# Chapter 2

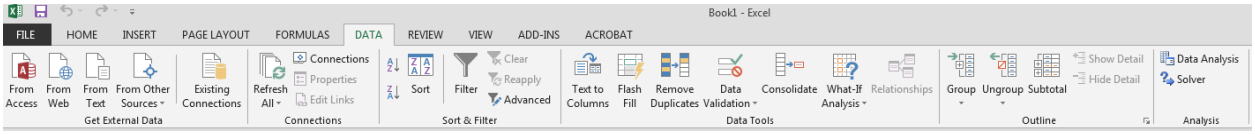
# Descriptive Statistics 1: Elementary Data Presentation and Description

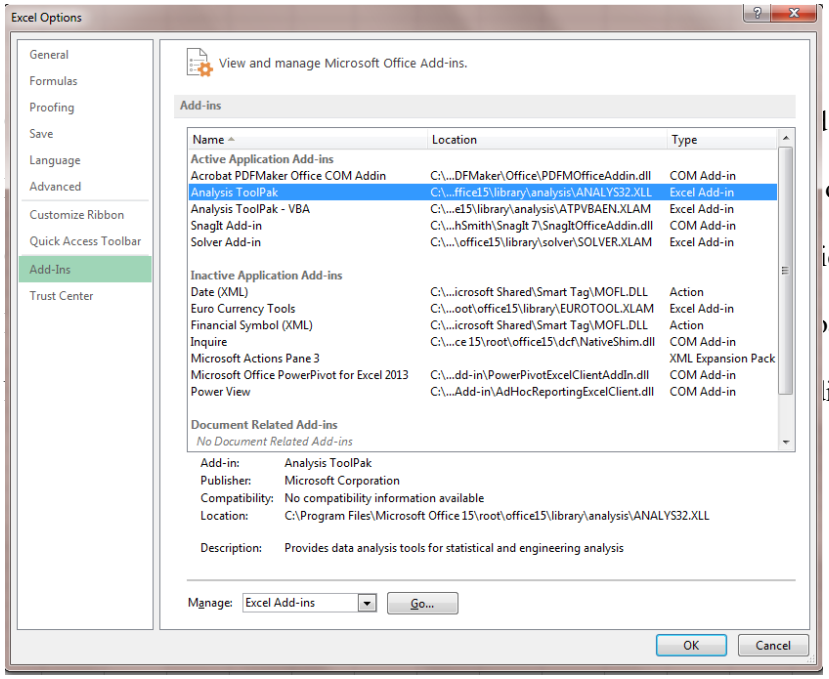
In this chapter we'll consider some of the most common descriptive measures for numerical data, beginning with measures of center (or *central tendency*) and measures of spread (or *dispersion*). We'll also examine ways of presenting data visually to communicate essential data set characteristics.

Graphic displays are essential to understand and interpret complex sets of data in order to be able to make business decisions easier. A first step in exploring and analyzing data is to reduce data to a graphic picture that is clear, concise and consistent with the message of the original data. In this chapter, guidelines are provided for selecting appropriate graphical representations for data sets.

## Data Analysis in Excel

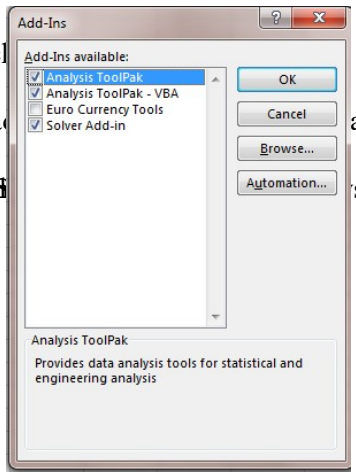
Many of the statistical techniques presented in this text can be performed in Excel using a tool called **Data Analysis**. To access this feature, select the **Data** tab along the top of an Excel worksheet. If the Data Analysis feature has been uploaded into your Excel package, it will be found in the **Analysis** section at the top right of the Data tab page. If **Data Analysis** does not appear in the Analysis section, it must be added.




**To add in Data Analysis:**

1. l in the drop down menu.
2. of the left menu.
3. ic
4. ain.
5. sis.



# 2.1 Measures of Central Location or Central Tendency

The measures defined here are **mean**, **median**, and **mode**.

