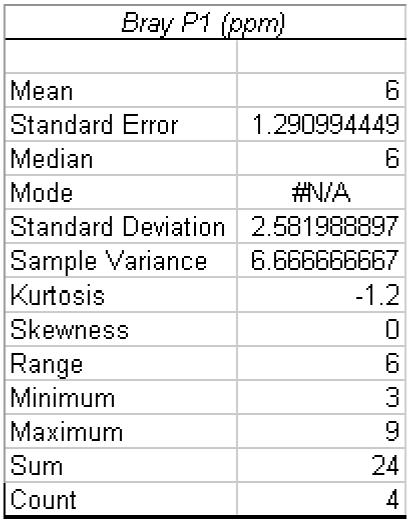
Meaning of descriptive statistics

This section explains the statistics generated by the Descriptive Statistics procedure in Analysis ToolPak. In addition to these statistics, we also discuss the CV, which is the statistic we added to the output in the previous section. If you are already familiar with these statistics, skip ahead to the next chapter. Each statistic will be defined and its calculation discussed. Some mathematical notation is used, but explanations of the notation are given to keep the equations understandable.

In this section, we will use a small data set comprised of 4 Bray P1 soil test values:

Sample No.	Soil test level
	(ppm)
1	3
2	5
3	7
4	9

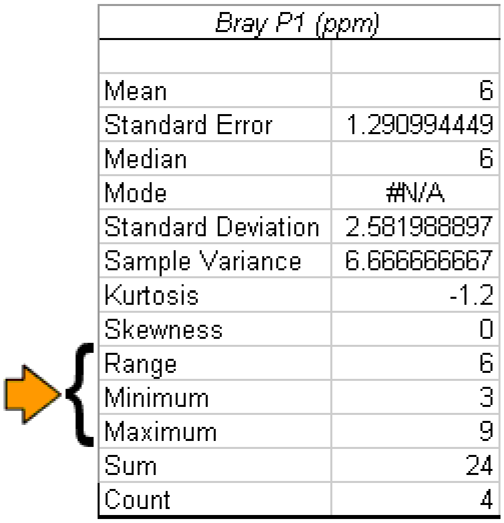
Running the Descriptive Statistics procedure in Analysis ToolPak on this small data set produces the following output:



Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

Let's see how each of these statistics were calculated.

Minimum, maximum, and range



Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

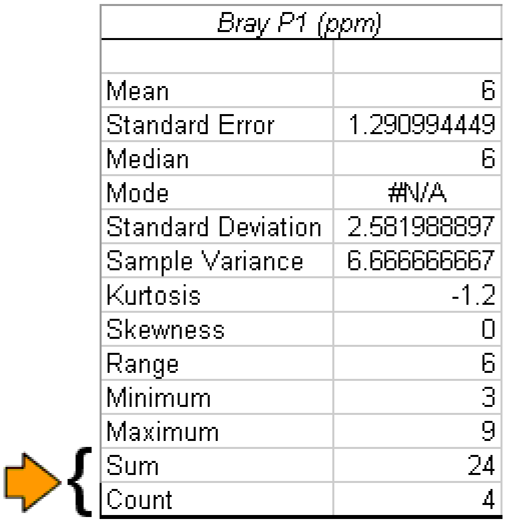
The **minimum** is simply the lowest value in a set of data. The **maximum** is the highest value. The range is the difference between the maximum and minimum:

[2-1] range = maximum – minimum.

In our example data set, the minimum is 3, the maximum is 9, and the range is:

$$9 - 3 = 6 \text{ ppm.}$$

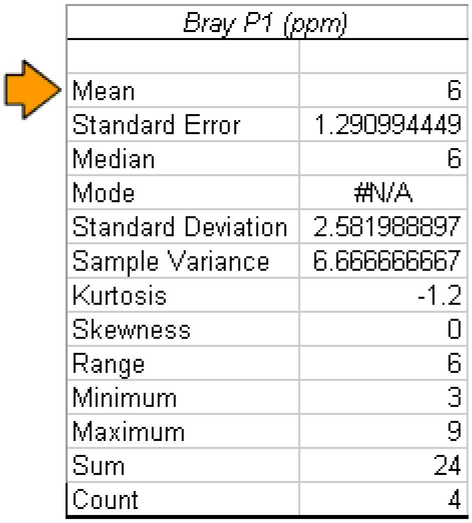
Sum and count



Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

The **sum** is the result of adding together all of the values in the data set. The **count** is the number of observations. For our data set, the sum is 3 + 5 + 7 + 9 = 24. The count is 4, since there are 4 numbers in the set.

Mean



Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

The **mean**, better know as the average, is the sum of all values divided by the total number of observations, or using the terminology above, the *sum* divided by the *count*. Using our Bray P1 data set, the mean is:

$$\text{Mean} = \frac{3+5+7+9}{4} = 6 \text{ ppm}$$

Often, a more generalized formula is presented for calculating the mean. We will build our way up to it. First, let the variable n represent the number of observations, or *count*. In our example, we are averaging 4 numbers, so $n = 4$. This changes the above formula to:

$$\text{Mean} = \frac{3+5+7+9}{n} = 6 \text{ ppm},$$

where $n = 4$.

Next, we can use special notation to denote each number we want to include in our average. The notation used throughout the statistical literature for an individual observation is y_{i}. Let's see how this works. **Table 2.1** shows the observations and their corresponding notation.

Table 2.1. Notation for each observation.

Observation number(i):	$i = 1$	$i = 2$	$i = 3$	$i = 4$
Bray P1value (ppm):	3	5	7	9
Observation (y_{i}):	y_1	y_2	y_3	y_4

We can now re-write our equation with this new notation. We build upon our previous equation:

$$\text{Mean} = \frac{y_1 + y_2 + y_3 + y_4}{n},$$

where $n = 4$.

In statistical notation, the mean is usually denoted by $\bar{y}$, called “y bar”. We continue to build our equation which now becomes:

[2-2]
$$\bar{\text{y}} = \frac{y_1 + y_2 + y_3 + y_4}{n}$$

Because we have 4 numbers in our calculation for the mean ($n = 4$), the variable i must necessarily start with 1 and go up to 4 (Table 8), or written another way, i will range from 1 to n. So if we were to write an explanation for the numerator of the above equation, $y_1 + y_2 + y_3 + y_4$, it would be:

“Add together, or sum, all observations (y_i), starting with the first one (y_1) and ending with the last one (y_n).”

Because this statement is rather lengthy, statisticians use a shorthand notation. The notation used is the Greek letter Sigma, Σ, which is a Greek “S” and stands for <u>s</u>um in mathematics. Using this **summation notation** for the above statement, we have:

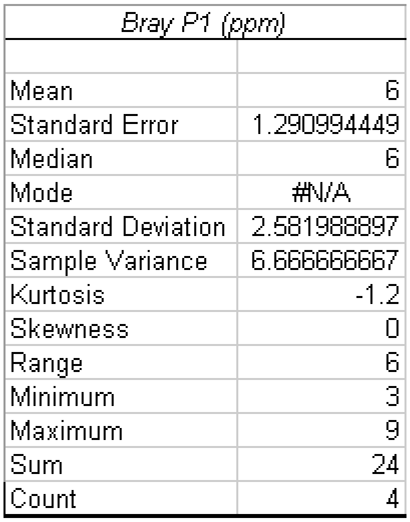
$$\sum_{i=1}^{n} y_i$$

In this notation, $i = 1$ under the Sigma tells us to begin with observation y_1. The Sigma tells us to sum the observations. The n on top of the Sigma symbol tells us when to stop summing (when we reach the last observation y_n, or in our case y_4). We can now plug this notation into equation [2-2] above to replace the more cumbersome $y_1 + y_2 + y_3 + y_4$:

[2-3]
$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

This is the equation for the mean found in all statistical texts and is a general equation that works regardless of how many numbers (n) go into the calculation of the mean.

Median



Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

The **median** of a set of data is the middle value when the measurements are arranged from lowest to highest. How we calculate this value depends upon whether our number of observations (*n*) is even or odd.

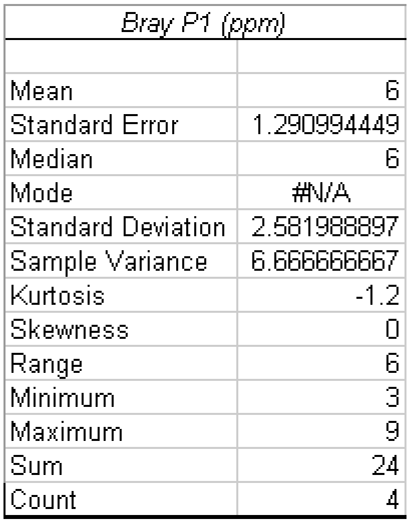
Odd number of observations

This is the simplest case. Let’s consider the following set of 5 Bray P1 values arranged from lowest to highest: 2, 4, 7, 8, and 10 ppm. The middle value, or median, is 7 ppm. Two numbers are below it and 2 numbers are above it.

Even number of observations

This is the case for our original small Bray P1 data set. The soil test levels are, arranged from smallest to largest: 3, 5, 7, and 9 ppm. Because there is an even number of observations, there is no middle value. We need a value between 5 and 7 ppm. In such cases, the median is the mean of the two values on either side of the middle, or (5+7)/2 = 6 ppm.

Mode



Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

The **mode** is the most frequently occurring value. The problem with this statistic is that it may not always be defined, or it may not have a unique solution. Let's take a look at some examples to demonstrate these limitations.

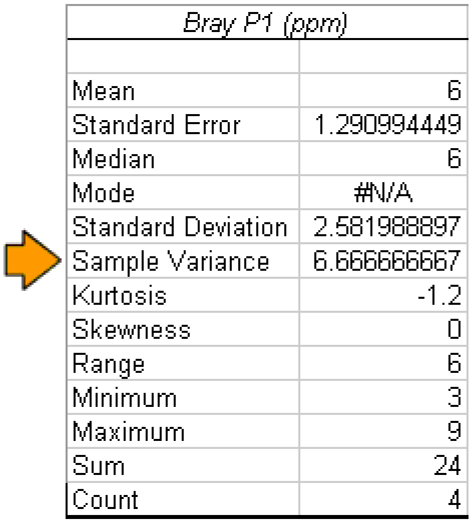
Let's first consider the following data set: 3, 5, 5, and 11 ppm Bray P1. The most frequently occurring value in this set is 5 ppm, which occurs twice. Our mode is therefore 5 ppm.

Next consider the data set 3, 5, 5, 7, 7, and 9 ppm. Notice that both the numbers 5 and 7 ppm occur more than once and both occur with the same greatest frequency. Therefore, there are two possible modes: 5 and 7 ppm.

Now consider the Bray P1 data set we have been using: 3, 5, 7, and 9 ppm Bray P1. No number appears more than once, so there is no most frequently occurring value. Therefore, no mode is defined for this data set. Excel displays **#N/A** for the mode when this occurs.

So why use the mode? Where it is defined, the mode does at least identify the largest cluster of repeated values in the data set.

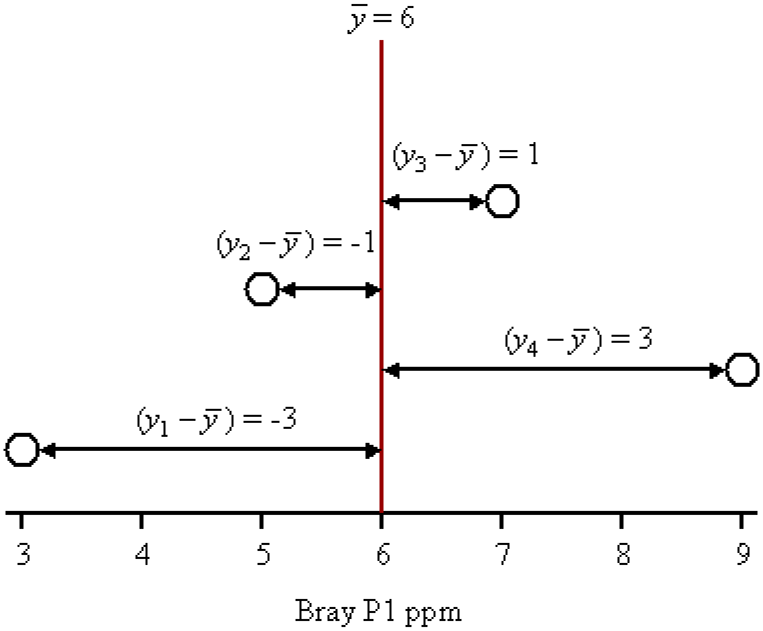
Sample variance



Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

Variance is a measure of the variability in our data set. Our discussion of variance begins with the range. For a given mean soil test, a smaller range means that the values in the data set are more closely clustered. For instance, 200 observations with a range of 5 ppm Bray P1 are much more tightly grouped than the same observations with a range of 81 ppm. The problem with using the range as our sole measure of variability is that it is merely the difference between the minimum and maximum value. It doesn't consider the variability between these endpoints. What we need is some measurement that accounts for the variability of all the points in a set of data.

One way to measure the variability of each point is to calculate it's distance from some reference point. Points farther away from the reference are considered more variable than those closer to it. A logical choice for the needed reference is a point that represents the central tendency of the distribution. Traditionally, statisticians have chosen the mean, which for our Bray P1 data set is 6 ppm and is shown as a vertical line in the figure below. The difference between an observation (shown as a circle) and the mean is termed **deviation**. In the figure, deviations are represented by double headed arrows. We see that observations that contribute to greater variability in our data set have greater deviations.



To calculate deviations, we will use the y_i notation for an observation and $\bar{y}$ as the notation for the mean. We assign values in **Table 2.2**.

Table 2.2. Notation for each observation in the Bray P1 data set.

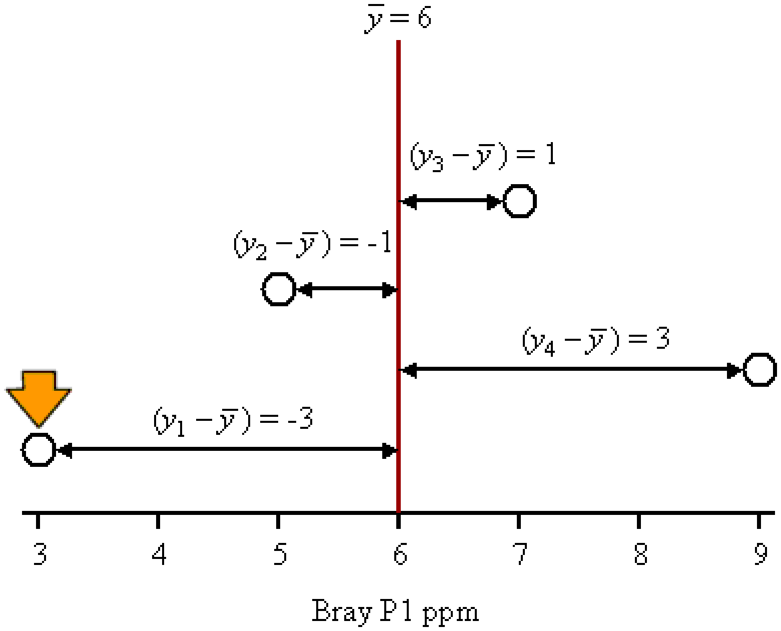
Observation number(i):	$i = 1$	$i = 2$	$i = 3$	$i = 4$
Value (ppm Bray P1):	3	5	7	9
Observation (y_i):	y_1	y_2	y_3	y_4

Deviation is defined as:

[2-4] $$(y_i - \bar{y})$$

where i = 1 to n.

The deviation of our first value (y_1 = 3 ppm) from the mean (6 ppm) is 3 – 6 = –3 ppm.



Deviation is limited to a single observation. However, our objective is to come up with one number that quantifies the variability of the entire data set. This number needs to consider the deviations of all the observations. We may initially think that we could simply add up all of the deviations of the individual points. Using summation notation, we would have the following formula:

$$\sum_{i=1}^{n} \left(y_i - \bar{y}\right)$$

However, when we try this, we find (using the figure above) that –3 –1 + 1 + 3 = 0. Zero is not a very accurate characterization of our dispersion since we know we have variability. This result is not unique to our data set. In fact, simply summing together all of the deviations in a data set will always result in a value of 0. This occurs because the sum of the deviations below the mean (all of which are negative) always equals the sum of the deviations above the mean (all of which are positive). This leads to the fundamental equality:

[2-5] $$\sum_{i=1}^{n} \left(y_i - \bar{y}\right) = 0$$

So if simply adding together the deviations doesn't provide a good characterization of variability, what does? There are many possible approaches, but what statisticians have agreed upon as a standard method of quantifying variation consists of three steps:

1. First, square all of the individual deviations.
2. Next, add together all of the squared deviations.
3. Divide the sum by the number of "free deviations" which is one less than the number of observations, or ($n - 1$). This number is termed the **degrees of freedom** and is abbreviated **df**.

Let's build the equation step by step.

In step 1, each deviation is squared, or multiplied by itself. This converts all of the negative deviations to positive ones, since a negative number times itself equals a positive number. A general equation for this step is based on equation [2-4]:

$$\left(y_i - \overline{y}\right)^2$$

In step 2, all of these squared deviations are added together for the *n* observations in the data set. Using summation notation, we have:

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

In step 3, we divide the above equation by the degrees of freedom. This is a term used in statistics to describe whether or not a value is "free" to be anything, or whether it "must" be a certain number. Let's restrict our discussion to deviations. Suppose that we have 4 observations in our data set. We therefore have 4 deviations from the mean. They can be anything. Our only stipulation is that equation [2-5] must hold true. The first deviation is "free" to be anything, say 4. The second and third deviations are also "free" to be anything say 8 and –5, respectively. But this is where the "freedom" ends. The final standard deviation "must" make the sum of all of the deviations equal 0. The fourth deviation therefore must be a specific value. It has to satisfy the following equation:

$$4 + 8 - 5 + d_4 = 0$$

where d_4 is our fourth deviation.

We move all of the numbers on the left side of the equation over to the right side, leaving only d_4 on the left. Remember that when you move a number to the other side of the "=" sign, you have to change its sign (+ goes to – and vice versa):

$$d_4 = -4 - 8 + 5$$

Therefore, $d_4 = -7$. This is the only value that the fourth deviation can be.

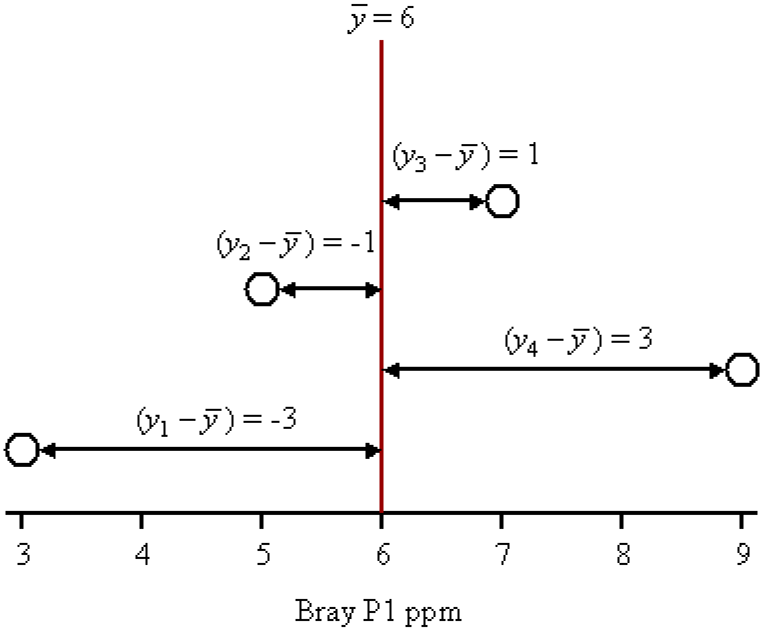
So for our 4 deviations, only 3 were "free" to be any value. If we did this for other numbers of deviations, we would find that the number of "free deviations" or degrees of freedom is always one less than the number of observations, or (n - 1).

We are now ready to write our final equation, which is called the **variance** and is abbreviated $\boldsymbol{s^2}$:

[2-6]
$$s^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{(n-1)}$$

As the deviations become greater, the variance also becomes greater. Therefore, larger variances mean greater dispersion within the data set.

To see how this equation works, again consider the deviations shown in the figure below. Remember that n = 4:

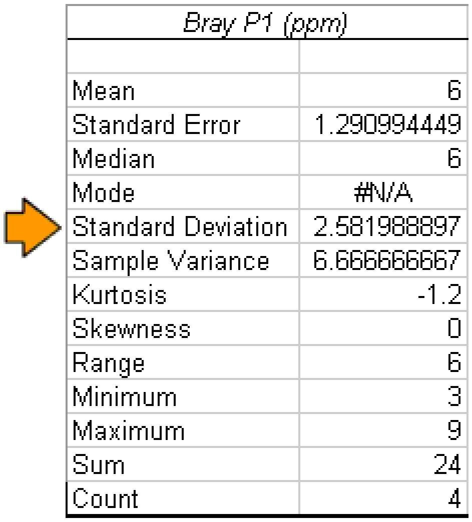


$$s^2 = \frac{(-3)^2 + (-1)^2 + (1)^2 + (3)^2}{4-1} = \frac{9+1+1+9}{3} = \frac{20}{3} = 6.67 \text{ ppm}^2$$

Notice that squaring each deviation also squares the units, so the unit of our variance is, in this case, ppm^2.

Notice that the variance meets our initial goal: one number that quantifies the variability of all of the observations in our data set.

Standard deviation



Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

Although the variance is a useful statistic for quantifying variability, its squared units are a bit cumbersome. It is difficult for us to understand, in the case of the Bray P1 example above, what 6.67 ppm^2 looks like. We would like our measure of variability to be in the same units as our original data. This is made possible by another measure of variability called the **standard deviation**, abbreviated ***s***. It is simply the positive square root of the variance.