**Demonstration Exercise 2.1**

The prime interest rate (%) at the close of each of the past twelve months was:

4.2, 4.0, 4.1, 5.2, 4.7, 5.4, 5.2, 5.8, 5.4, 5.2, 5.6, 5.2

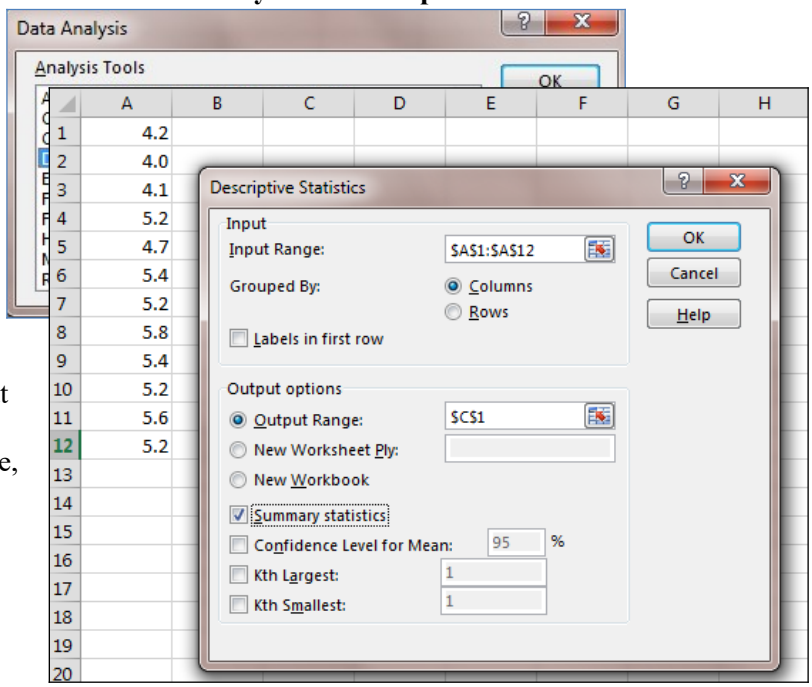
a) Determine the mean, the median and the mode for the interest rates. Treat the data as a population.
b) Interpret each of these three measures of central tendency.
c) Show the values on the number line and mark the location of the mean, the median and the mode. Show the mean as the balance point for the data.

1. Input the data into a column in Excel or open the Excel file **Demo Exercise 2-1** from the student website.

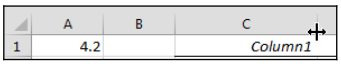
2. Determine the mean, median, and mode by using the **Data Analysis** capability in the **Analysis Toolpak** in Excel.

**Note on Analysis Toolpak:**
**Check to see if the Analysis Toolpak has been installed by selecting the Data tab to see if there is a Data Analysis button on the right side of the ribbon. If you do not see it, select** File ⇨ Options ⇨ Add-ins. At the bottom of the dialog box, you will see Manage Excel Add-ins. Click on the Go... button. Check the boxes for the Analysis Toolpak and the Solver Add-ins and OK. You should now see the Data Analysis and Solver options.

3. Select **Data ⇨ Data Analysis ⇨ Descriptive Statistics** and **OK.**




4. Input the data range, the cell for upper left of the output range, and select summary statistics and OK.

5. The resulting output is the summary statistics for this data set. Extend the column width (by double-clicking on the column label 

to see the labels for the output:

| *Column1* | |
|---|---|
| | |
| Mean | 5 |
| Standard Error | 0.174512 |
| Median | 5.2 |
| Mode | 5.2 |
| Standard Deviation | 0.604528 |
| Sample Variance | 0.365455 |
| Kurtosis | -0.88246 |
| Skewness | -0.68142 |
| Range | 1.8 |
| Minimum | 4 |
| Maximum | 5.8 |
| Sum | 60 |
| Count | 12 |

7. The mean of 5.0 represents the center value in the data set in the sense that it provides a balance point for the data. The median, 5.2, is the 50/50 marker—at least half the values (8 of 12 in this case) are at or above 5.2 and at least half the values (8 of 12 in this case) are at or below. The mode of 5.2 represents the most frequently occurring interest rate in the data set.

## 2.2 Measures of Dispersion

In data description, the mean, the median, or the mode give only a partial picture of a data set. It's often helpful, and sometimes essential, to accompany such measures of center with a measure of dispersion or variation. By reporting the **range** of the data— the difference between the smallest and the largest value in the data set— we've painted a much clearer picture of the values involved.

In contrast to the range, the **mean absolute deviation** (**MAD**) provides a much more comprehensive measure of dispersion. Specifically, the mean absolute deviation measures the average distance (or deviation) of the values in the data set from the data set mean.

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

Although the MAD is a straightforward, easily interpreted value, it's not the most frequently used measure of data dispersion in statistics. Much more common are the *variance* and the *standard deviation*— two very closely related descriptors of dispersion that possess more desirable properties than the MAD (as we'll see later in the text).