Let's consider the Bray P1 soil test example above. The standard deviation is:

$$s = \sqrt{s^2} = \sqrt{6.67\ \text{ppm}^2} = 2.58\ \text{ppm}$$

Now we have a measure of variability that is in our original units and is more understandable.

A common use of the standard deviation is to estimate what percent of the total observations in a data set fall in certain ranges centered around the mean. The standard deviation, added to and subtracted from the mean, defines the maximum and minimum of this range. For instance, let's consider the range defined by the mean ($\overline{y}$ = 6 ppm) and standard deviation (s = 2.58 ppm) or our small data set:

minimum = $\overline{y} - s = 6 - 2.58 = 3.42$

maximum = $\overline{y} + s = 6 + 2.58 = 9.42$

Now we see what percent of our total number of samples falls in this range. In our data set 3, 5, 7, 9 ppm, only two of the four observations, 5 and 7, fall between 3.42 and 9.42. This represents 50% of our total observations.

Instead of writing $\overline{y} - s$ and $\overline{y} + s$ every time, which is cumbersome, we can combine these expressions into $\overline{y} \pm s$ to define the maximum and minimum of this range.

We can also examine wider ranges. Instead of looking only at one standard deviation on either side of the mean, we could look at twice or three times the standard deviation. A range defined by two times the standard deviation would be written $\bar{y} \pm 2s$ and would be calculated, using our example, as:

minimum = $\bar{y} - 2s = 6 - 2(2.58) = 0.84$

maximum = $\bar{y} + s = 6 + 2(2.58) = 11.16$

In our data set 3, 5, 7, 9 ppm, all four (100%) or the observations fall within this range.

So what use is this? Most times, we will use the standard deviation combined with the mean to predict, rather than back calculate, what percent of the population falls in such ranges. We don't want to go back to the data set each time to do this, and sometimes we won't have ready access to it. Making an estimate often comes close enough, and involves a lot less effort. The following general estimates can be made, assuming that the way soil test levels are distributed resembles a bell curve (dicussed in more detail later):

1. The interval $\bar{y} \pm s$ contains approximately 68% of the observations
2. The interval $\bar{y} \pm 2s$ contains approximately 95% of the observations
3. The interval $\bar{y} \pm 3s$ contains nearly all of the observations

This rule of thumb is called the **empirical rule**, and it is used often to gain a sense of the variability in a data set. We know from our previous calculations that the actual percentage of observations falling into these intervals can be different from that predicted. For instance, we found 50% of the observations fell in the interval $\bar{y} \pm s$ where the general estimate from the empirical rule predicted 68%. Actual and predicted percentages get closer with larger data sets distributed more nearly like a bell shape.

Because the standard deviation is so important for understanding variability, it should be included in reports whenever possible. The interval commonly used in scientific literature is one standard deviation about the mean, or $\bar{y} \pm s$. For instance, if you are recording the average soil test level of a field, instead of just writing down 6 ppm (for our example), write 6 ± 2.58 ppm, to remind you about the variability of the data set.

CV

The **CV**, or **coefficient of variation**, is the standard deviation (s) divided by the mean ($\bar{y}$) and expressed as a percentage:

[2-7]

$$CV = \left(\frac{s}{\bar{y}}\right)\times 100$$

The CV therefore expresses variability in a way that is scaled to the magnitude of the mean. It is a useful statistic for comparing the variability of two or more data sets that have different means. For instance, let's say that we want to compare two sets of soil test data, both of which have a standard deviation of 5 ppm Bray P1. The first data set has a mean of 10 ppm, while the second has a mean of 100 ppm. The CV of the first data set is:

$$(5 \text{ ppm} / 10 \text{ ppm}) \times 100 = 50\%.$$

The CV of the second data set is:

$$(5 \text{ ppm} / 100 \text{ ppm}) \times 100 = 5\%.$$

Comparing these two data sets, we can say that, relative to the mean, the first data set averaging 10 ppm Bray P1 is more variable (CV = 50%) than the second set averaging 100 ppm (CV = 5%).

Standard Error

The data set we have been using, 3, 5, 7, 9 ppm Bray P1, represents 4 samples taken from a field. The mean of this data set is 6 ppm. Based on these four samples, we might assume that 6 ppm represents the actual average Bray P1 level for the entire field.

If we could put these soil samples back in the field and randomly take 4 new samples, we might get 10, 15, 20, and 25 ppm. The mean of these four samples is 17.5 ppm. This is another estimate of the actual average Bray P1 level for the entire field.

We could go out a third time, gather another four samples, and get 8, 10, 12, and 14 ppm. The mean of these four samples is 11 ppm. This is a third estimate of the actual average Bray P1 level for the entire field.

We could keep doing this and doing this, thousands of times if money were no object. In each case we have two means we are calculating. We have the mean of the four samples we take each time, which is called the **sample mean**. In our scenario, we calculated three sample means: 6, 17.5, and 11 ppm. Each one is an estimate of the actual average Bray P1 soil test for the entire field. The actual average, or **population mean**, could only be obtained by measuring the Bray P1 level of every cubic inch of soil in the field. Since we cannot do this, we estimated the population mean with our three sample means.

We can consider the three sample means as a separate data set. Just like the data sets of the samples, the new data set comprised of just the sample means has its own mean and standard deviation. The mean of the sample means is:

$$\frac{6+17.5+11}{3} = 11.5 \text{ ppm.}$$

This is our estimate for the population mean, or the actual mean Bray P1 soil test level that exists in the field.

The standard deviation of our data set of sample means is a bit different than the one we calculated earlier for the samples themselves. It is defined as the standard deviation of the population (σ, the lower case Greek letter sigma) divided by the square root of the number of samples going into each sample mean (n) which is 4 in our case, since each sample mean was comprised of 4 samples. The standard deviation of the sample means is called the standard error, and is expressed as:

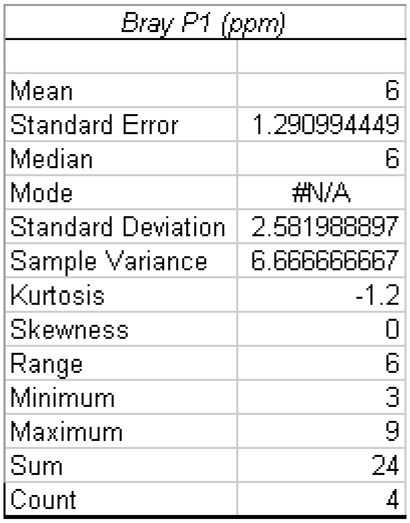
$$\text{Standard Error} = \sigma / \sqrt{n} \qquad [2\text{-}8]$$

The standard deviation of the population, σ, is often unknown. Our best estimate in our single data set we have been using is the standard deviation of the 4 samples in our data set (s = 2.58). Since our sample mean (6 ppm) was made up of four observations (3, 5, 7, 9 ppm), then n = 4. An estimate of the standard error, based on the standard deviation of our sample, is then:

$$\text{Estimated standard error: } \frac{2.58}{\sqrt{4}} = \frac{2.58}{2} = 1.29 \text{ ppm.}$$

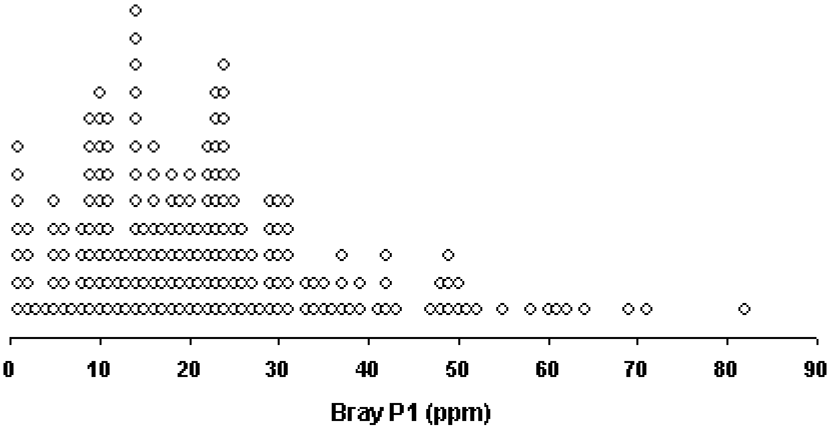
This is our estimate for the standard deviation of a data set comprised of sample means.

Skewness

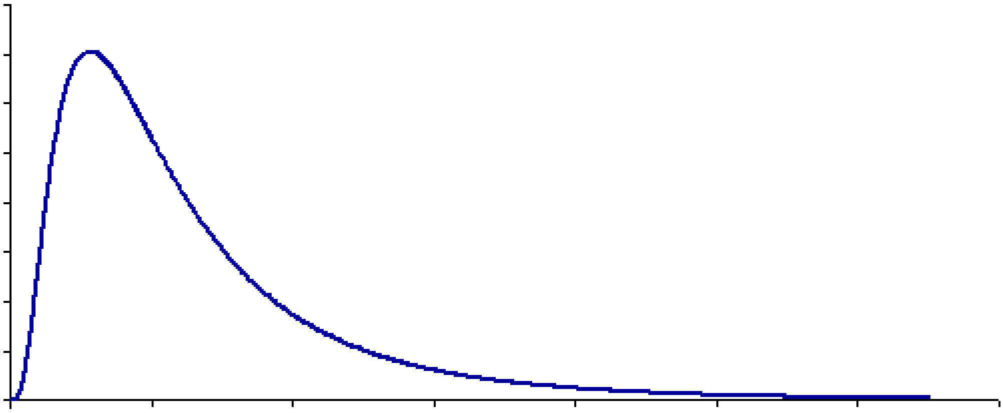


Bray P1 (ppm)	
Mean	6
Standard Error	1.290994449
Median	6
Mode	#N/A
Standard Deviation	2.581988897
Sample Variance	6.666666667
Kurtosis	-1.2
Skewness	0
Range	6
Minimum	3
Maximum	9
Sum	24
Count	4

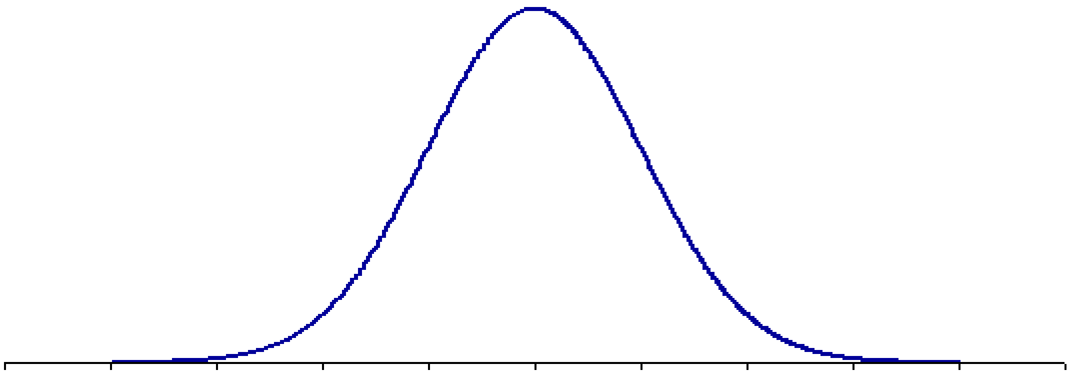
Skewness describes the shape of a distribution. Let's consider a large data set from a research field. Researchers collected 198 samples from one field and determined the Bray P1 level of each one. When we plot all of the points, we come up with the graph below. Points stacked on top of one another represent values that are repeated. For instance on the left side, the first stack of points represents 7 of the 198 samples that tested 2 ppm Bray P1.



If we were to draw a smooth curve over the top of these points, it would look something like the graph below.



Notice the long tail to the right, toward higher values. This distribution is said to be **positively skewed**. If the long tail were on the left side, toward smaller values, the distribution would be **negatively skewed**. A distribution like the one below, has **zero skewness**. It is bell shaped and the tail on one side is not longer than the tail on the other. This type of distribution is called a **normal distribution**.



Skewness can be quantified by the coefficient of skewness. This coefficient is calculated somewhat the same as the variance. Remember that the variance used the square of the deviations. The coefficient of skewness uses the cube of the deviations in the following equation:

$$\text{Coefficient of skewness} = \frac{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^3}{ns^3} \qquad [2\text{-}9]$$

Where y_i is an observation, $\overline{y}$ is the mean, n is the number of observations, and s is the standard deviation. Using the small Bray P1 data set we have been using, the coefficient of skewness is:

$$s^2 = \frac{(-3)^3 + (-1)^3 + (1)^3 + (3)^3}{4(2.58)^3} = \frac{-27 - 1 + 1 + 27}{68.7} = \frac{0}{68.7} = 0$$

This is the value reported in the output from our Descriptive Statistics procedure in Analysis ToolPak. We have to be careful interpreting the coefficient of skewness for our small data set. It is more meaningful with higher numbers of observations. What our coefficient of skewness indicates is that the data we have appears to be normally distributed with no extreme values that create a tail to any one side of the distribution.

Notes

Notes

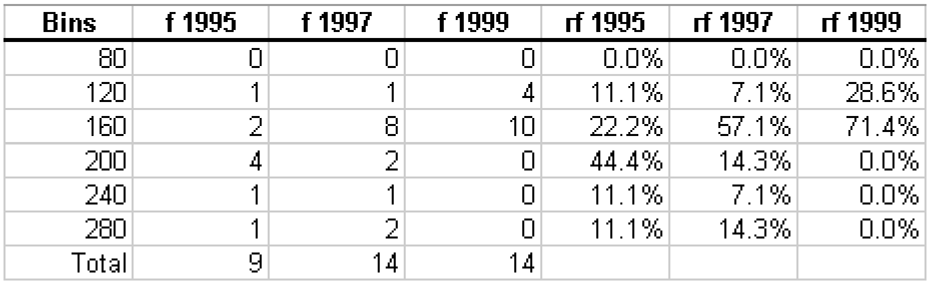
CHAPTER 3

Distributions

Generating distributions

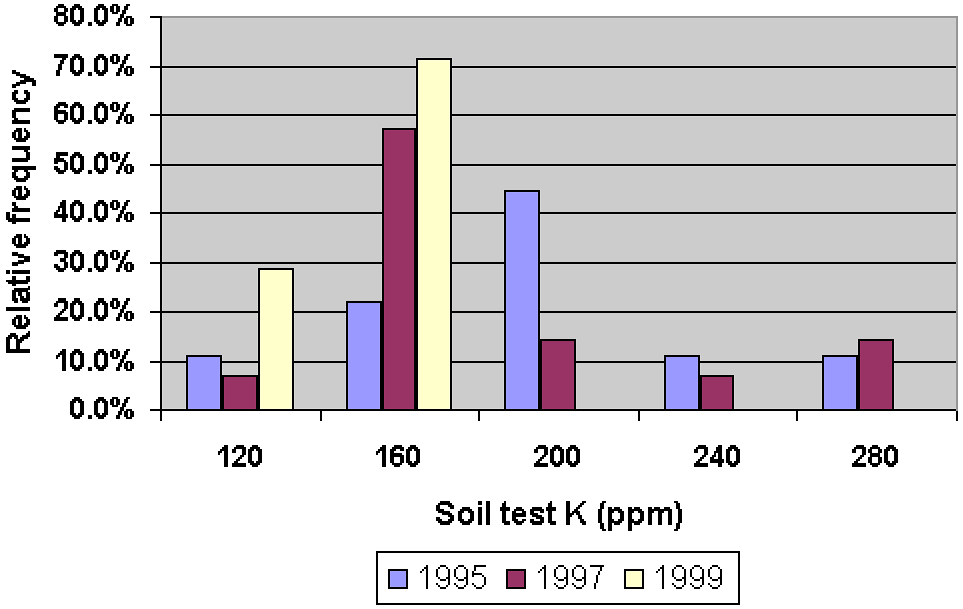
Creating histograms

In the previous chapter, we generated some descriptive statistics like the variance, standard deviation, and coefficient of variation, that quantified the variability in our data sets. In addition to quantifying variability, we also want to visualize it. To do this, we will generate two types of distributions.



Bins	f 1995	f 1997	f 1999	rf 1995	rf 1997	rf 1999
80	0	0	0	0.0%	0.0%	0.0%
120	1	1	4	11.1%	7.1%	28.6%
160	2	8	10	22.2%	57.1%	71.4%
200	4	2	0	44.4%	14.3%	0.0%
240	1	1	0	11.1%	7.1%	0.0%
280	1	2	0	11.1%	14.3%	0.0%
Total	9	14	14			

We will then create a graph to visualize how the distribution in soil test K levels has changed over time.



Generating distributions

Creating a frequency distribution

When we distribute something, we divide it into portions. In statistics, the place where we put a portion is called a **class**. We are interested in dividing all of the observations in our data set into distinct classes. The result is a distribution. To create a distribution, we must do three things:

1. Create a set of distinct classes that are equal in size
2. Create enough classes to encompass both the minimum and maximum values
3. Count how many observations fall into each class

The classes we create should have agronomic meaning, if possible. For our soil test example, we are interested in organizing soil tests according to the classes in the potassiuum recommendation table:

Table 4.1. Potassium recommendations for corn

Category:	Very low	Low	Medium	High	Very High
Soil test range:	0-40	41-80	81-120	121-160	161+
Amount of potassium to apply (lb K_2O/acre)	185	135	80	25	0

We notice that the soil test ranges in the table not only have agronomic meaning but are also evenly spaced, each being 40 ppm wide. These classes therefore are a good choice for creating a distribution. The next step involves creating enough classes to encompass both the minimum and maximum soil test levels of each data set we want to analyze.

Because we want to see how the distribution in soil test levels has changed over time, we need to be sure to construct our classes carefully. We want one set of classes that are appropriate for all three data sets we will analyze.

From the previous chapter, we found the following minimum and maximum values for each of the three years of soil test K data:

Table 4.2. Minimum and maximum values in each soil test K data set (from Chapter 2)

	Year		
	1995	1997	1999
Minimum (ppm):	119	111	97
Maximum (ppm):	254	279	155

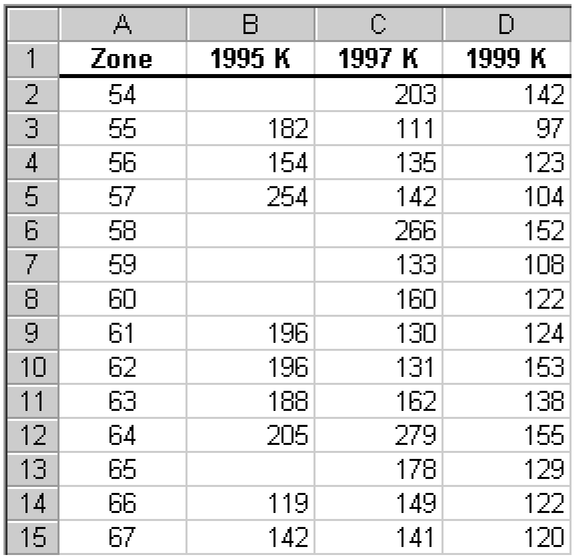
From this table we see that the lowest minimum of the three years is 97 ppm. The first class therefore needs to be 81-120 ppm. So what will be our highest class? To find this, we create a table of equally spaced classes, each 40 ppm wide to see what the last class must be to encompass the 279 ppm maximum value in our data set. From the table below, we find that the last required category is 241-280 ppm. A total of 5 categories are needed.

Table 4.3. Categories for soil test K data.

Category		
Min.		Max.
---- (ppm) ----		
81	-	120
121	-	160
161	-	200
201	-	240
241	-	280

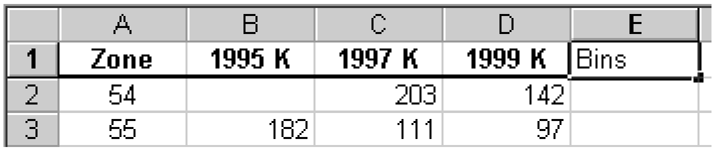
In Excel, our first step in creating a distribution is to define the limits of our classes. First of all, Excel uses the term **Bins** for classes. Second, Excel only wants the upper limits of the classes, or bins, specified. This means that to create the classes in Table 4.3, we only need to enter the data in the column labeled Max.

1. Start Microsoft Excel. On the Standard toolbar, Click File | Open and navigate to C:\PPIStat. Open file **Ex02**.



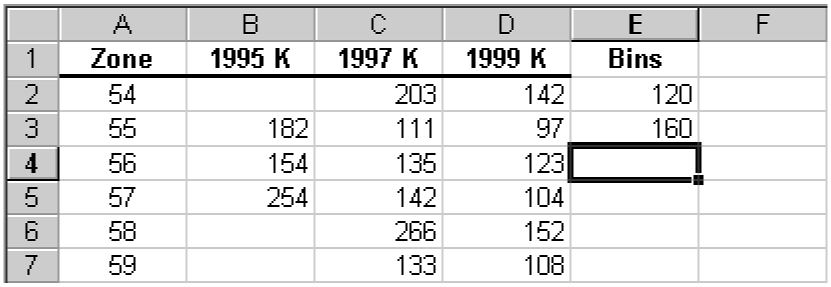
	A	B	C	D
1	**Zone**	**1995 K**	**1997 K**	**1999 K**
2	54		203	142
3	55	182	111	97
4	56	154	135	123
5	57	254	142	104
6	58		266	152
7	59		133	108
8	60		160	122
9	61	196	130	124
10	62	196	131	153
11	63	188	162	138
12	64	205	279	155
13	65		178	129
14	66	119	149	122
15	67	142	141	120

2. In cell E1, type **Bins** then press **Enter**.



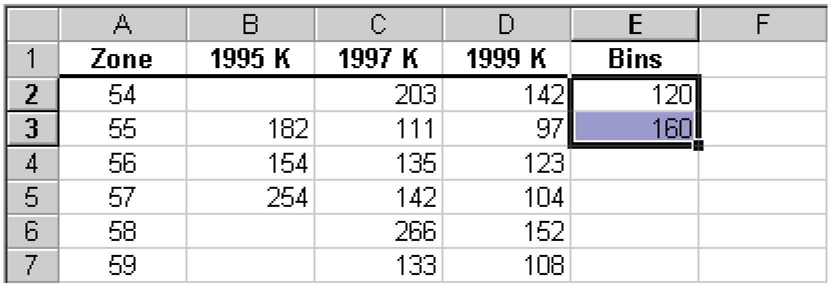
	A	B	C	D	E
1	**Zone**	**1995 K**	**1997 K**	**1999 K**	Bins
2	54		203	142	
3	55	182	111	97	

3. Enter **120** in cell **E2** and **160** in cell **E3**.



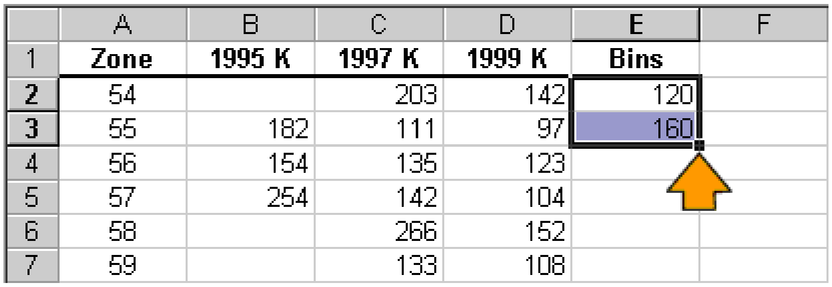
	A	B	C	D	E	F
1	Zone	1995 K	1997 K	1999 K	Bins	
2	54		203	142	120	
3	55	182	111	97	160	
4	56	154	135	123		
5	57	254	142	104		
6	58		266	152		
7	59		133	108		

4. Select cells **E2:E3**.



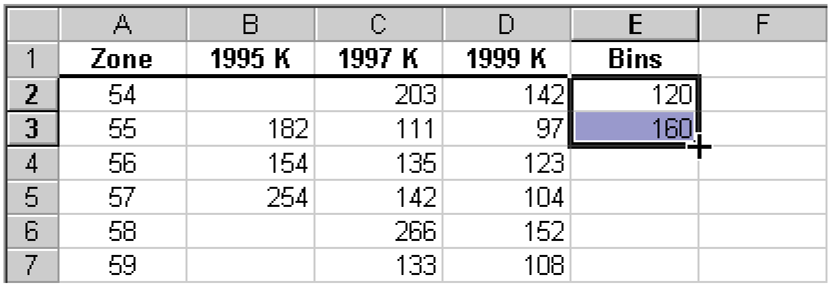
	A	B	C	D	E	F
1	Zone	1995 K	1997 K	1999 K	Bins	
2	54		203	142	120	
3	55	182	111	97	160	
4	56	154	135	123		
5	57	254	142	104		
6	58		266	152		
7	59		133	108		

The fill handle is located at the bottom right hand corner of the selected area.



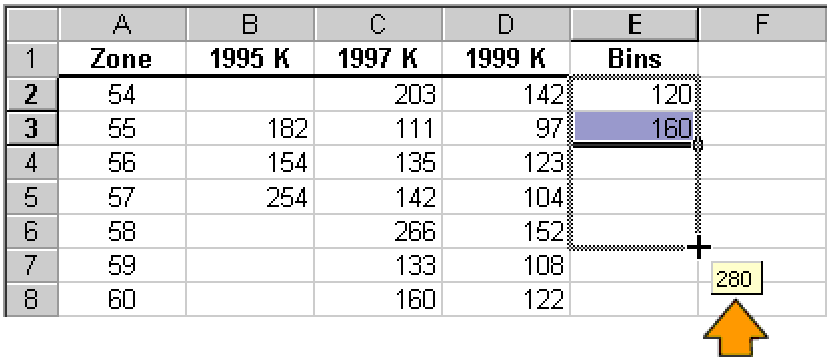
	A	B	C	D	E	F
1	Zone	1995 K	1997 K	1999 K	Bins	
2	54		203	142	120	
3	55	182	111	97	160	
4	56	154	135	123		
5	57	254	142	104		
6	58		266	152		
7	59		133	108		

5. Move the cursor over the fill handle. This will change the cursor to a cross.



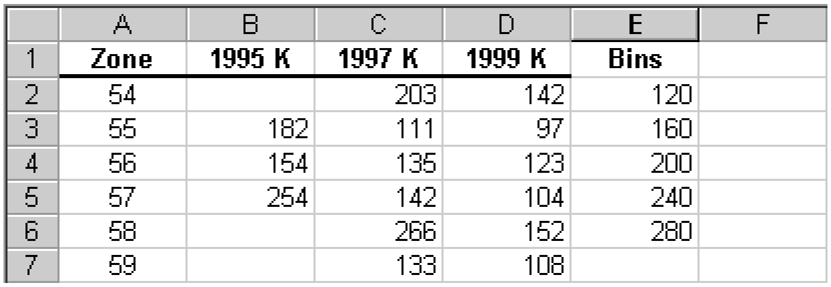
	A	B	C	D	E	F
1	Zone	1995 K	1997 K	1999 K	Bins	
2	54		203	142	120	
3	55	182	111	97	160	
4	56	154	135	123		
5	57	254	142	104		
6	58		266	152		
7	59		133	108		

6. Click and drag the cursor from cell **E3** to the bottom of cell **E6**.



	A	B	C	D	E	F
1	Zone	1995 K	1997 K	1999 K	Bins	
2	54		203	142	120	
3	55	182	111	97	160	
4	56	154	135	123		
5	57	254	142	104		
6	58		266	152		
7	59		133	108		
8	60		160	122		

You will notice that as you begin to drag the cursor, a box to the bottom right of the fill handle appears. This box contains the value to be placed into the last cell highlighted during the dragging process. In this case, it tells us that the value 280 will be placed into cell E6. This is as high as we want to go with our bins. It is the upper limit of the last category we need to encompass all of the soil test levels in our data sets. Using the fill handle has created a column of numbers that represent all of the upper class limits in Table 4.3.



	A	B	C	D	E	F
1	Zone	1995 K	1997 K	1999 K	Bins	
2	54		203	142	120	
3	55	182	111	97	160	
4	56	154	135	123	200	
5	57	254	142	104	240	
6	58		266	152	280	
7	59		133	108		

To review how bins work in Excel, let's look at the first few entries in our newly created bins column. The first number, 120, is the upper limit for the first category. It will include all values in each data set up to and including 120. This means that soil tests ranging from 0-120 will be put into this category. The next bin number is 160. This category will include all soil tests that are above 120 but less than or equal to 160. A summary of how Excel interprets bin numbers is in the table below.

Table 4.4. Example bins and their interpretation

An Excel Bin of:	Means:
120	Everything equal to or below 120 (for first category only)
160	Everything above 120 up to an including 160
200	Everything above 160 up to an including 200
Etc.	

The next step in creating a distribution is to count how many observations fall into each of the classes we have established. If we have created our classes correctly, each observation should fall into one and only one class. At this point, we are interested solely in the *number of observations*, termed the **frequency**. For instance, if we scanned our 1995 data set, we would discover that there are 4 samples with soil test levels in the 161-200 ppm class. The frequency of this class is therefore 4. If we went through all observations, we would come up with the following distribution, termed a **frequency**

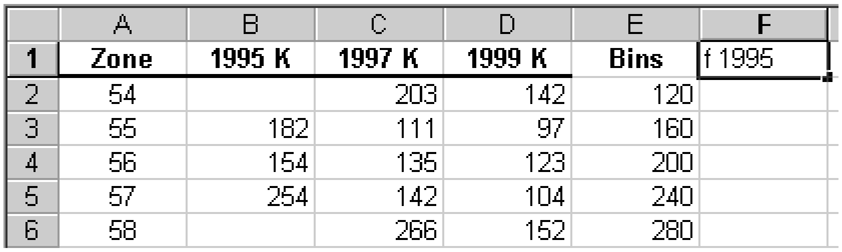
distribution, since it tells how many samples fall into each class. This distribution is fundamental and provides the starting point for other types of distributions.

Table 4.5. Frequency distribution.

Category			1995 K
Min.		Max.	Frequency
---- (ppm) ----			
81	-	120	1
121	-	160	2
161	-	200	4
201	-	240	1
241	-	280	1
		Total:	9

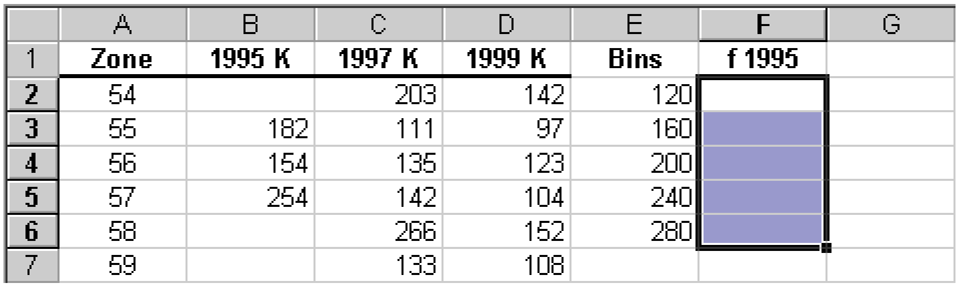
Obviously, it is not desirable to do all of this counting by hand, so we will use the Frequency function in Excel to create this distribution for us.

7. Type **f 1995** in cell **F1** to remind you that this column contains the frequency distribution of the 1995 soil test K data.



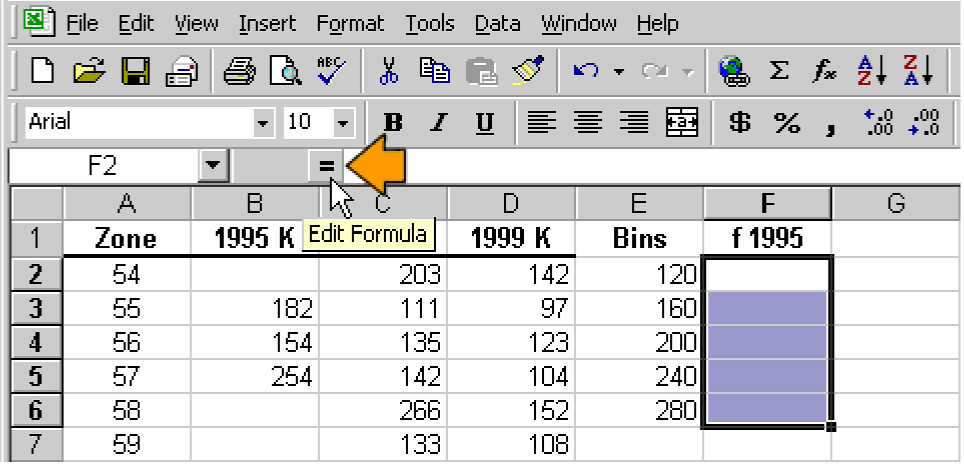
	A	B	C	D	E	F
1	Zone	1995 K	1997 K	1999 K	Bins	f 1995
2	54		203	142	120	
3	55	182	111	97	160	
4	56	154	135	123	200	
5	57	254	142	104	240	
6	58		266	152	280	

8. Select cells **F2:F6**.



	A	B	C	D	E	F	G
1	Zone	1995 K	1997 K	1999 K	Bins	f 1995	
2	54		203	142	120		
3	55	182	111	97	160		
4	56	154	135	123	200		
5	57	254	142	104	240		
6	58		266	152	280		
7	59		133	108			

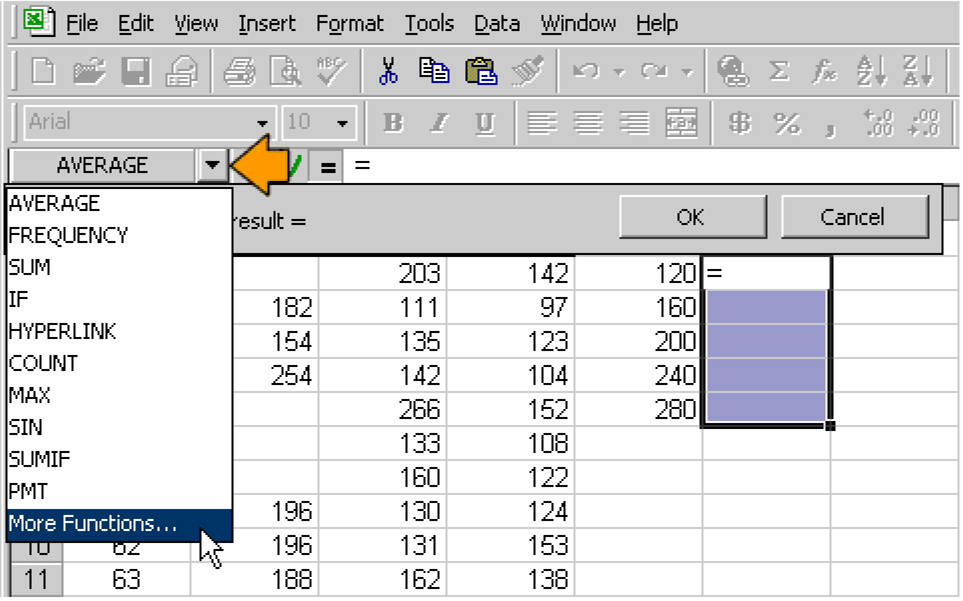
9. On the Formula bar, click the **Edit formula** button, which is labeled with an **=**.



	A	B	C	D	E	F	G
1	Zone	1995 K		1999 K	Bins	f 1995	
2	54		203	142	120		
3	55	182	111	97	160		
4	56	154	135	123	200		
5	57	254	142	104	240		
6	58		266	152	280		
7	59		133	108			

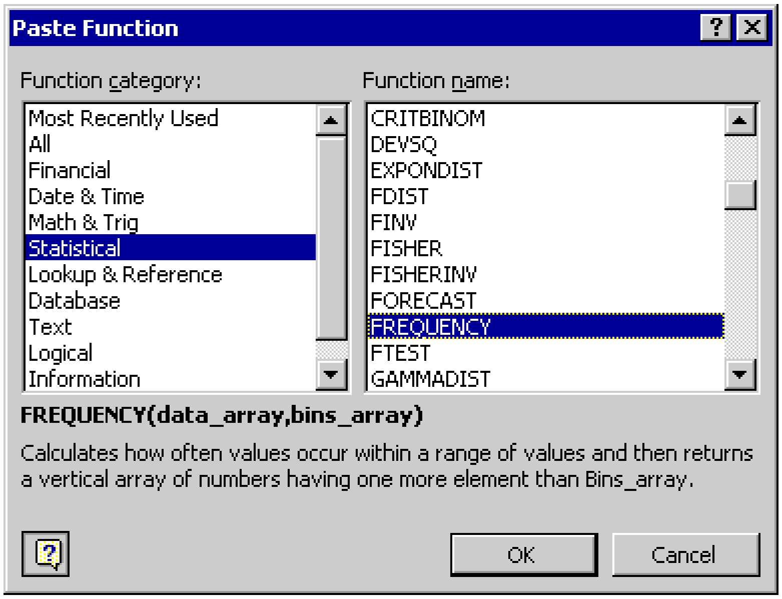
Hint: If you cannot find the formula bar, on the Standard Toolbar click **View | Formula Bar**. A check should appear to the left of the Formula Bar item on the drop down menu and the Formula Bar should appear.

10. Click the **down arrow** to the right of the current function name. In the drop down list, click **More Functions**.



	A	B	C	D	E	F	G
			203	142	120	=	
		182	111	97	160		
		154	135	123	200		
		254	142	104	240		
			266	152	280		
			133	108			
			160	122			
		196	130	124			
10	62	196	131	153			
11	63	188	162	138			

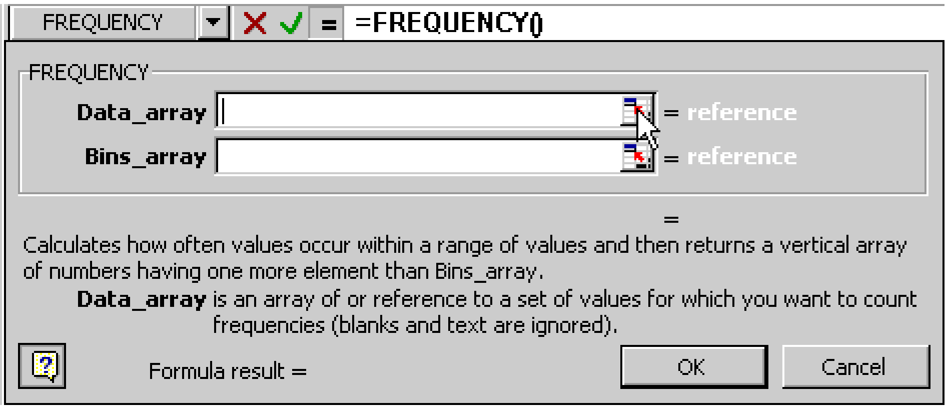
11. In the Paste Function dialog box, select **Statistical** in the Function category pane. Select **FREQUENCY** in the Function name pane. Click **OK**.



FREQUENCY(data_array,bins_array)

Calculates how often values occur within a range of values and then returns a vertical array of numbers having one more element than Bins_array.

12. Click in the blank box to the right of the **Data_array** label then click the ▣ button to the right of the text box.



=FREQUENCY()

Data_array = reference

Bins_array = reference

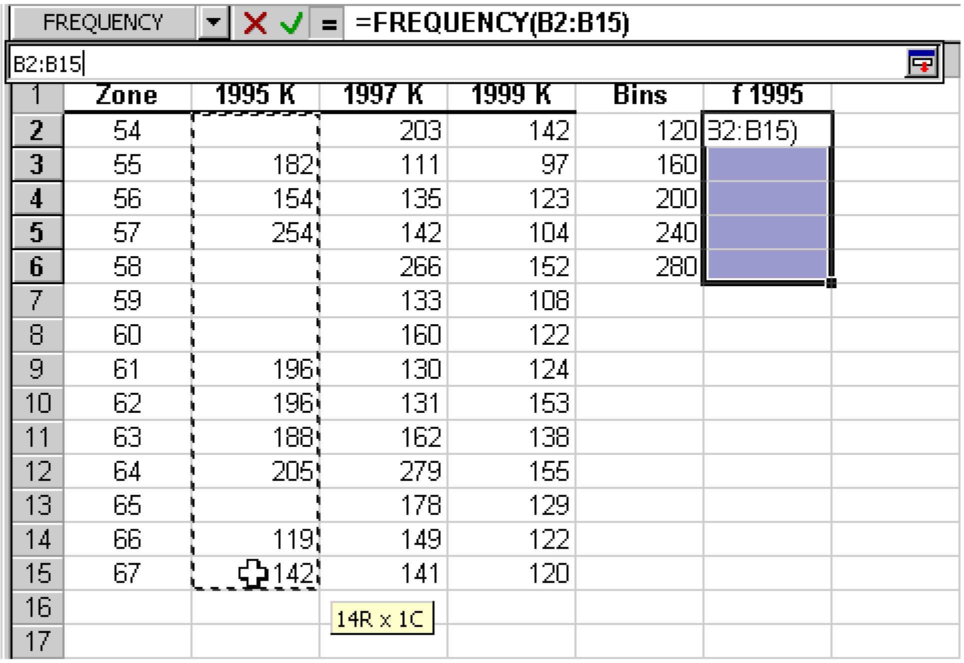
=

Calculates how often values occur within a range of values and then returns a vertical array of numbers having one more element than Bins_array.

Data_array is an array of or reference to a set of values for which you want to count frequencies (blanks and text are ignored).