The calculation of **variance f**or a *population*:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

If we treat a *sample*, then both the calculation and the notation change slightly:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The **standard deviation**— the positive square root of the variance— is often used to report data dispersion.

For a population of values:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

If we treat the survey data as a sample, the standard deviation expression is:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

**Demonstration Exercise 2.2 Measures of Dispersion.**

ABC Products has 10 sales reps whose jobs involve a fair amount of overseas travel. The following set of values shows the number of overseas trips made during the past year for each of the reps:

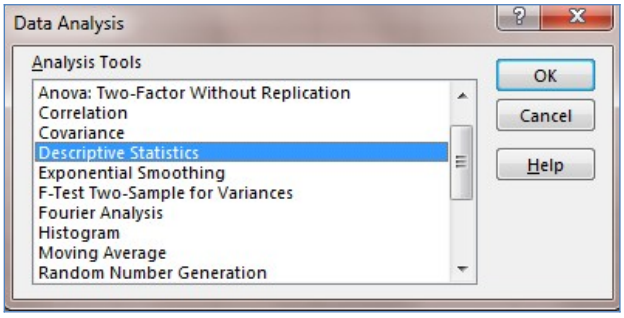
10, 30, 15, 35, 30, 40, 45, 10, 30, 5

Compute the range, the MAD, the variance, and the standard deviation for the data. Interpret each of these measures of dispersion. *Treat the data as a population*.

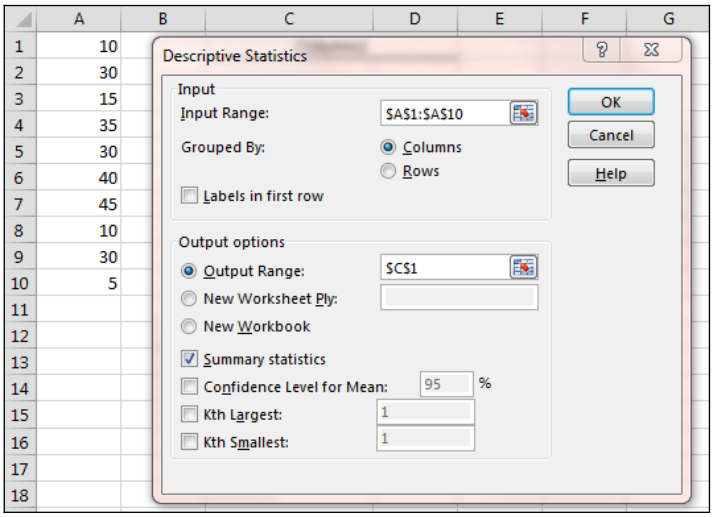
**Note on Analysis Toolpak:**
**Check to see if the Analysis Toolpak has been installed by selecting the Data tab to see if there is a Data Analysis button on the right side of the ribbon. If you do not see it, select** File ⇨ Options ⇨ Add-ins. At the bottom of the dialog box, you will see Manage Excel Add-ins. Click on the Go... button. Check the boxes for the Analysis Toolpak and the Solver Add-ins and OK. You should now see the Data Analysis and Solver options.

1. Input the data into a column in Excel.

2. Select **Data ⇨ Data Analysis ⇨ Descriptive Statistics** and **OK.**




3. Input the data range, the cell for the upper left of the output range, and select summary statistics and OK.



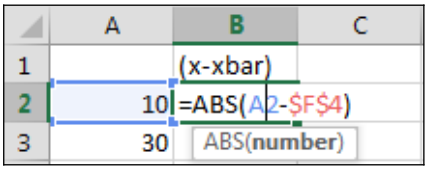

4. The output is shown as follows and shows the **Sample Variance**, the **Standard Deviation**, and the **Range**.

| *Column1* | |
|---|---|
| | |
| Mean | 25 |
| Standard Error | 4.409586 |
| Median | 30 |
| Mode | 30 |
| Standard Deviation | 13.94433 |
| Sample Variance | 194.4444 |
| Kurtosis | -1.48907 |
| Skewness | -0.15367 |
| Range | 40 |
| Minimum | 5 |
| Maximum | 45 |
| Sum | 250 |
| Count | 10 |

5. The **Mean** is 25 and the **Range** is 40. The **MAD** calculation is not included in the output. We need to calculate:

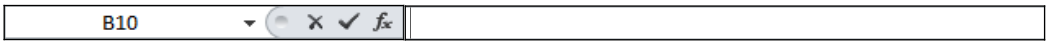
$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

6. The MAD calculation can be included in Excel by inserting the appropriate formula into the column to the right of the data. Input the absolute value of $(x - \bar{x})$ by inputting the formula: =ABS(A2-$E$4) into the first cell to the right of the first data point and Enter. A simple formula starts with an equal sign, telling Excel that a formula or function will follow. **ABS** is a function in Excel to take the absolute value of an expression.



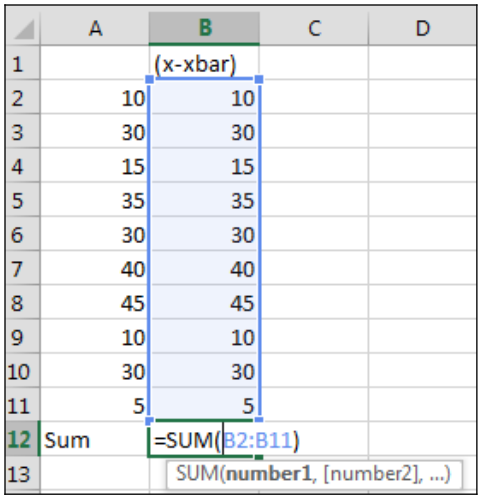

**INSERTING A FORMULA OR FUNCTION**

Functions can either be typed in directly into a cell or into the **Formula Bar**. The **Formula Bar** is located above the worksheet. The Formula Bar's unique features include access to the **Insert Function** dialog box when you click on the $f_x$ symbol. **Insert Function** is a built-in tool in Excel that assists with function choice and syntax.

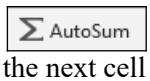



The **Cancel X** and the **Accept check mark** appear only when information is being input or edited into a cell.

The two values are input by selecting those cells by clicking on them. The reference to the mean is made absolute in this formula by typing $ before the row letter and $ before the column number (you can also click the function key **F4** when your cursor is inside the cell reference; every time you click F4, the absolute reference changes from the row reference to the column reference or both). The absolute reference means that when a formula is copied, the reference to that cell doesn't change, which is what happens normally when a formula is copied (called relative reference).

7. 


Copy the formula down by selecting the cell and dragging down the crosshair in the bottom right of the cell with the first formula.




Click on the next cell down and click the **AutoSum** button and Enter. This function selects the cells it thinks you want to sum. Verify that it is the correct range before Entering.

8. Divide the sum by *n* by inputting = and then click on the sum cell, then divide by (/) and typing the number of data points, in this case 10. The result is MAD, which = 12. *Interpretation*: The average difference between the number of trips made by each of the reps and the overall mean number of trips (25) is 12.

| | A | B |
|---|---|---|
| 1 | | (x-xbar) |
| 2 | 10 | 15 |
| 3 | 30 | 5 |
| 4 | 15 | 10 |
| 5 | 35 | 10 |
| 6 | 30 | 5 |
| 7 | 40 | 15 |
| 8 | 45 | 20 |
| 9 | 10 | 15 |
| 10 | 30 | 5 |
| 11 | 5 | 20 |
| 12 | Sum | 120 |
| 13 | Div N | =B12/10 |

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | (x-xbar) | | | |
| 2 | 10 | =ABS(A2-$E$4) | | *Column1* | |
| 3 | 30 | 5 | | | |
| 4 | 15 | 10 | | Mean | 25 |
| 5 | 35 | 10 | | Standard Error | 4.409586 |
| 6 | 30 | 5 | | Median | 30 |
| 7 | 40 | 15 | | Mode | 30 |
| 8 | 45 | 20 | | Standard Deviation | 13.94433 |
| 9 | 10 | 15 | | Sample Variance | 194.4444 |
| 10 | 30 | 5 | | Kurtosis | -1.48907 |
| 11 | 5 | 20 | | Skewness | -0.15367 |
| 12 | Sum | 120 | | Range | 40 |
| 13 | Div N | 12 | | Minimum | 5 |
| 14 | | | | Maximum | 45 |
| 15 | | | | Sum | 250 |
| 16 | | | | Count | 10 |

9. The variance and the standard deviation are supposed to be treated as a population. The Excel output treats the sample data as a sample. We will have to calculate these values associated with the population.