Formula result =

13. Select cells **B2:B15**.

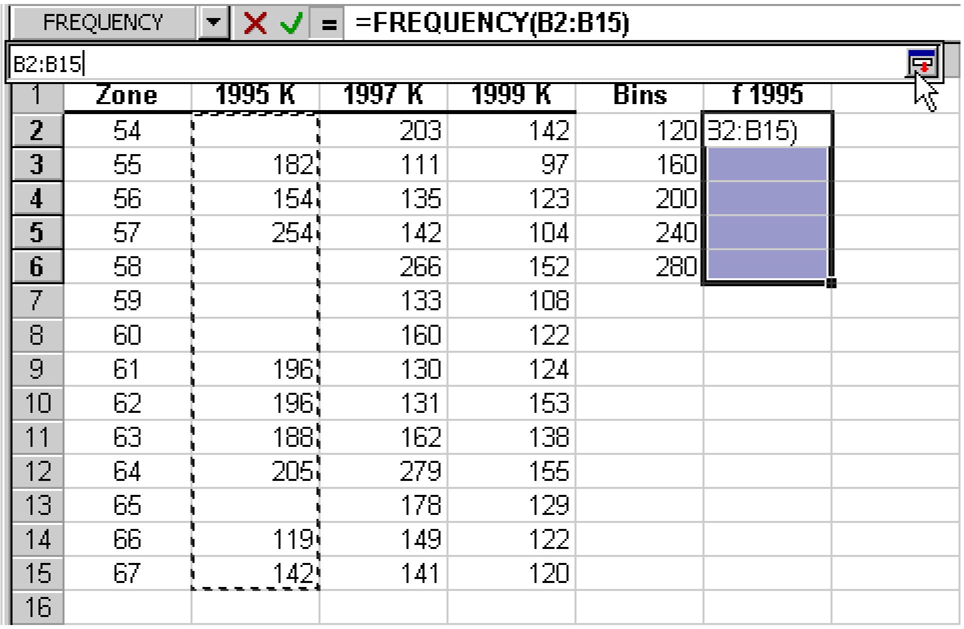


FREQUENCY =FREQUENCY(B2:B15)

B2:B15

1	Zone	1995 K	1997 K	1999 K	Bins	f 1995
2	54		203	142	120	32:B15)
3	55	182	111	97	160	
4	56	154	135	123	200	
5	57	254	142	104	240	
6	58		266	152	280	
7	59		133	108		
8	60		160	122		
9	61	196	130	124		
10	62	196	131	153		
11	63	188	162	138		
12	64	205	279	155		
13	65		178	129		
14	66	119	149	122		
15	67	142	141	120		
16						
17						

14. Click the button to the right of the text box.

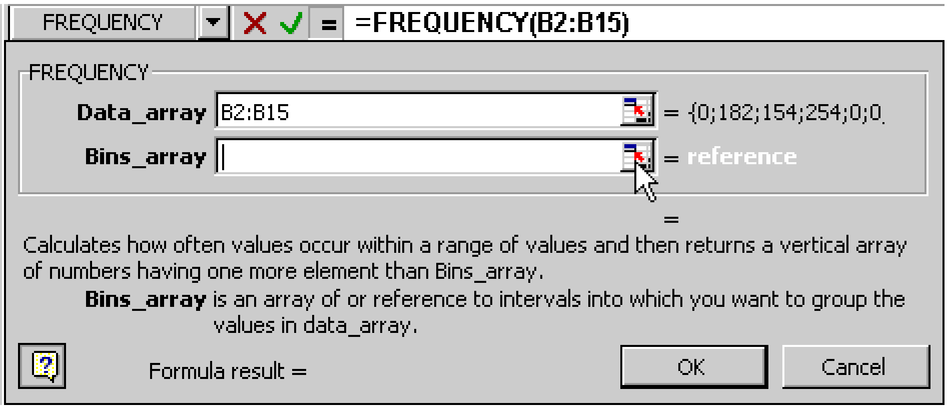


FREQUENCY =FREQUENCY(B2:B15)

B2:B15

1	Zone	1995 K	1997 K	1999 K	Bins	f 1995
2	54		203	142	120	32:B15)
3	55	182	111	97	160	
4	56	154	135	123	200	
5	57	254	142	104	240	
6	58		266	152	280	
7	59		133	108		
8	60		160	122		
9	61	196	130	124		
10	62	196	131	153		
11	63	188	162	138		
12	64	205	279	155		
13	65		178	129		
14	66	119	149	122		
15	67	142	141	120		
16						

15. Click the text box to the right of the label Bins_array then click the button.



=FREQUENCY(B2:B15)

FREQUENCY

Data_array B2:B15 = {0;182;154;254;0;0

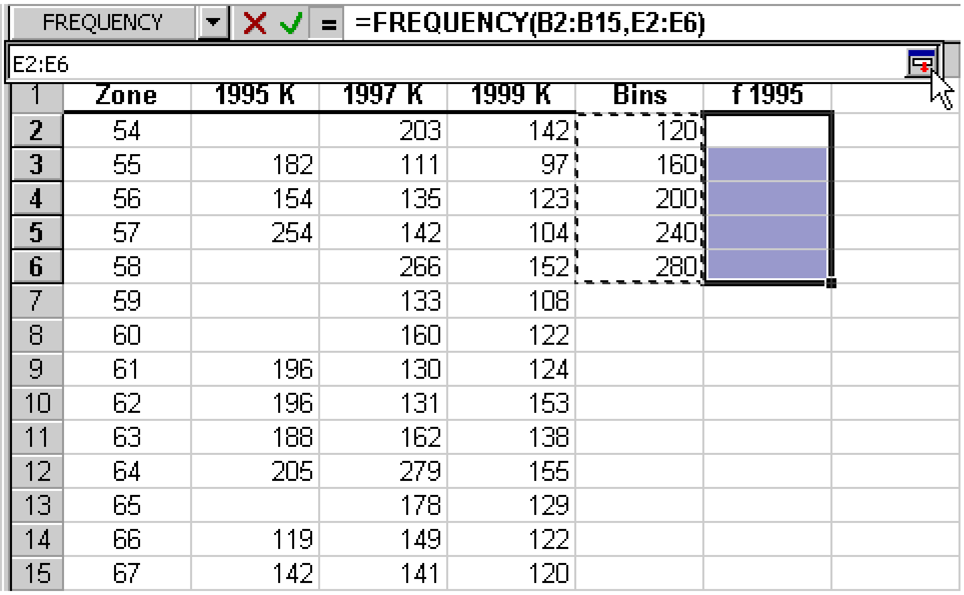
Bins_array = reference

Calculates how often values occur within a range of values and then returns a vertical array of numbers having one more element than Bins_array.

Bins_array is an array of or reference to intervals into which you want to group the values in data_array.

Formula result =

16. Select cells **E2:E6**.

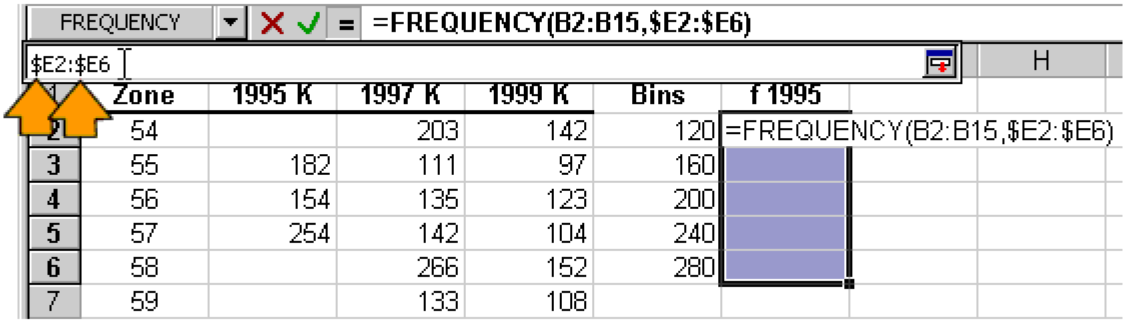


=FREQUENCY(B2:B15,E2:E6)

E2:E6

1	Zone	1995 K	1997 K	1999 K	Bins	f 1995
2	54		203	142	120	
3	55	182	111	97	160	
4	56	154	135	123	200	
5	57	254	142	104	240	
6	58		266	152	280	
7	59		133	108		
8	60		160	122		
9	61	196	130	124		
10	62	196	131	153		
11	63	188	162	138		
12	64	205	279	155		
13	65		178	129		
14	66	119	149	122		
15	67	142	141	120		

17. Use the mouse to move the cursor to the space just before the first E and type **$**. Then use the mouse to move the cursor to the space between the colon and the second E, and type another **$**. Click the button to the right of the text box.



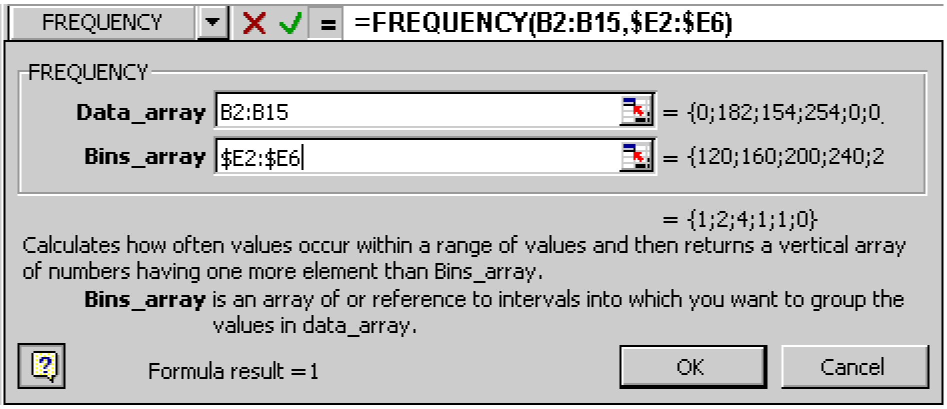
=FREQUENCY(B2:B15,$E2:$E6)

$E2:$E6

1	Zone	1995 K	1997 K	1999 K	Bins	f 1995	H
2	54		203	142	120	=FREQUENCY(B2:B15,$E2:$E6)	
3	55	182	111	97	160		
4	56	154	135	123	200		
5	57	254	142	104	240		
6	58		266	152	280		
7	59		133	108			

The **$** creates an absolute reference that does not change when cell contents are copied or moved to a new location in the spreadsheet. For instance, $E2:$E6 keeps column E fixed as the column that contains the bins. We will take advantage of this to create the remaining two distributions for the 1997 and 1999 soil test K data.

18. Hold down the **Ctrl** and **Shift** keys at the same time, then press **Enter**. DO NOT click OK.



FREQUENCY =FREQUENCY(B2:B15,$E2:$E6)

FREQUENCY

Data_array B2:B15 = {0;182;154;254;0;0

Bins_array $E2:$E6 = {120;160;200;240;2

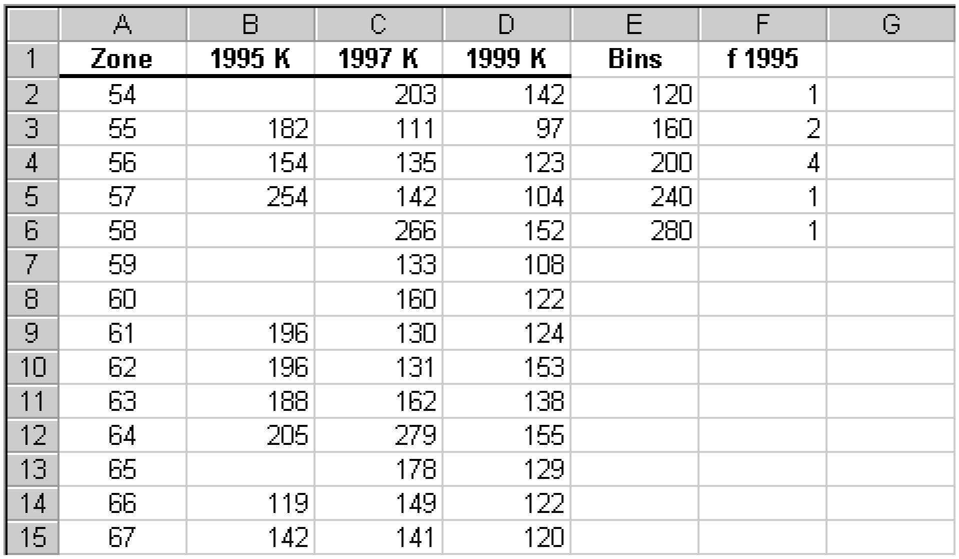
= {1;2;4;1;1;0}

Calculates how often values occur within a range of values and then returns a vertical array of numbers having one more element than Bins_array.

Bins_array is an array of or reference to intervals into which you want to group the values in data_array.

Formula result =1

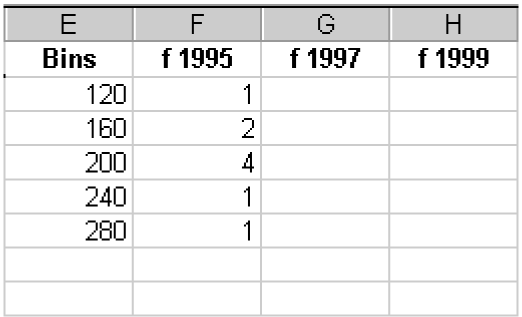
The Frequency function fills in the appropriate frequencies for each of the classes we created.



	A	B	C	D	E	F	G
1	**Zone**	**1995 K**	**1997 K**	**1999 K**	**Bins**	**f 1995**	
2	54		203	142	120	1	
3	55	182	111	97	160	2	
4	56	154	135	123	200	4	
5	57	254	142	104	240	1	
6	58		266	152	280	1	
7	59		133	108			
8	60		160	122			
9	61	196	130	124			
10	62	196	131	153			
11	63	188	162	138			
12	64	205	279	155			
13	65		178	129			
14	66	119	149	122			
15	67	142	141	120			

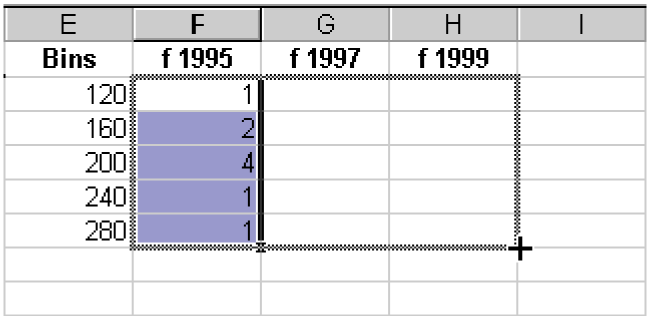
Now we must create frequency distributions for the 1997 and 1999 soil test K data.

19. Type **f 1997** in cell G1 and **f 1999** in cell H1.



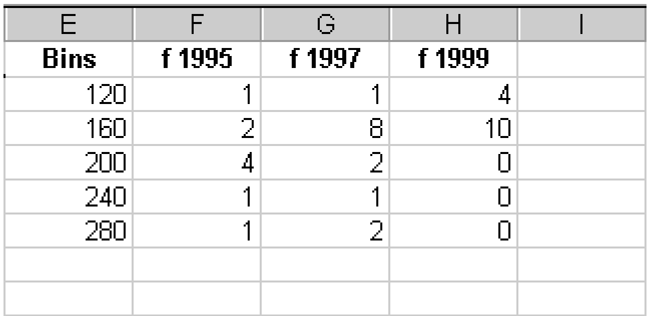
E	F	G	H
Bins	**f 1995**	**f 1997**	**f 1999**
120	1		
160	2		
200	4		
240	1		
280	1		

20. Select cells **F2:F6**. Move the cursor to the **fill handle** and **click and drag** the cursor to the bottom right hand corer of cell **H6**.



E	F	G	H	I
Bins	**f 1995**	**f 1997**	**f 1999**	
120	1			
160	2			
200	4			
240	1			
280	1			

We have now calculated frequency distributions for the soil test K data from 1997 and 1999. We were able to use the fill handle to do this because we used relative cell references for the data array and absolute references for the bins.



E	F	G	H	I
Bins	**f 1995**	**f 1997**	**f 1999**	
120	1	1	4	
160	2	8	10	
200	4	2	0	
240	1	1	0	
280	1	2	0	

The calculation of the frequency distributions for all 3 years of data is now complete.

Creating a relative frequency distribution

Rather than just look at numbers of samples, it may at times be more meaningful to look at what proportion of the samples fall into various classes. This is particularly useful when comparing distributions from several different fields or years, as we are interested in doing. Directly comparing numbers of samples makes little sense in such cases, because the total number of samples may vary from field to field and/or year to year. What we want instead is to describe how the *percent* of the total number of samples differs across our classes. This is termed the **relative frequency distribution.** For instance the 5 samples in the 0-120 ppm range for the 1999 soil test K data represent 28.6 percent of the total number of samples. We calculated this by dividing the frequency (4) by the total number of samples (14) and then multiplying by 100 to express the result as a percentage:

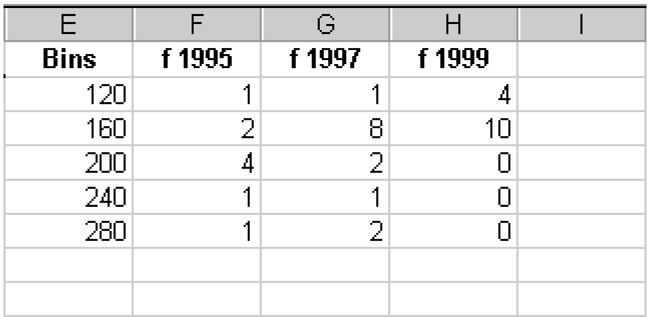
$$\left(\frac{4}{14}\right)\times 100 = 28.6\% .$$

The general formula for calculating a relative frequency is:

[3.1] $$\text{Relative frequency (\%)} = \left(\frac{\text{Frequency of a class}}{\text{Total number of samples}}\right)\times 100 .$$

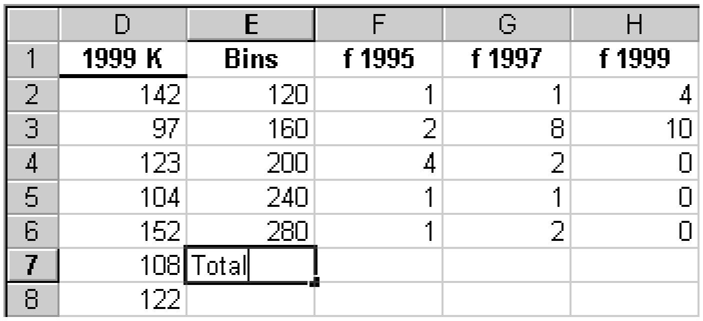
We now will create relative frequency distributions for the three years of soil test data. We begin by first calculating the total number of observations for each data set. These totals will become the basis for our relative frequency calculations.

21. Start Microsoft Excel. On the Standard toolbar, Click File | Open and navigate to C:\PPIStat. Open file **Ex03**.



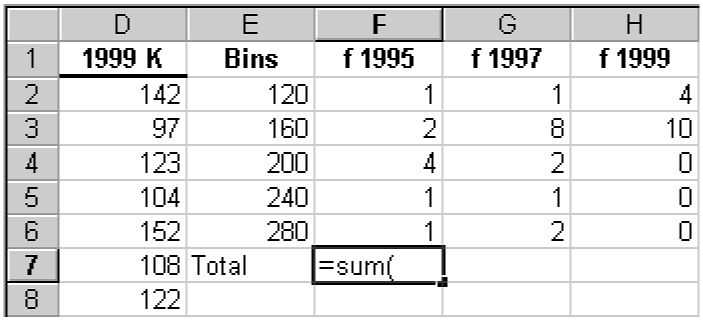
E	F	G	H	I
Bins	**f 1995**	**f 1997**	**f 1999**	
120	1	1	4	
160	2	8	10	
200	4	2	0	
240	1	1	0	
280	1	2	0	

22. Type **Total** into cell **E7** and press **Enter**.



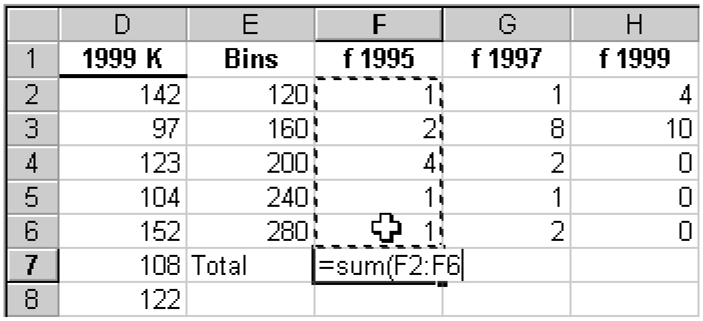
	D	E	F	G	H
1	**1999 K**	**Bins**	**f 1995**	**f 1997**	**f 1999**
2	142	120	1	1	4
3	97	160	2	8	10
4	123	200	4	2	0
5	104	240	1	1	0
6	152	280	1	2	0
7	108	Total			
8	122				

23. In cell **F7**, type **=sum(**.



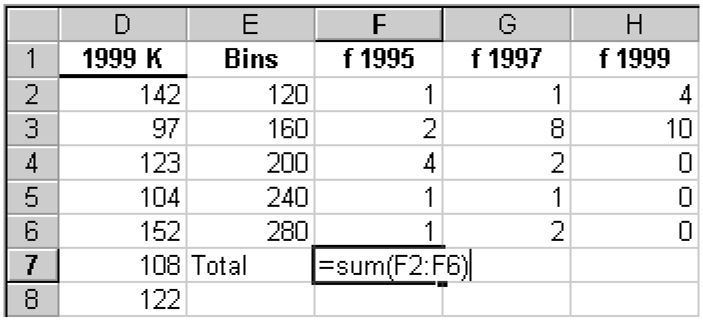
	D	E	F	G	H
1	**1999 K**	**Bins**	**f 1995**	**f 1997**	**f 1999**
2	142	120	1	1	4
3	97	160	2	8	10
4	123	200	4	2	0
5	104	240	1	1	0
6	152	280	1	2	0
7	108	Total	=sum(		
8	122				

24. Select cells **F1:F6**.



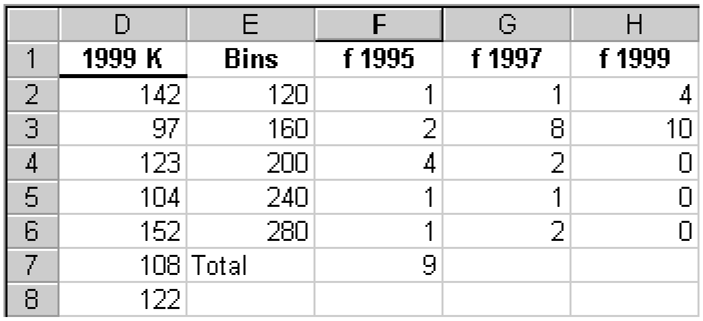
	D	E	F	G	H
1	**1999 K**	**Bins**	**f 1995**	**f 1997**	**f 1999**
2	142	120	1	1	4
3	97	160	2	8	10
4	123	200	4	2	0
5	104	240	1	1	0
6	152	280	1	2	0
7	108	Total	=sum(F2:F6		
8	122				

25. Type **)** and press **Enter**.



	D	E	F	G	H
1	1999 K	Bins	f 1995	f 1997	f 1999
2	142	120	1	1	4
3	97	160	2	8	10
4	123	200	4	2	0
5	104	240	1	1	0
6	152	280	1	2	0
7	108	Total	=sum(F2:F6)		
8	122				

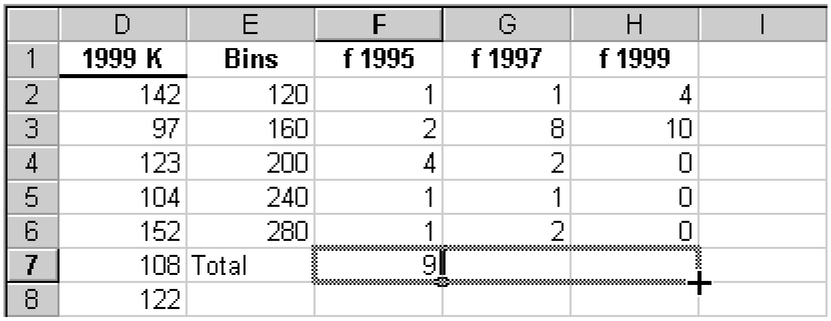
The total number of soil samples for 1995 appears in cell F7. The sum function added together all of the observations in the selected range.