10. Starting with the variance, $(x-\bar{x})^2$. Input the formula value to the left, type ^ 2. This indicates the square of as done in the previous steps

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

| | A | B | C |
|---|---|---|---|
| 1 | | (x-xbar) | (x-xbar)^2 |
| 2 | 10 | 15 | =B2^2 |
| 3 | 30 | 5 | 25 |
| 4 | 15 | 10 | 100 |
| 5 | 35 | 10 | 100 |
| 6 | 30 | 5 | 25 |
| 7 | 40 | 15 | 225 |
| 8 | 45 | 20 | 400 |
| 9 | 10 | 15 | 225 |
| 10 | 30 | 5 | 25 |
| 11 | 5 | 20 | 400 |
| 12 | Sum | 120 | 1750 |
| 13 | Div N | 12 | 175 |

insert a column to calculate

starting with =, click on the cell (indicates an exponent) and input the cell value. Sum those values and divide by 10 (the number of data points).

11. The **variance** is calculated as 175. *Interpretation*: The average *squared* difference between the number of trips made by each of the reps and the mean number of trips (25) is 175.

12. The **standard deviation** of the population is the square root of the variance we calculated in the previous steps. Input = and then SQRT( and then click on the cell above that calculated the variance. SQRT is another function in Excel. The value of the population standard deviation is 13.2. The value of the cell can be formatted by changing the number of decimal places using the toolbar buttons:

*Interpretation*: Roughly speaking, the number of trips made by each of the sales reps is, on average, about 13.2 trips away from the overall mean of 25 trips. As is typically the case, the standard deviation of 13.2 is greater than the MAD, which here is 12..

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

| | A | B | C |
|---|---|---|---|
| 1 | | (x-xbar) | (x-xbar)^2 |
| 2 | 10 | 15 | 225 |
| 3 | 30 | 5 | 25 |
| 4 | 15 | 10 | 100 |
| 5 | 35 | 10 | 100 |
| 6 | 30 | 5 | 25 |
| 7 | 40 | 15 | 225 |
| 8 | 45 | 20 | 400 |
| 9 | 10 | 15 | 225 |
| 10 | 30 | 5 | 25 |
| 11 | 5 | 20 | 400 |
| 12 | Sum | 120 | 1750 |
| 13 | Div N | 12 | 175 |
| 14 | St Dev | | =SQRT(C13) |

| | A | B | C |
|---|---|---|---|
| 1 | | (x-xbar) | (x-xbar)^2 |
| 2 | 10 | 15 | 225 |
| 3 | 30 | 5 | 25 |
| 4 | 15 | 10 | 100 |
| 5 | 35 | 10 | 100 |
| 6 | 30 | 5 | 25 |
| 7 | 40 | 15 | 225 |
| 8 | 45 | 20 | 400 |
| 9 | 10 | 15 | 225 |
| 10 | 30 | 5 | 25 |
| 11 | 5 | 20 | 400 |
| 12 | Sum | 120 | 1750 |
| 13 | Div N | 12 | 175 |
| 14 | St Dev | | 13.2 |

## 2.3 Frequency Distributions

It's often useful to present data in a *partially* summarized form that makes it easy to see important data set features. One possibility is to display the data as a **frequency distribution**. Here we'll simply identify the unique value possibilities for members of the data set and count the number of times that each of these values appears, showing results in a simple table format.

**Demonstration Exercise 2.3: Frequency Distributions**.

Klobes International Construction is a prime contractor for major construction projects in Europe and South America. Recent labor problems in both regions have begun to cause delays of from one to two weeks for 18 current projects. Below is a table showing the estimated delay in days for each project:

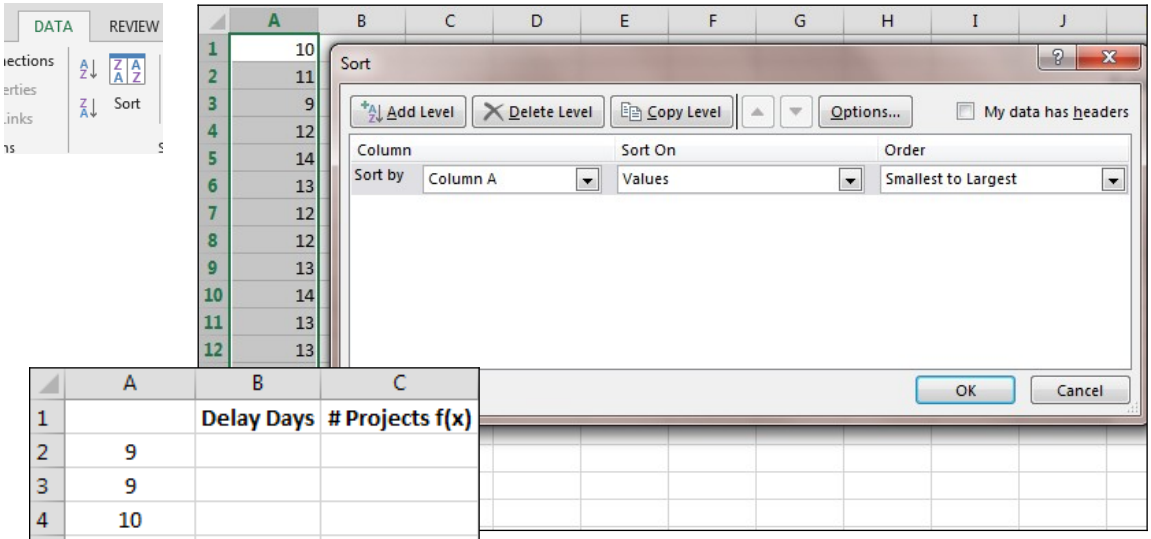
**Estimated Projects Delays (days)**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 10 | 11 | 9 | 12 | 14 | 13 | 12 | 12 | 13 |
| 14 | 13 | 13 | 10 | 9 | 12 | 11 | 13 | 11 |