	D	E	F	G	H
1	1999 K	Bins	f 1995	f 1997	f 1999
2	142	120	1	1	4
3	97	160	2	8	10
4	123	200	4	2	0
5	104	240	1	1	0
6	152	280	1	2	0
7	108	Total	9		
8	122				

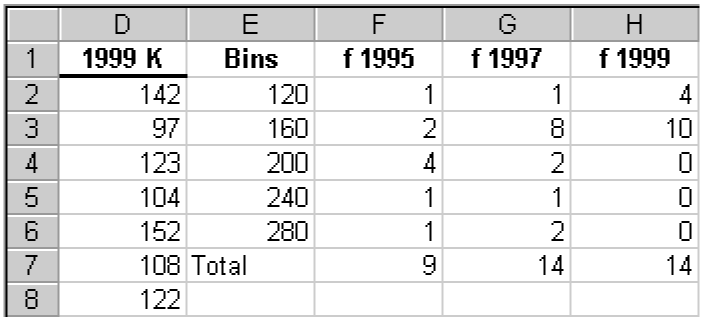
We next create totals for the 1997 and 1999 data by using the fill handle to copy the sum formula from cell F7 to cells G7 and H7.

26. Click the **fill handle** and drag the cursor to the right side of cell **H7**.



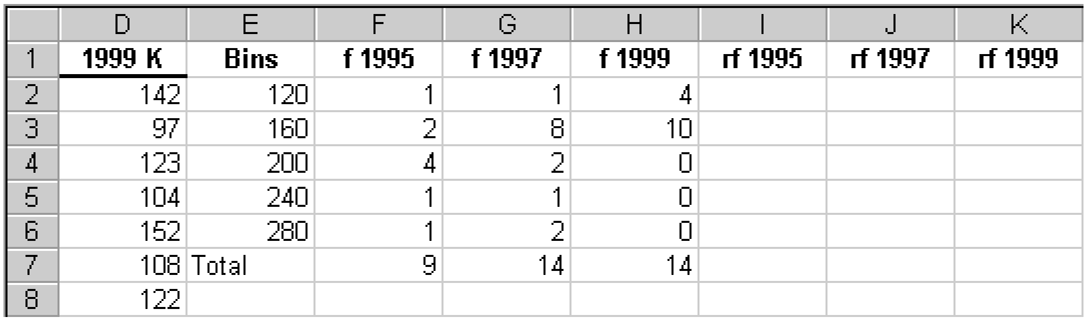
	D	E	F	G	H	I
1	1999 K	Bins	f 1995	f 1997	f 1999	
2	142	120	1	1	4	
3	97	160	2	8	10	
4	123	200	4	2	0	
5	104	240	1	1	0	
6	152	280	1	2	0	
7	108	Total	9			
8	122					

We now have the total number of observations calculated for each of the three years of data.



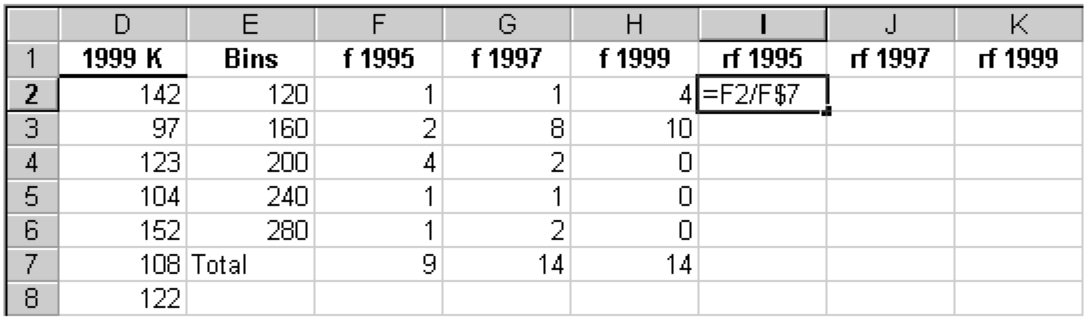
	D	E	F	G	H
1	1999 K	Bins	f 1995	f 1997	f 1999
2	142	120	1	1	4
3	97	160	2	8	10
4	123	200	4	2	0
5	104	240	1	1	0
6	152	280	1	2	0
7	108	Total	9	14	14
8	122				

27. Type **rf 1995**, **rf 1997** and **rf 1999** in cells **I1**, **J1**, and **K1**, respectively. These labels are a reminder that these columns contain relative frequencies.



	D	E	F	G	H	I	J	K
1	1999 K	Bins	f 1995	f 1997	f 1999	rf 1995	rf 1997	rf 1999
2	142	120	1	1	4			
3	97	160	2	8	10			
4	123	200	4	2	0			
5	104	240	1	1	0			
6	152	280	1	2	0			
7	108	Total	9	14	14			
8	122							

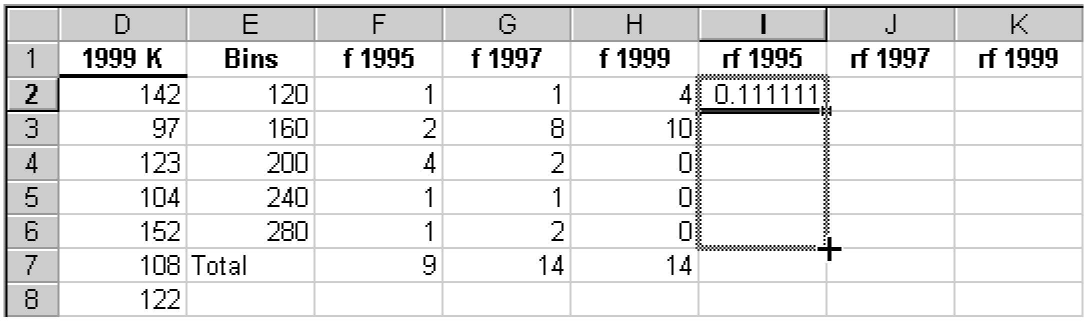
28. Type **=F2/F$7** in cell **I2**.



	D	E	F	G	H	I	J	K
1	1999 K	Bins	f 1995	f 1997	f 1999	rf 1995	rf 1997	rf 1999
2	142	120	1	1	4	=F2/F$7		
3	97	160	2	8	10			
4	123	200	4	2	0			
5	104	240	1	1	0			
6	152	280	1	2	0			
7	108	Total	9	14	14			
8	122							

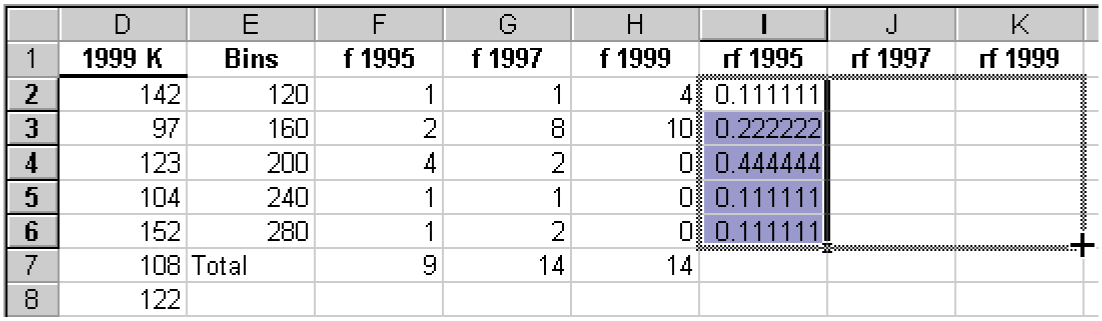
This formula will divide the frequency in cell F2 by the total number of samples in cell F7. The absolute reference $7 keeps the row for the denominator frozen at row 7, the row containing the totals. The column reference F is kept relative so that totals in columns G and H will be accessed when we use the fill handle later.

29. Select cell **I2**, click on the **fill handle**, and drag the cursor to the bottom of cell **I6**.



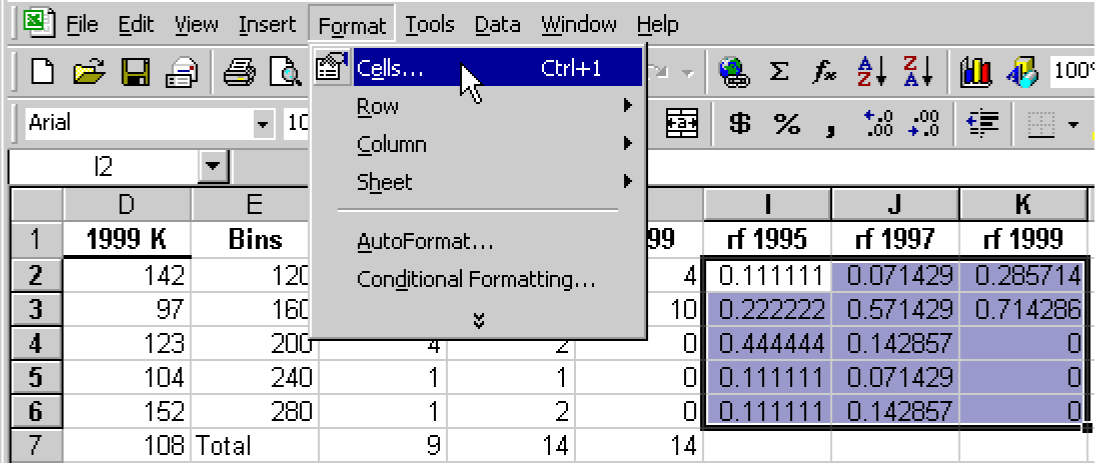
	D	E	F	G	H	I	J	K
1	1999 K	Bins	f 1995	f 1997	f 1999	rf 1995	rf 1997	rf 1999
2	142	120	1	1	4	0.111111		
3	97	160	2	8	10			
4	123	200	4	2	0			
5	104	240	1	1	0			
6	152	280	1	2	0			
7	108	Total	9	14	14			
8	122							

30. Select cells **I2:I6**, click on the **fill handle**, and drag the cursor to the bottom right hand corner of cell **K6**.



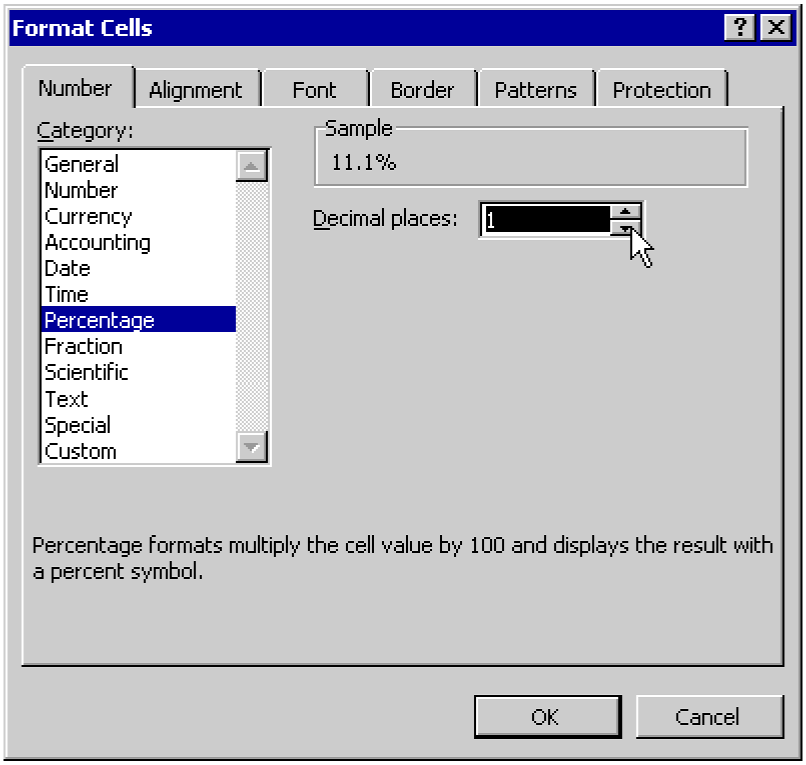
	D	E	F	G	H	I	J	K
1	1999 K	Bins	f 1995	f 1997	f 1999	rf 1995	rf 1997	rf 1999
2	142	120	1	1	4	0.111111		
3	97	160	2	8	10	0.222222		
4	123	200	4	2	0	0.444444		
5	104	240	1	1	0	0.111111		
6	152	280	1	2	0	0.111111		
7	108	Total	9	14	14			
8	122							

31. Select cells **I2:K6** and click **Format | Cells** on the Standard toolbar.



	D	E				I	J	K
1	1999 K	Bins			99	rf 1995	rf 1997	rf 1999
2	142	120			4	0.111111	0.071429	0.285714
3	97	160			10	0.222222	0.571429	0.714286
4	123	200	4	2	0	0.444444	0.142857	0
5	104	240	1	1	0	0.111111	0.071429	0
6	152	280	1	2	0	0.111111	0.142857	0
7	108	Total	9	14	14			

32. In the Format Cells dialog box, select **Percentage** in the Category pane on the Number tab. Select **1** for **Decimal places**. Click **OK**.

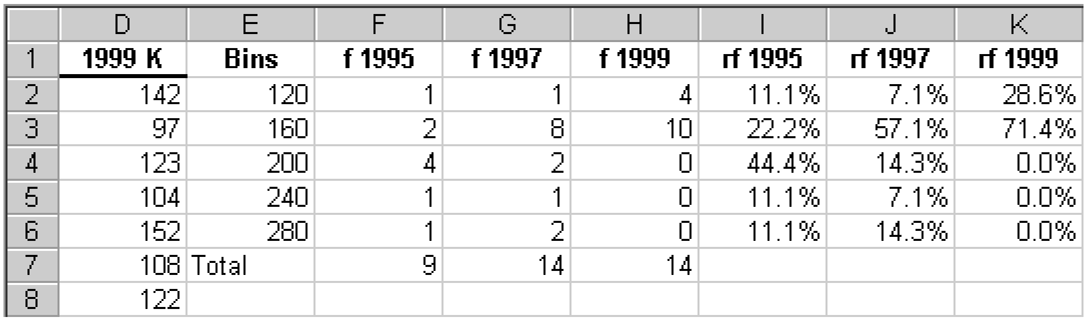


Sample
11.1%

Decimal places: 1

Percentage formats multiply the cell value by 100 and displays the result with a percent symbol.

We now have relative frequency distributions for the three years of soil test K data.

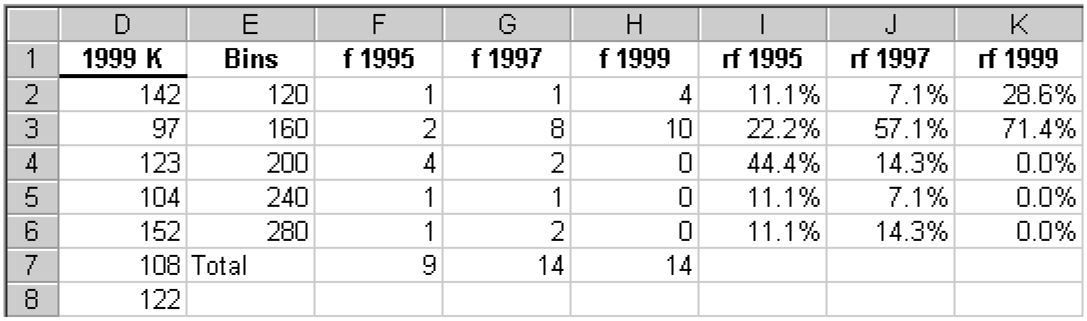


	D	E	F	G	H	I	J	K
1	**1999 K**	**Bins**	**f 1995**	**f 1997**	**f 1999**	**rf 1995**	**rf 1997**	**rf 1999**
2	142	120	1	1	4	11.1%	7.1%	28.6%
3	97	160	2	8	10	22.2%	57.1%	71.4%
4	123	200	4	2	0	44.4%	14.3%	0.0%
5	104	240	1	1	0	11.1%	7.1%	0.0%
6	152	280	1	2	0	11.1%	14.3%	0.0%
7	108	Total	9	14	14			
8	122							

Creating histograms

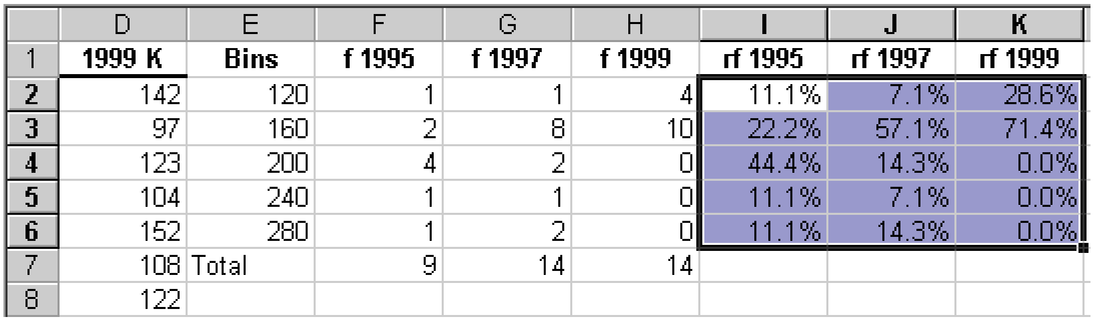
In this section, we will create a graph of the relative frequency distributions we created previously. A graph of a distribution is called a **histogram**. Histograms are typically bar charts, with the categories on the horizontal, or **x-axis**, and frequency or relative frequency on the vertical, or **y-axis**. Bars are created for each class, or bin. We will use the graphing features of Excel to create our histograms.

1. Start Microsoft Excel. On the Standard toolbar, Click File | Open and navigate to C:\PPIStat. Open file **Ex04**.



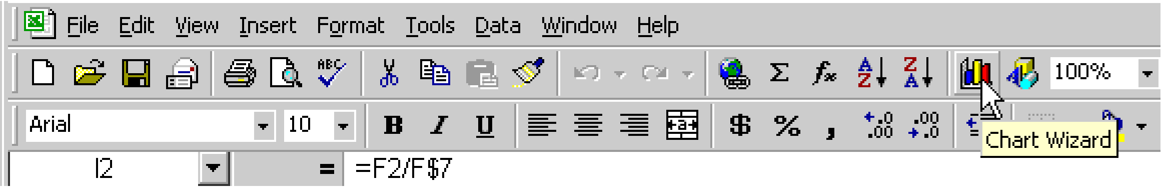
	D	E	F	G	H	I	J	K
1	1999 K	Bins	f 1995	f 1997	f 1999	rf 1995	rf 1997	rf 1999
2	142	120	1	1	4	11.1%	7.1%	28.6%
3	97	160	2	8	10	22.2%	57.1%	71.4%
4	123	200	4	2	0	44.4%	14.3%	0.0%
5	104	240	1	1	0	11.1%	7.1%	0.0%
6	152	280	1	2	0	11.1%	14.3%	0.0%
7	108	Total	9	14	14			
8	122							

2. Select cells **I2:K6**.



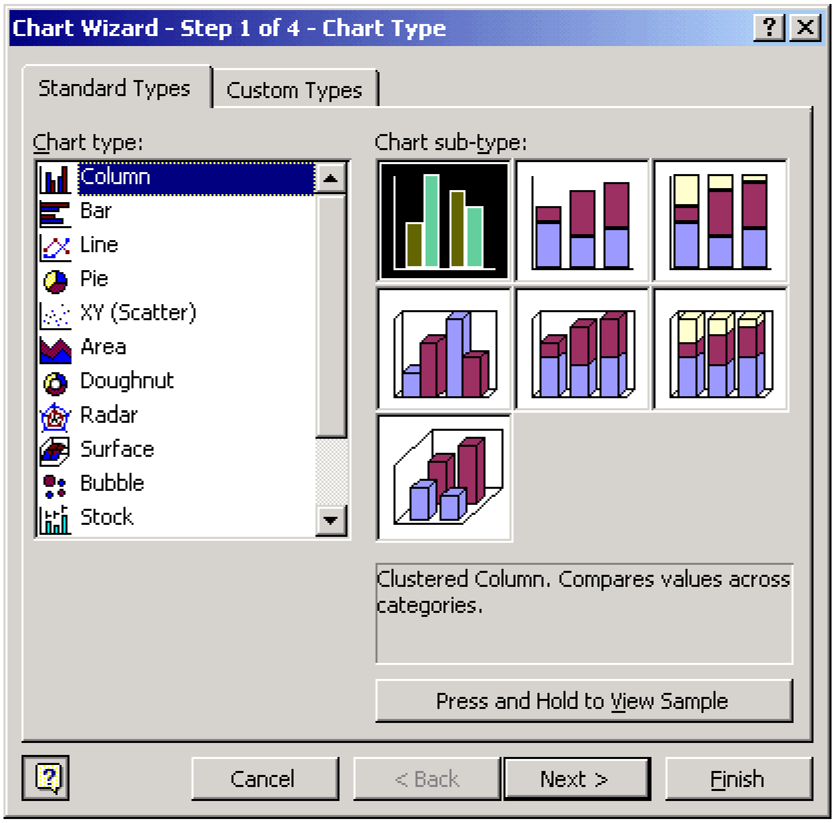
	D	E	F	G	H	I	J	K
1	1999 K	Bins	f 1995	f 1997	f 1999	rf 1995	rf 1997	rf 1999
2	142	120	1	1	4	11.1%	7.1%	28.6%
3	97	160	2	8	10	22.2%	57.1%	71.4%
4	123	200	4	2	0	44.4%	14.3%	0.0%
5	104	240	1	1	0	11.1%	7.1%	0.0%
6	152	280	1	2	0	11.1%	14.3%	0.0%
7	108	Total	9	14	14			
8	122							

3. On the Standard toolbar, click the **Chart Wizard** button.



I2 = =F2/F$7

4. In the Chart Wizard dialog box (Step 1 of 4 – Chart Type), select **Column** in the Chart type pane on the Standard Types tab. Select **Clustered Column** as the Chart sub-type. Press **Next**.



Clustered Column. Compares values across categories.

5. In the Chart Wizard dialog box (Step 2 of 4 – Chart Source Data), on the Data Range tab, select **Columns** as the option for **Series in**.

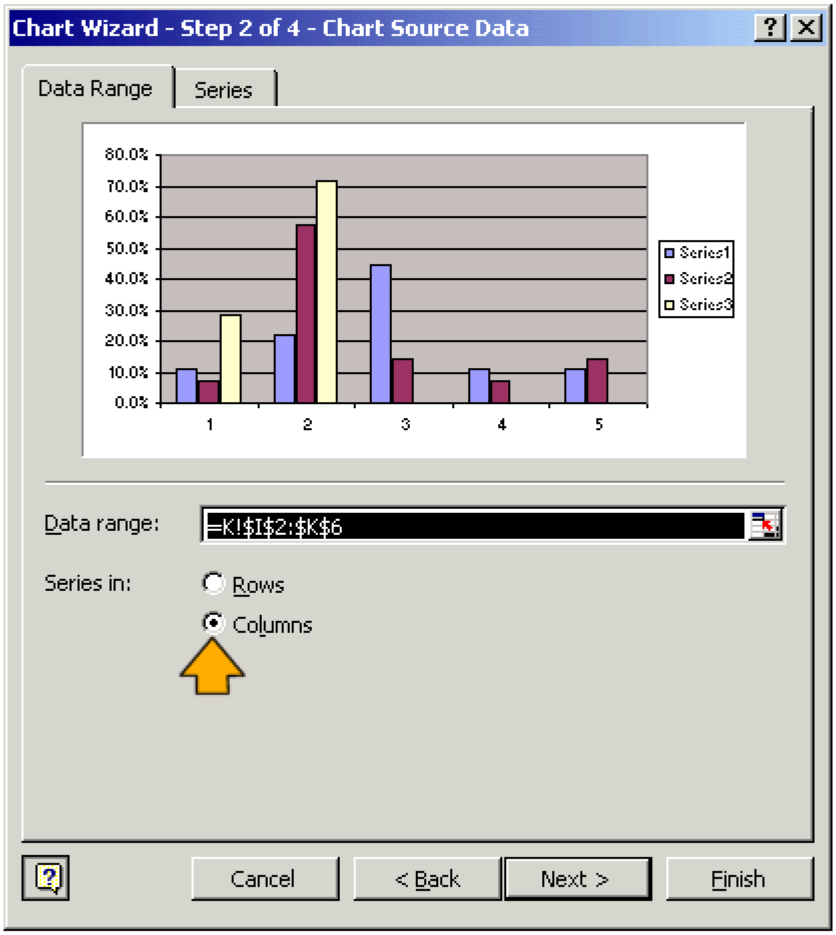
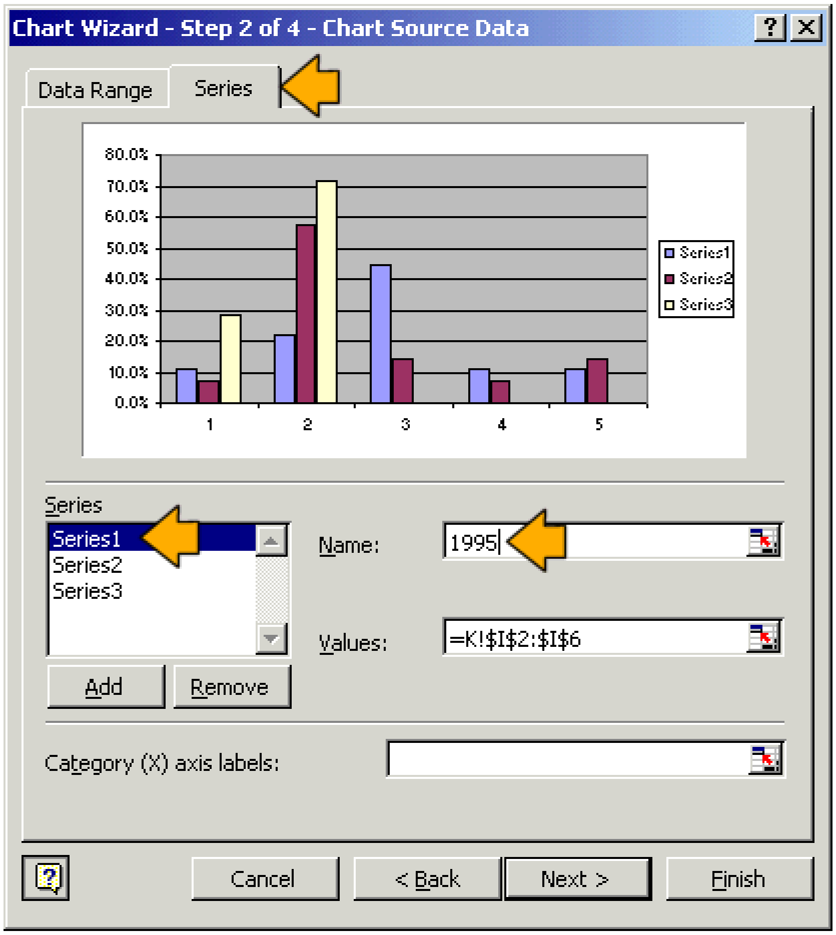


Chart Wizard - Step 2 of 4 - Chart Source Data

Data range: =K!$I$2:$K$6

Series in: ○ Rows ● Columns

6. Click on the **Series** tab. Select **Series1** in the Series pane. Click on the **Name** field. Type **1995**.

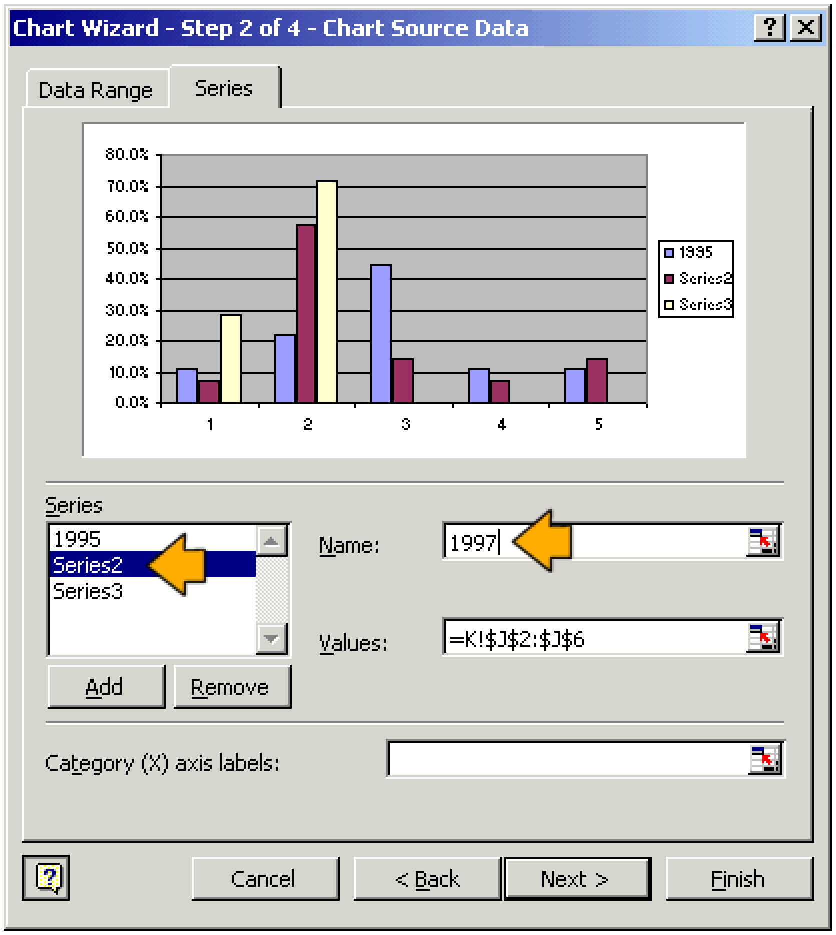


Name: 1995

Values: =K!$I$2:$I$6

Category (X) axis labels:

7. Select **Series2** in the Series pane. Click on the **Name** field. Type **1997**.



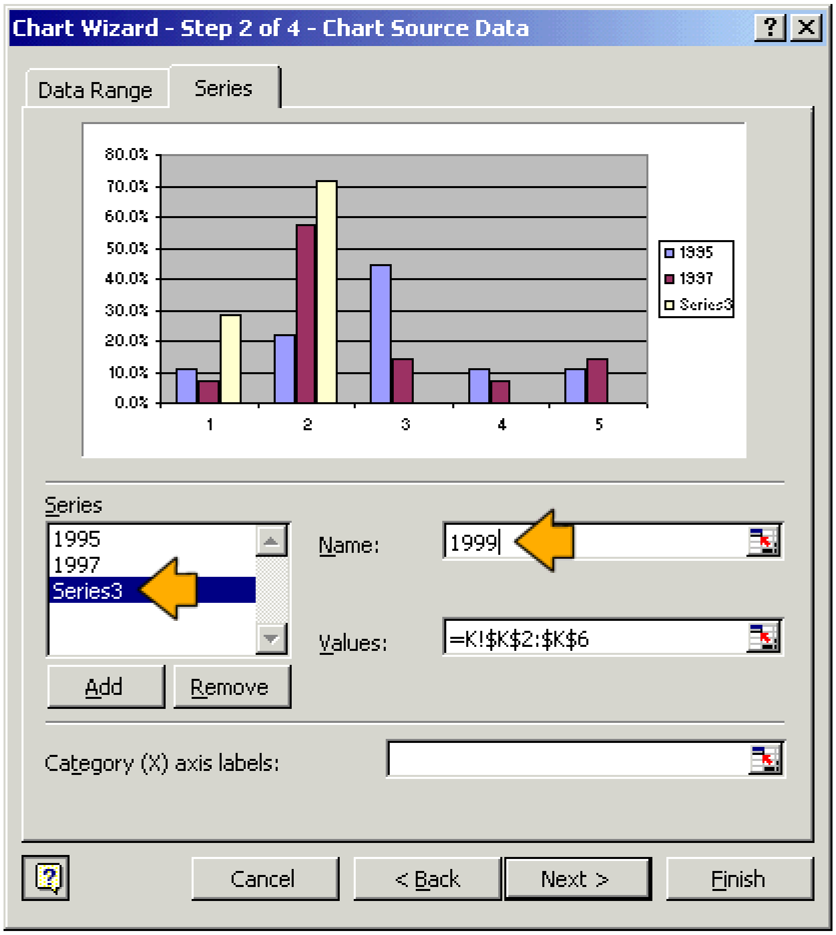
Series: 1995, Series2, Series3

Name: 1997

Values: =K!$J$2:$J$6

Category (X) axis labels:

8. Select **Series3** in the Series pane. Click on the **Name** field. Type **1999**.



Series

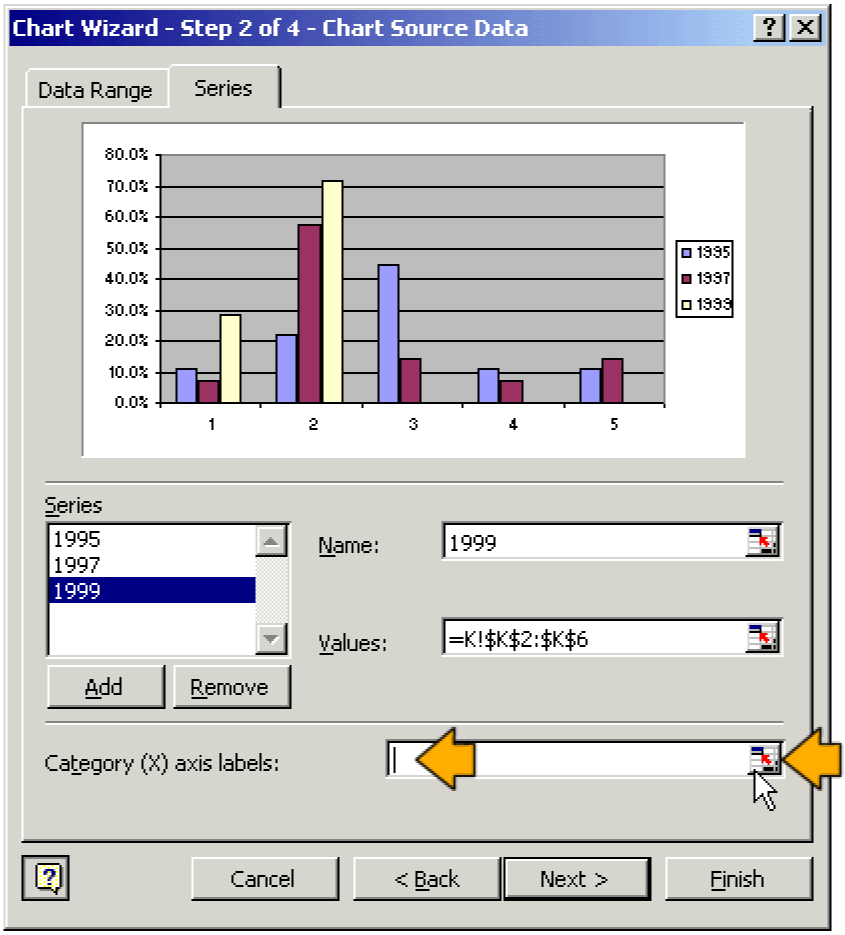
- 1995
- 1997
- Series3

Name: 1999

Values: =K!$K$2:$K$6

Category (X) axis labels:

9. Click on the text box to the right of Category (X) axis labels. Click on the ▦ button.



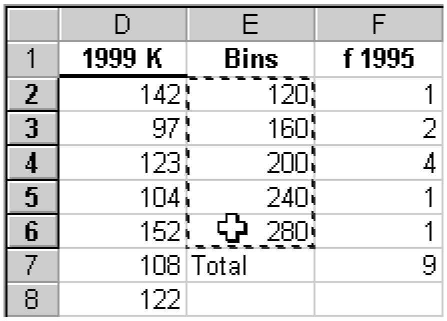
Series: 1995, 1997, 1999

Name: 1999

Values: =K!$K$2:$K$6

Category (X) axis labels:

10. Select cells **E2:E6**.



	D	E	F
1	1999 K	Bins	f 1995
2	142	120	1
3	97	160	2
4	123	200	4
5	104	240	1
6	152	280	1
7	108	Total	9
8	122		

11. Click the button to on the right side of the Chart Wizard dialog box. After the dialog box returns to its original size, click **Next**.

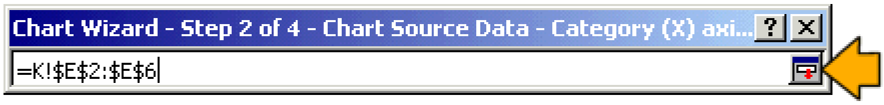


Chart Wizard - Step 2 of 4 - Chart Source Data - Category (X) axi...

=K!$E$2:$E$6

12. In the Chart Wizard dialog box (Step 3 of 4 – Chart Options), select the **Titles** tab. Type **Soil test K (ppm)** for the Category (X) axis. Type **Relative frequency** for the Value (Y) axis.

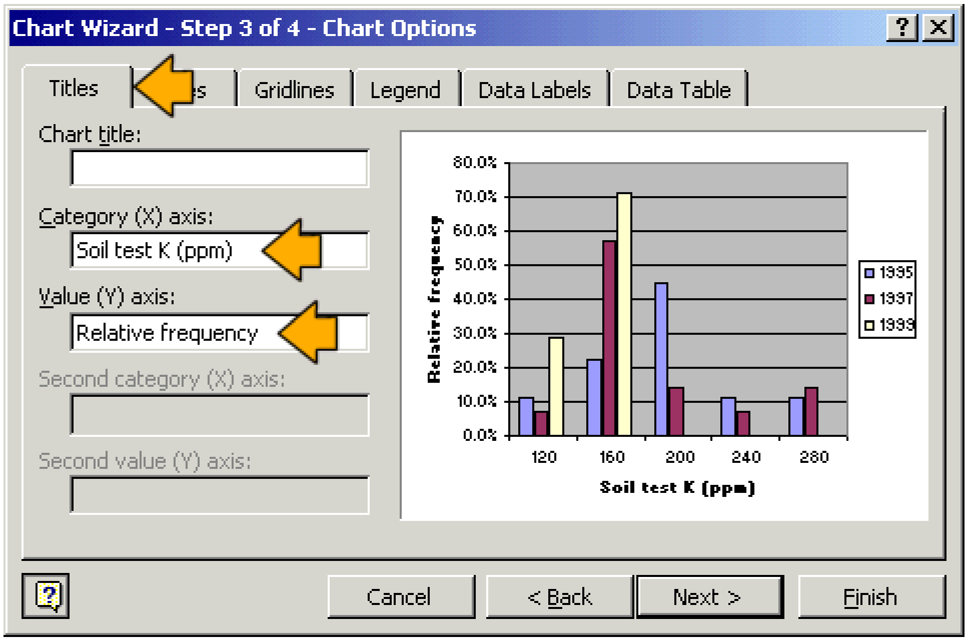


Chart Wizard - Step 3 of 4 - Chart Options

Chart title:

Category (X) axis: Soil test K (ppm)

Value (Y) axis: Relative frequency

Second category (X) axis:

Second value (Y) axis:

13. Select the **Legend** tab. Select **Bottom** in the Placement pane. Click **Next**.

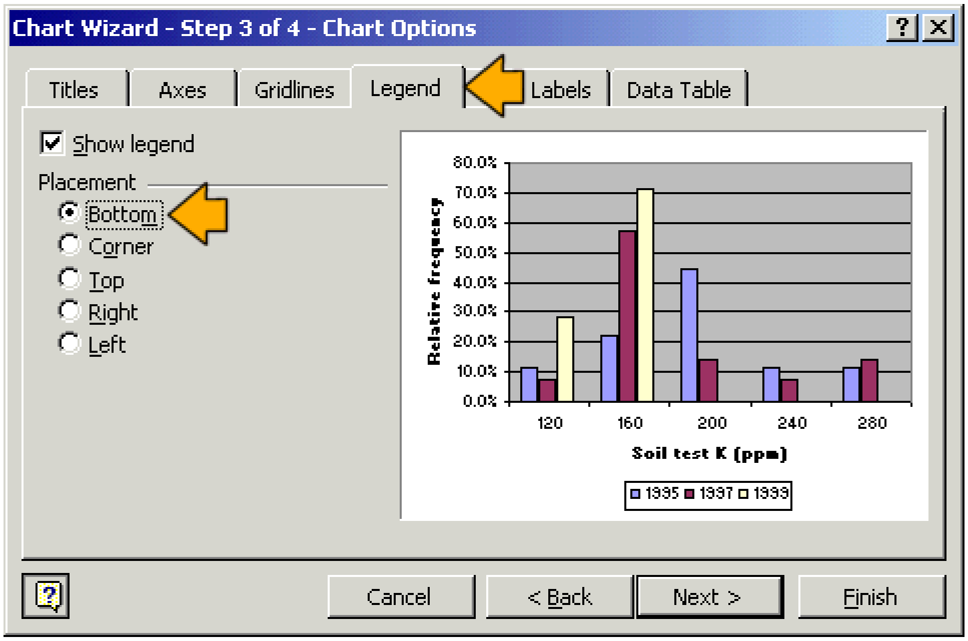


Chart Wizard - Step 3 of 4 - Chart Options

☑ Show legend

Placement

● Bottom
○ Corner
○ Top
○ Right
○ Left

14. In the Chart Wizard dialog box (Step 4 of 4 – Chart Location), select **As new sheet** and type **Histograms** in the text box. Click **Finish**.