a) Show the data in a frequency table and construct the corresponding bar chart.
b) Show the frequency polygon for the data.
c) Using the vocabulary of the previous section, describe the shape of the distribution.

1. Input the data into a column in Excel.

2. Determine the possible values for project delay time. It would be helpful to sort the data to make it easier to locate those values. You can sort the data by clicking on the first cell of data in a column and selecting **Data** ⇨ **Sort** from the ribbon. The data cells in that column will automatically be selected and the following dialog box will appear. Click OK and the data will be sorted smallest to largest.



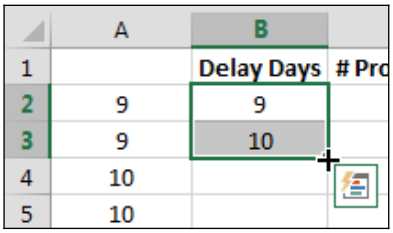

| | A | B | C |
|---|---|---|---|
| 1 | | Delay Days | # Projects f(x) |
| 2 | 9 | | |
| 3 | 9 | | |
| 4 | 10 | | |
| 5 | 10 | | |
| 6 | 11 | | |
| 7 | 11 | | |
| 8 | 11 | | |
| 9 | 12 | | |
| 10 | 12 | | |
| 11 | 12 | | |
| 12 | 12 | | |
| 13 | 13 | | |
| 14 | 13 | | |
| 15 | 13 | | |
| 16 | 13 | | |
| 17 | 13 | | |
| 18 | 14 | | |
| 19 | 14 | | |

3. The possible values for project delay time are 9 through 14 days.

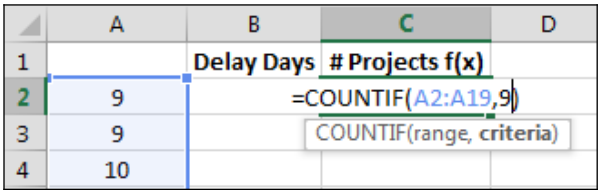
4. Counting the number of projects associated with each delay time gives the frequency values. This data set could be analyzed manually but the method described will be helpful for larger data sets. Insert a row at the top of your data set by rt-clicking on the number 1 button on left and select Insert. Input the titles as follows:

5. Counting the number of projects associated with each delay time gives the frequency values. This data set could be analyzed manually but the method described will be helpful for larger data sets. Insert 9 as the first number under Delay Days.

6. The Delay Day number is 10. If you want to fill in a range of cells with a specific number or a series of numbers, type and enter the data to be repeated. Select the cell(s) to be copied and then move the cursor to the lower right of the cell until the mouse pointer turns to a cross hair. Drag in whatever direction you want to copy the information. If an extended series is desired, type the first two numbers or time increments in the series, then select those two cells and drag down the lower right crosshair until the series is completed, in this case 14.




7. To count the cells with a particular value, you can use the COUNTIF function, syntax shown below. The range of the data set is selected and then the criteria defined, in this case 9 and Enter. The count of 9's is shown as 2.