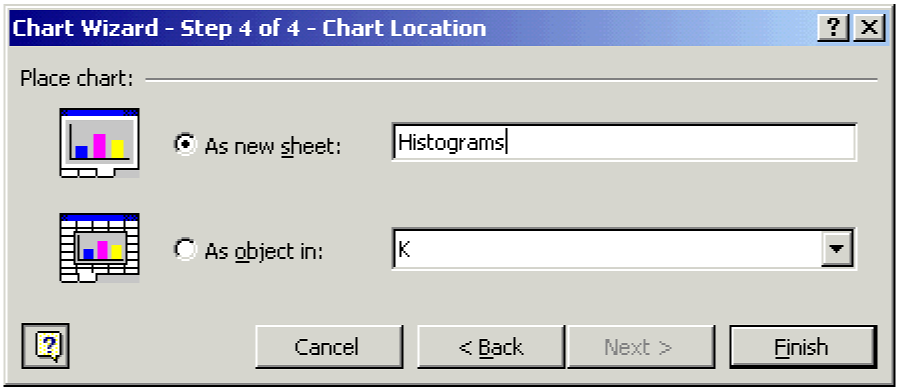


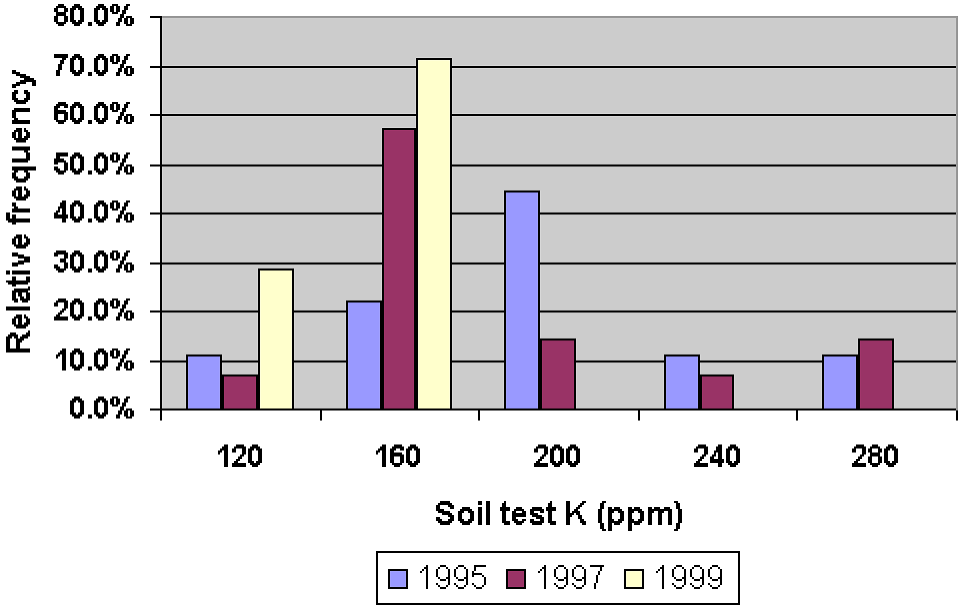
Chart Wizard - Step 4 of 4 - Chart Location

Place chart:

● As new sheet: Histograms

○ As object in: K

The Chart Wizard creates the graph below and places it in a new sheet labeled Histograms.



The graph contains 3 relative frequency distributions, one for each of the three years of soil test K data.

From this graph, we can easily see how the distribution in soil test levels has changed over time. In 1995 (the leftmost bar in each category), the greatest percentage of samples was in the 161-200 ppm class (labeled 200 on the x-axis). In 1996 (the middle bar in the first 2 classes and the right bar in the upper 3 classes), samples spanned all five classes, just as in 1995. However, most of the samples were located in the 121-160 ppm class, which is one class lower than in 1995. In 1999 (the rightmost bars in the lower 2 categories), samples spanned only the lower two categories. Like 1997, most of the samples were in the 121-160 ppm class. The remainder were in the next lowest class.

Notes

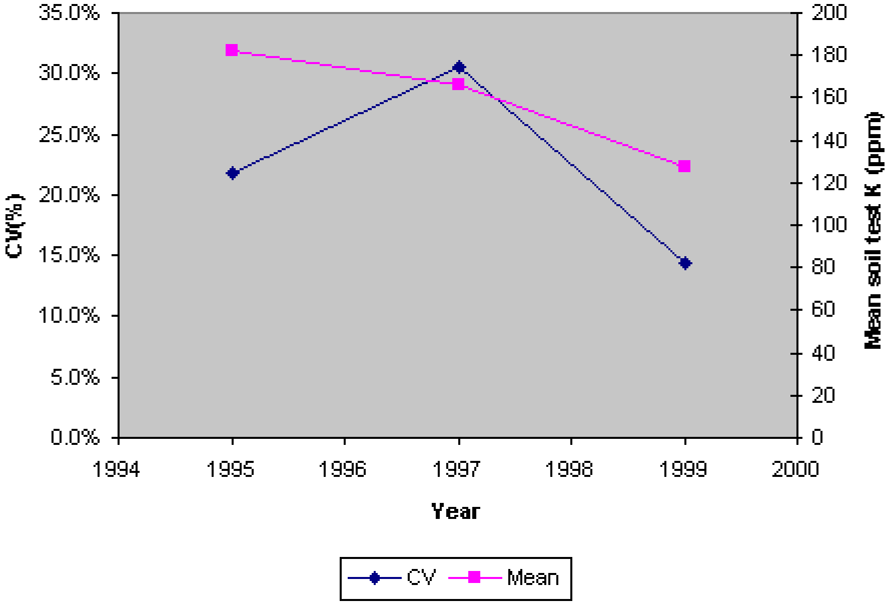
CHAPTER 4

Graphing Trends

Graphing descriptive statistics

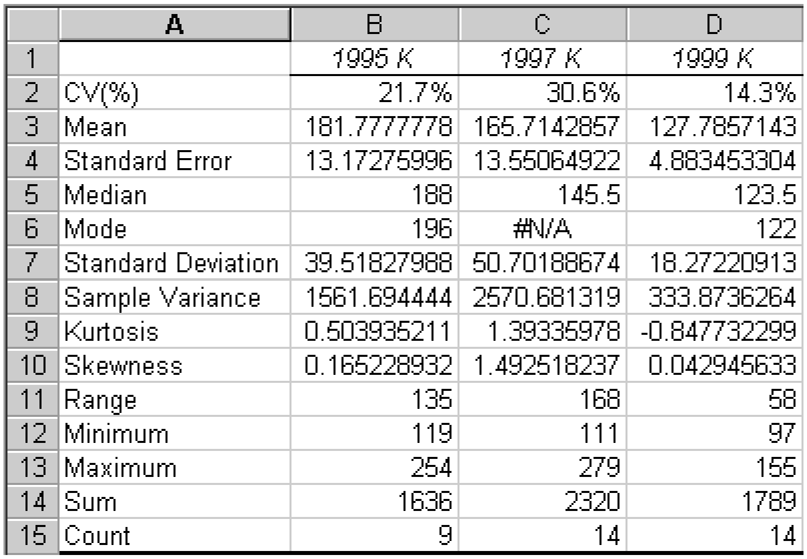
In this chapter, we will focus on graphing trends in two of our descriptive statistics: the mean and the coefficient of variation (CV). We want to put both on the same graph so we can tell at a glance what has been happening to: 1) the overall soil test level in the field, and 2) the variability in soil test levels. We will learn how to create the graph below.

In this chapter, we are not yet concerned about determining if these trends are statistically significant. We simply want to visualize them.



Graphing descriptive statistics

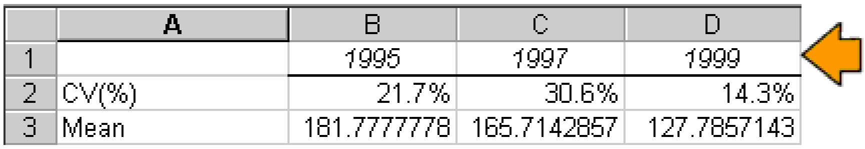
1. Start Microsoft Excel. On the Standard toolbar, Click File | Open and navigate to C:\PPIStat. Open file **Ex05**.



	A	B	C	D
1		*1995 K*	*1997 K*	*1999 K*
2	CV(%)	21.7%	30.6%	14.3%
3	Mean	181.7777778	165.7142857	127.7857143
4	Standard Error	13.17275996	13.55064922	4.883453304
5	Median	188	145.5	123.5
6	Mode	196	#N/A	122
7	Standard Deviation	39.51827988	50.70188674	18.27220913
8	Sample Variance	1561.694444	2570.681319	333.8736264
9	Kurtosis	0.503935211	1.39335978	-0.847732299
10	Skewness	0.165228932	1.492518237	0.042945633
11	Range	135	168	58
12	Minimum	119	111	97
13	Maximum	254	279	155
14	Sum	1636	2320	1789
15	Count	9	14	14

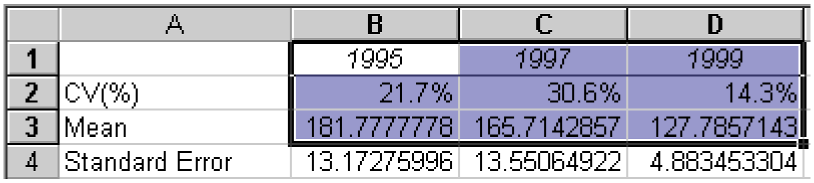
We want to graph the CV and the Mean over the 3 years (1995-9). We will have to first change the year column labels in row 1 to numbers, so we can plot the data.

2. Type **1995**, **1997**, and **1999** to cells **B1**, **C1**, and **D1**, respectively.



	A	B	C	D
1		*1995*	*1997*	*1999*
2	CV(%)	21.7%	30.6%	14.3%
3	Mean	181.7777778	165.7142857	127.7857143

3. Select cells **B1:D3**.



	A	B	C	D
1		*1995*	*1997*	*1999*
2	CV(%)	21.7%	30.6%	14.3%
3	Mean	181.7777778	165.7142857	127.7857143
4	Standard Error	13.17275996	13.55064922	4.883453304

4. Click on the **Chart Wizard** button, , on the Standard toolbar. In the Chart Wizard dialog box (Step 1 of 4 – Chart Type), select **XY (Scatter)** in the **Chart type** pane on the Standard Types tab. Select **Scatter with data points connected by lines** as the Chart sub-type. Click **Next**.

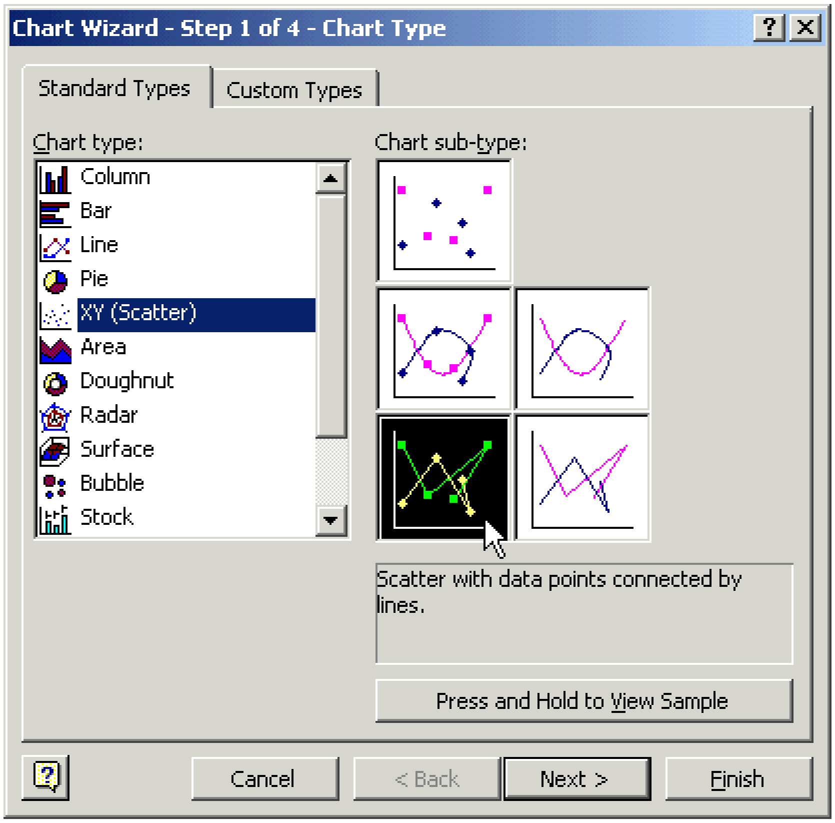
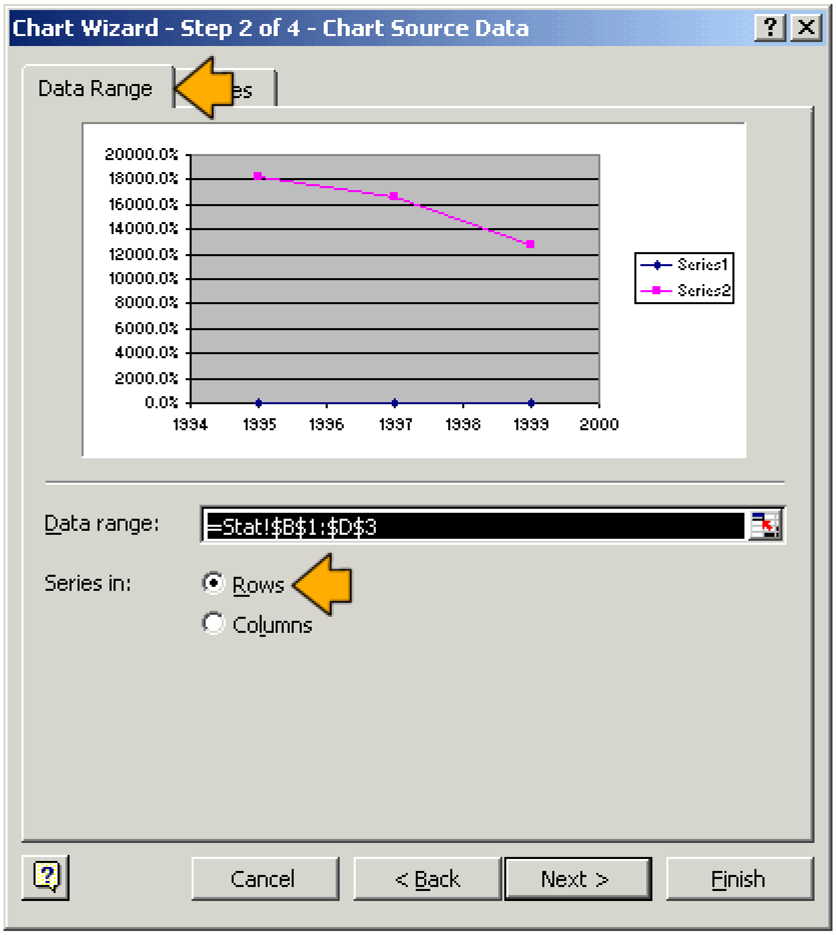


Chart Wizard - Step 1 of 4 - Chart Type

Chart type: XY (Scatter)

Scatter with data points connected by lines.

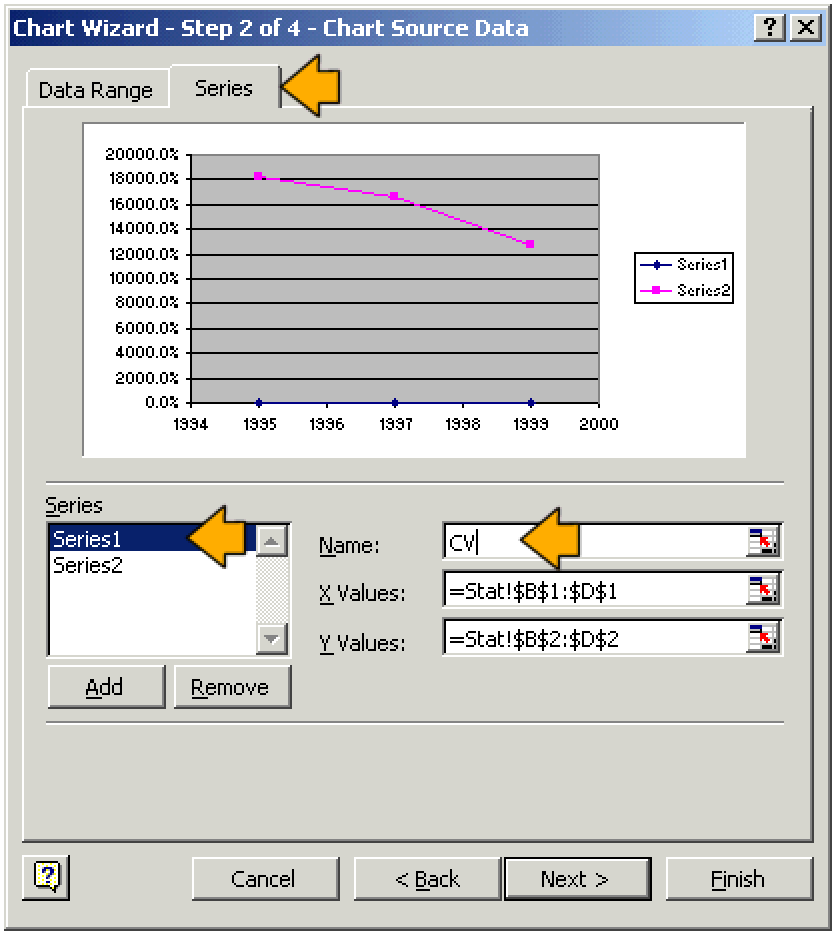
5. In the Chart Wizard dialog box (Step 2 of 4 – Chart Source Data), click on the **Data Range** tab. Select **Rows** for **Series in**.



Data range: =Stat!$B$1:$D$3

Series in: ● Rows ○ Columns

6. Click on the **Series** tab. Select **Series1** in the Series pane. Click on the **Name** field. Type **CV**.

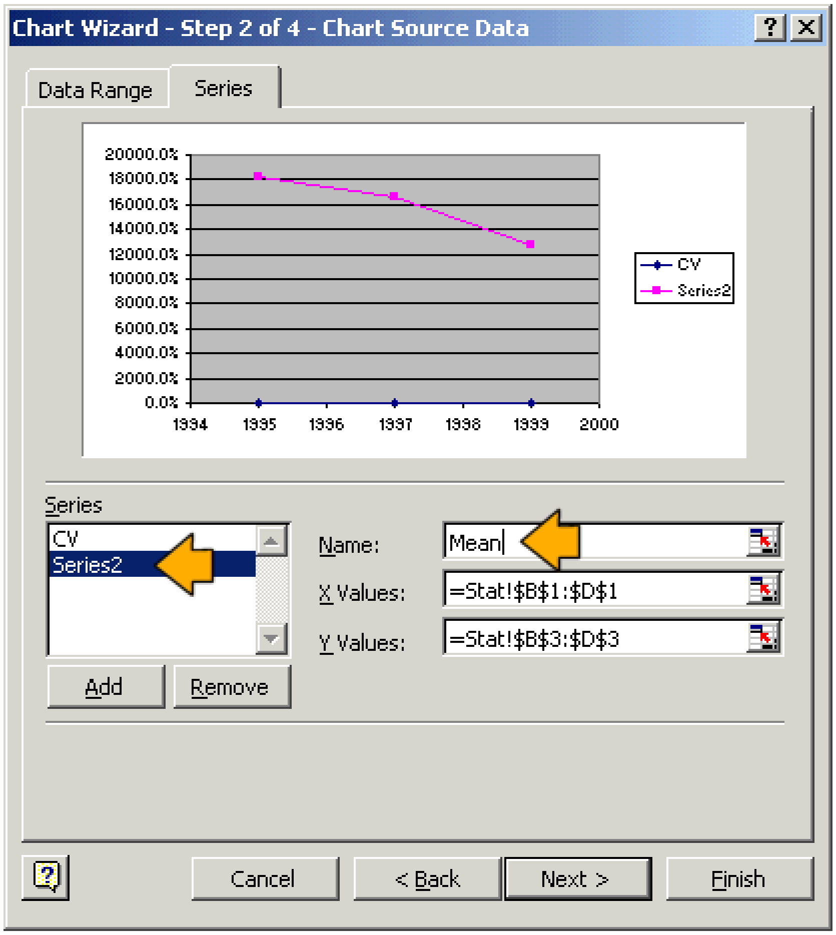


Name: CV

X Values: =Stat!$B$1:$D$1

Y Values: =Stat!$B$2:$D$2

7. Select **Series2** in the Series pane. Click on the **Name** field. Type **Mean**. Click **Next**.



Name: Mean

X Values: =Stat!$B$1:$D$1

Y Values: =Stat!$B$3:$D$3

8. In the Chart Wizard dialog box (Step 3 of 4 – Chart Options), click the **Titles** tab, type **Year** for **Value (X) axis** and **CV(%)** for **Value (Y)** axis.

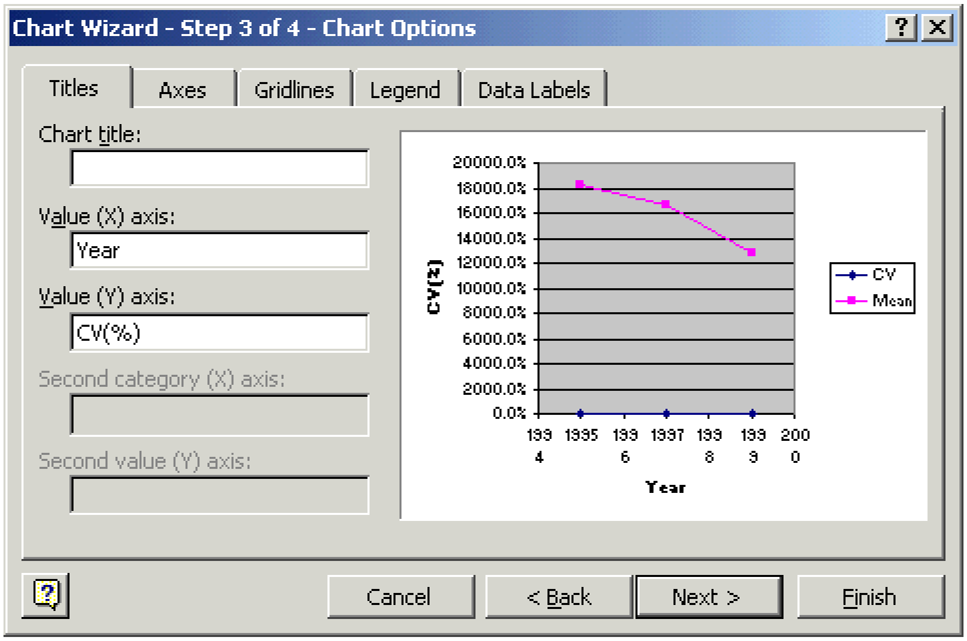


Chart Wizard - Step 3 of 4 - Chart Options

Chart title:

Value (X) axis: Year

Value (Y) axis: CV(%)

Second category (X) axis:

Second value (Y) axis:

9. Click on the **Gridlines** tab and uncheck **Major gridlines**.

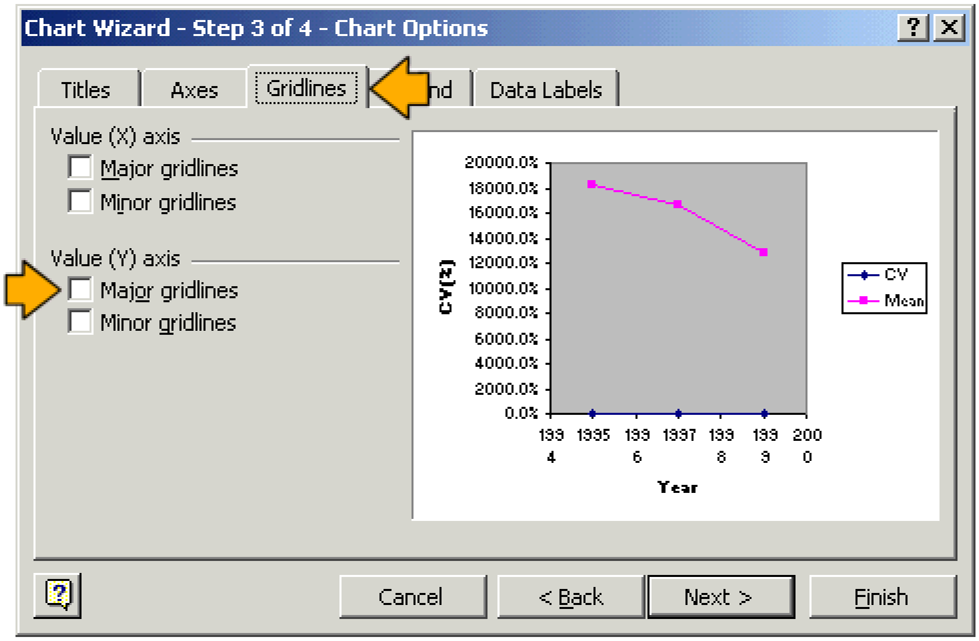


Chart Wizard - Step 3 of 4 - Chart Options

Value (X) axis

□ Major gridlines

□ Minor gridlines

Value (Y) axis

□ Major gridlines

□ Minor gridlines

10. Click on the **Legend** tab. Be sure the **Show legend** option is checked. Select the **Bottom** placement. Click **Next**.

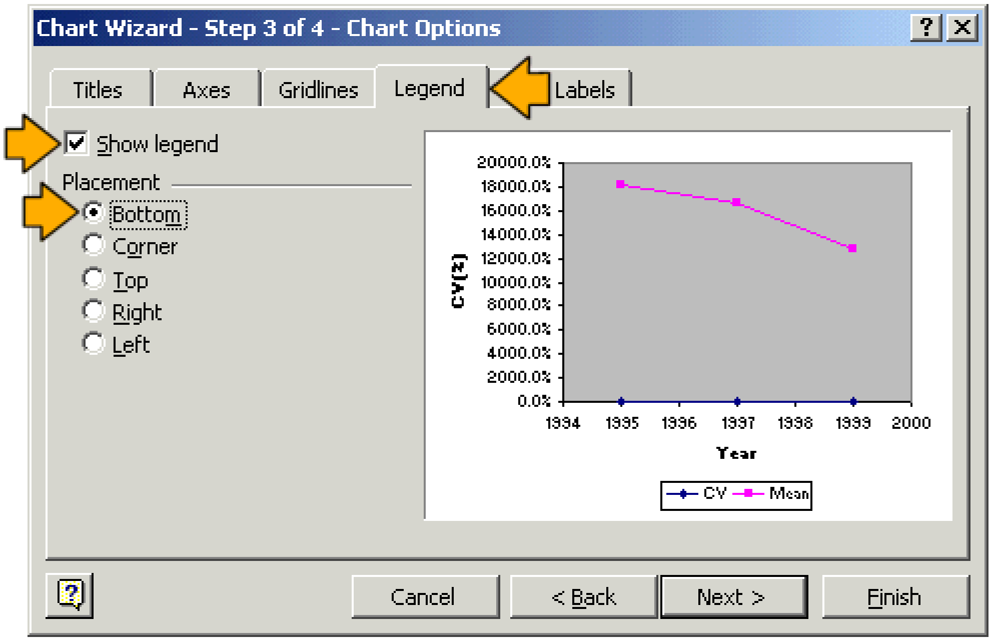


Chart Wizard - Step 3 of 4 - Chart Options

☑ Show legend

Placement

● Bottom
○ Corner
○ Top
○ Right
○ Left

11. In the Chart Wizard dialog box (Step 4 of 4 – Chart Location), select **As new sheet** and type **Trend** in the text box. Click **Finish**.

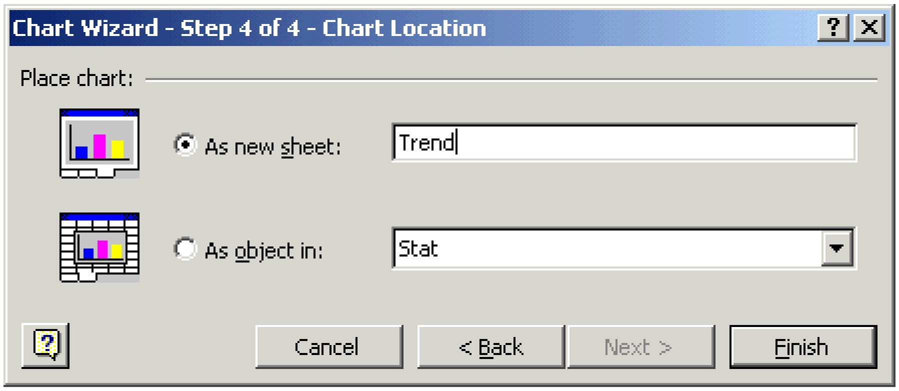


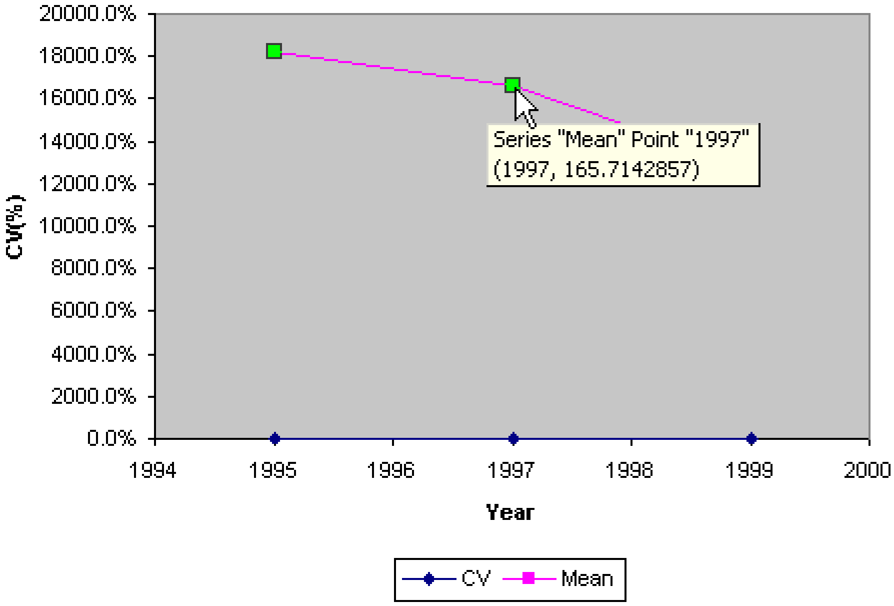
Chart Wizard - Step 4 of 4 - Chart Location

Place chart:

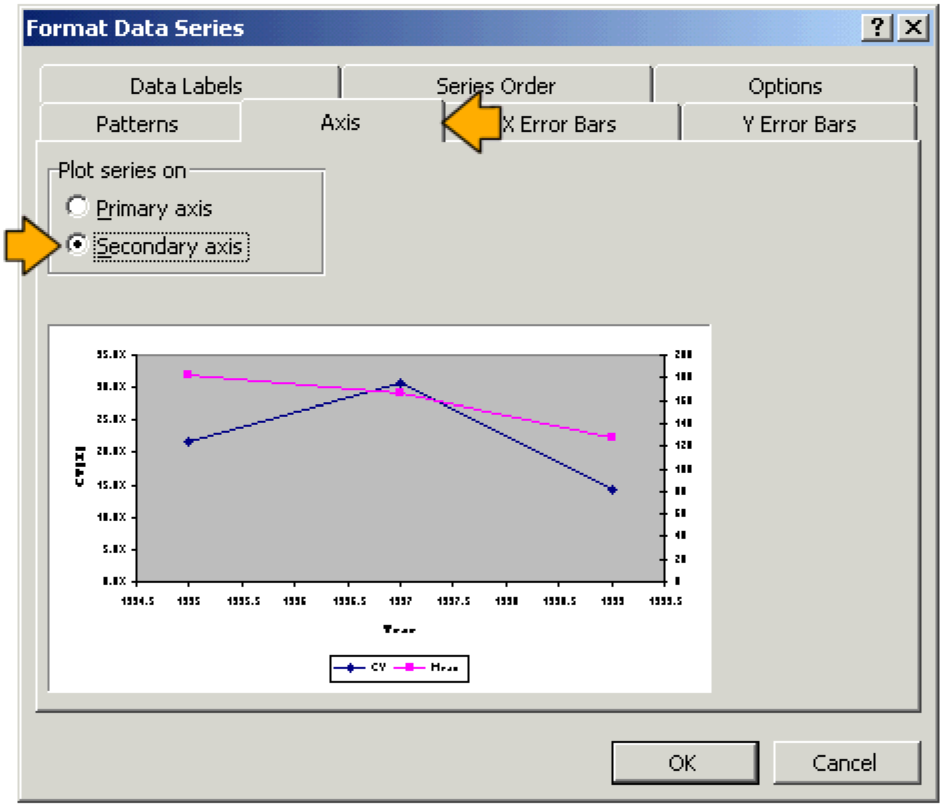
● As new sheet: Trend

○ As object in: Stat

12. On the newly created graph, **double click** on one of the points on the upper line (this is the line of the mean).



13. In the Format Data Series dialog box, click the **Axis** tab and select **Secondary axis** in the **Plot series on** pane. Click **OK**.

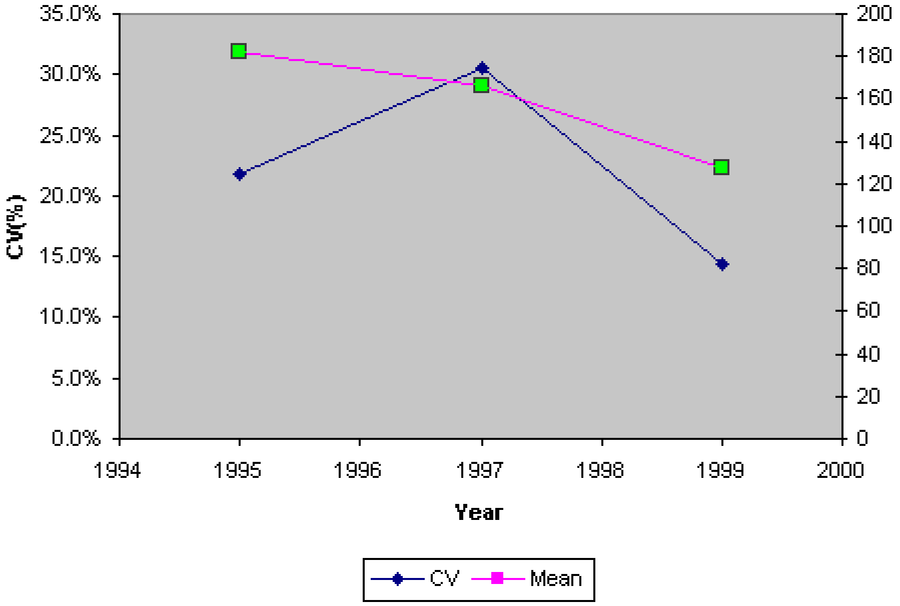

Format Data Series

Plot series on

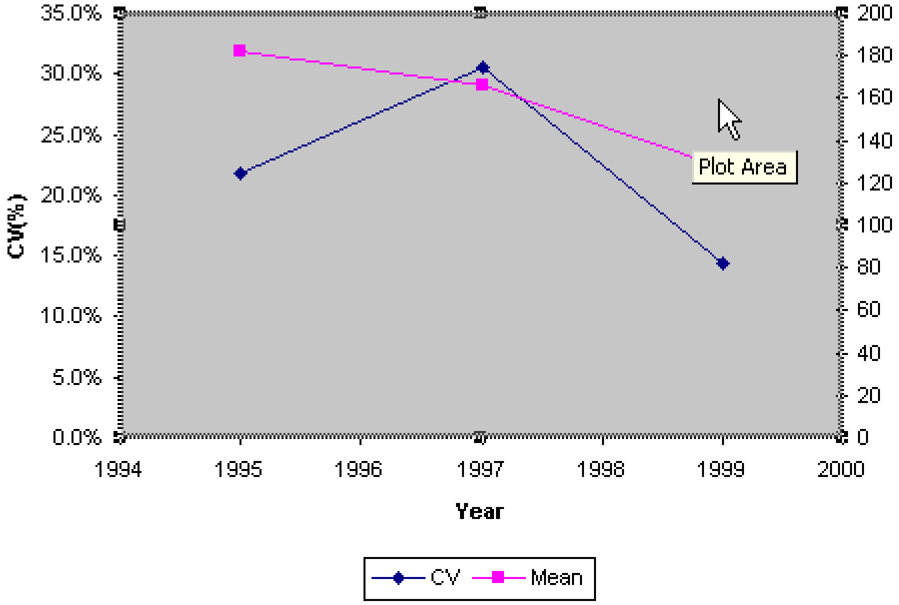
○ Primary axis

● Secondary axis

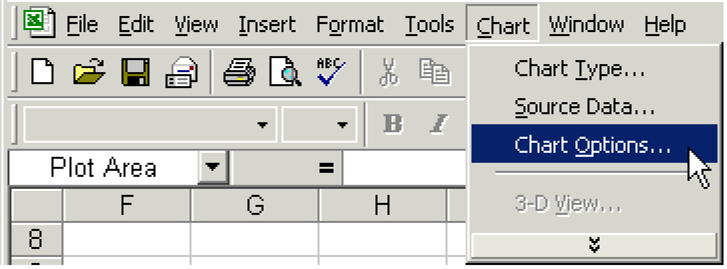
This creates a new y-axis on the right side of the graph, scaled to the mean soil test K levels. The y-axis on the left is now also properly scaled to the CV data.



14. Select the **Plot Area** by clicking inside the graph border.



15. On the Standard toolbar, click **Chart | Chart Options**.



16. In the Chart Options dialog box, click the **Titles** tab and type **Mean soil test K (ppm)** for the Second value (Y) axis.

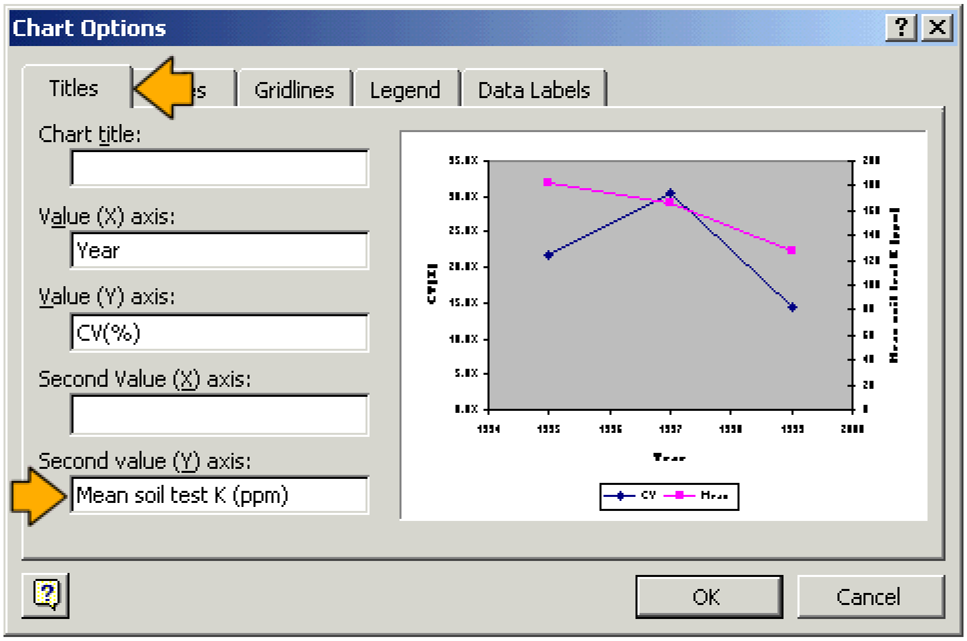


Chart title:

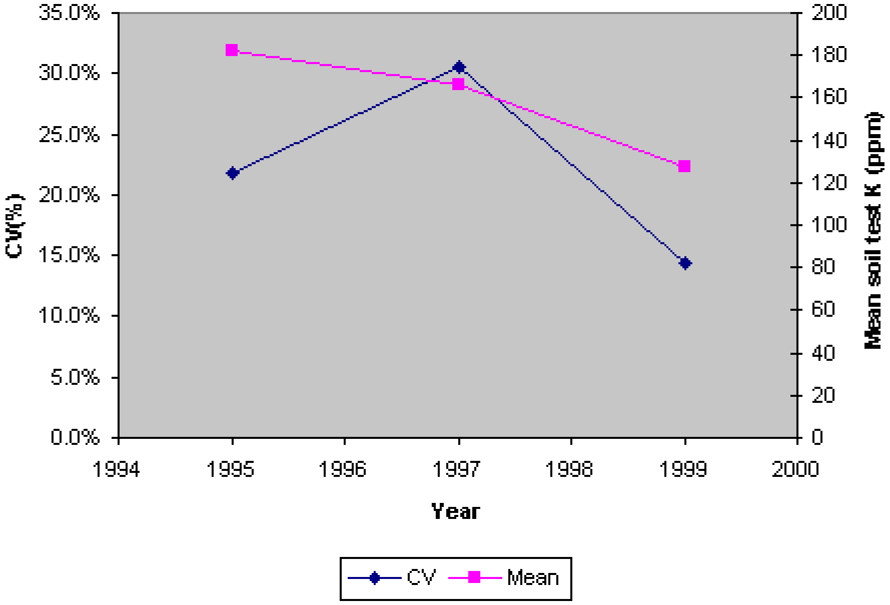
Value (X) axis: Year

Value (Y) axis: CV(%)

Second Value (X) axis:

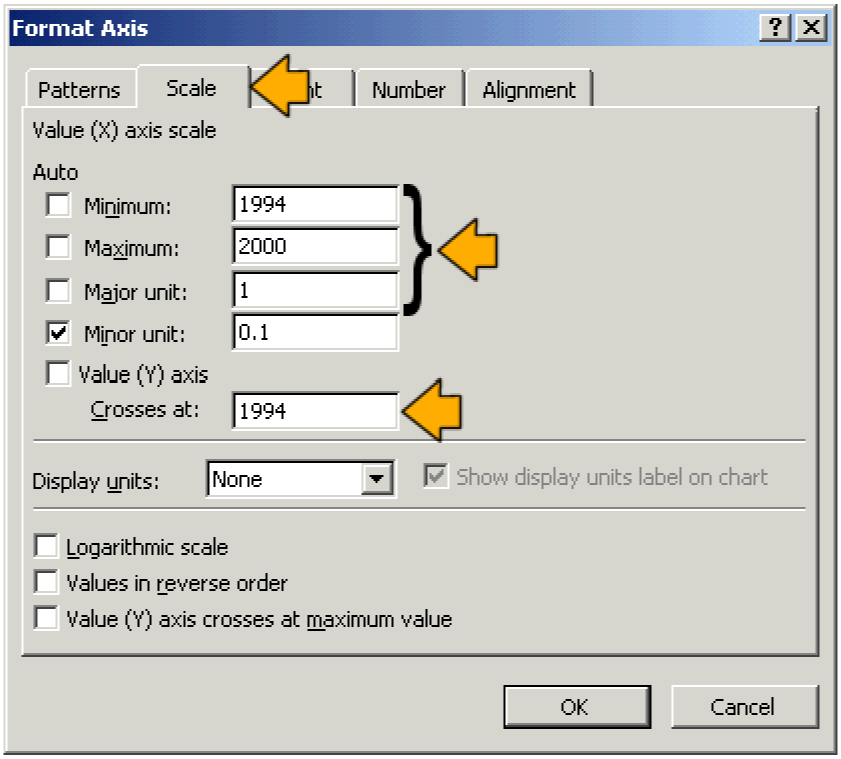
Second value (Y) axis: Mean soil test K (ppm)

We now have a graph that has labeled axes for both the mean soil test K levels and the CVs.



You may need to rescale the x-axis. Sometimes ticks on this axis are created for every 0.5 years, which doesn't make much sense for our example.

17. In the Format Axis dialog box, click on the **Scale** tab, type **1994** for the **Minimum**, **2000** for the **Maximum**, and **1** as the **Major unit**. Type **1994** in the **Value (Y) axis Crosses at** field. Click **OK**.



Value (X) axis scale

Auto

☐ Minimum: 1994

☐ Maximum: 2000

☐ Major unit: 1

☑ Minor unit: 0.1

☐ Value (Y) axis Crosses at: 1994

Display units: None ☑ Show display units label on chart

☐ Logarithmic scale

☐ Values in reverse order

☐ Value (Y) axis crosses at maximum value

This scales the x-axis from years 1994 to 2000 and keeps entire years as intervals on the axis.

Notes

Notes