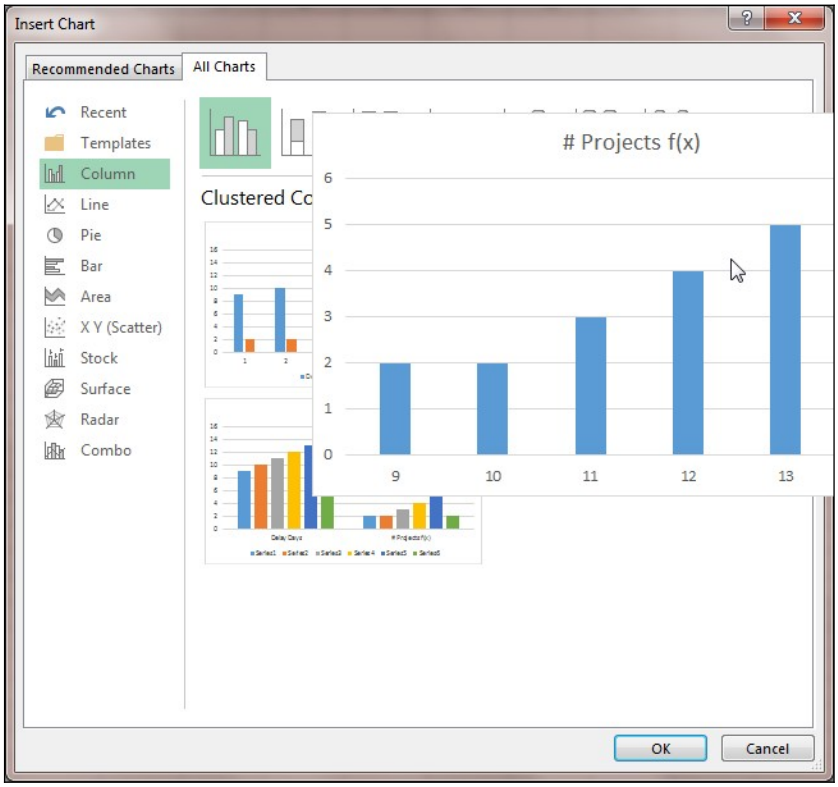

8. The process of constructing this table is easier overall if the function can be copied with accurate results. The range of cells would have to be made absolute in order for that reference not be changed when the function is copied. Also, instead of typing 9, you can click on the appropriate cell next to the function. The function will look like the following:

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | Delay Days | # Projects f(x) | |
| 2 | 9 | =COUNTIF($A$2:$A$19,B2) | | |
| 3 | 9 | 10 | COUNTIF(range, criteria) | |
| 4 | 10 | 11 | 3 | |
| 5 | 10 | 12 | 4 | |
| 6 | 11 | 13 | 5 | |
| 7 | 11 | 14 | 2 | |
| 8 | 11 | | | |
| 9 | 12 | | | |
| 10 | 12 | | | |

9. Copy the function down. The frequency table looks like the following:

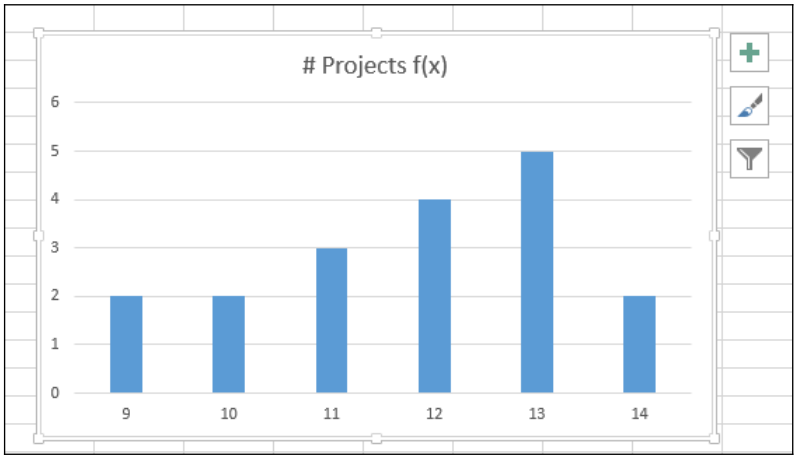
| B | C |
|---|---|
| **Delay Days** | **# Projects f(x)** |
| 9 | 2 |
| 10 | 2 |
| 11 | 3 |
| 12 | 4 |
| 13 | 5 |
| 14 | 2 |

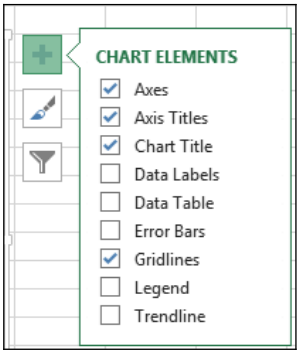
10. Display the frequency table as a bar chart. Select the **Delay Days** title and data and **# Projects f(x)** title and data. Select **Insert** and under the Chart options, select **2-D Column** chart and then select **More Column Charts**.

11. 


Select the column chart that best displays the data which is the one shown below:

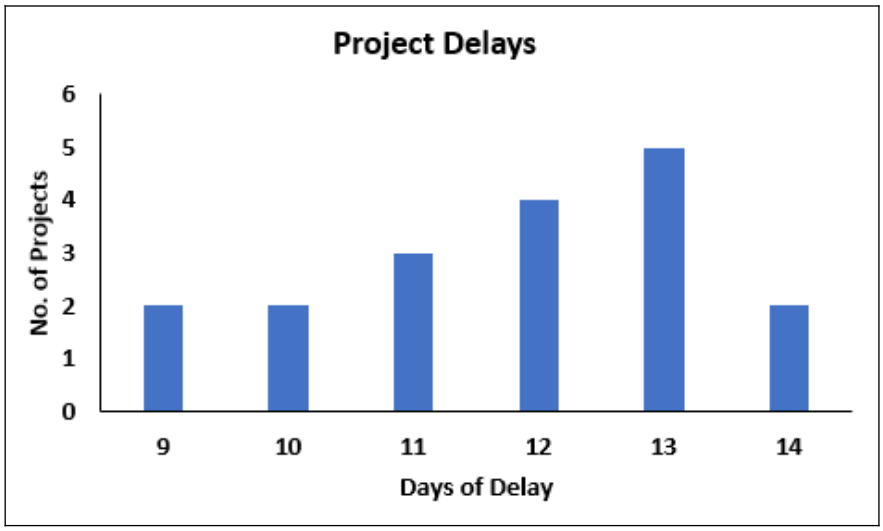
12. Select the column chart that best displays the data which is the one shown below. Click OK and the chart is inserted into the worksheet. The chart can then be formatted and labeled.



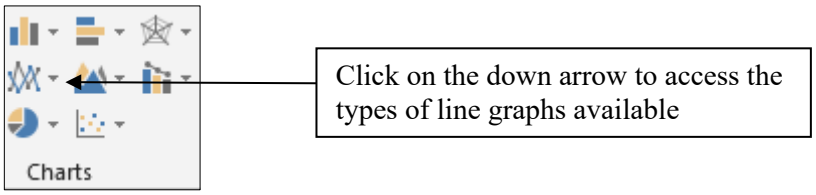




13. Click on the chart to select it and select the plus button on the right side. Select **Axis Titles**.

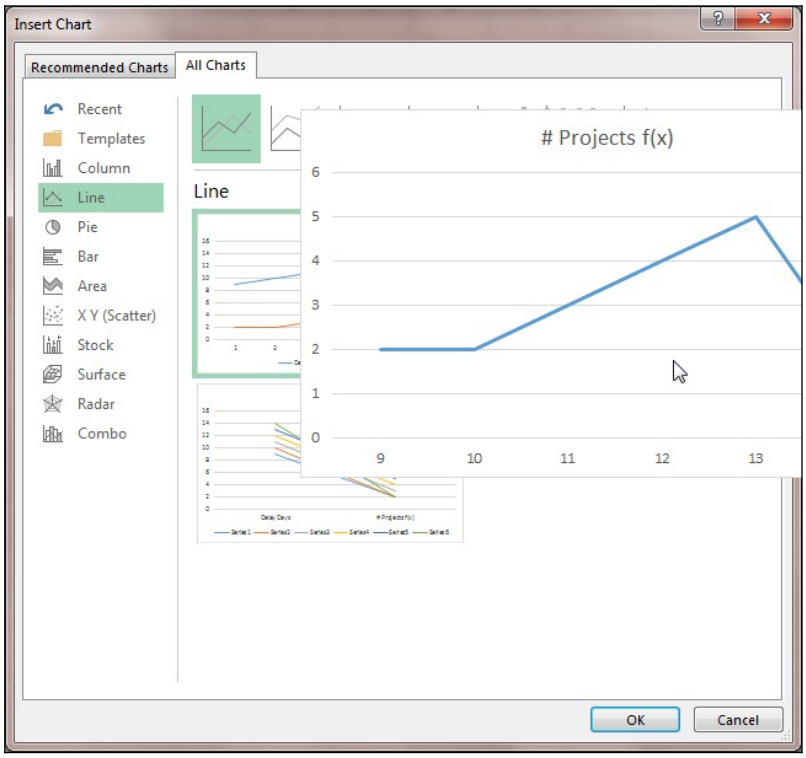
14. Text boxes now appear that you can click inside and edit. You can edit the chart title, right click on the gridlines and delete. The x-axis label should be input as **Days of Delay**, the y-axis label input as **No. of Projects**, and the chart title edited to be **Project Delays**. Colors, text, and other aspects of the chart can be changed (see Chapter 1 for additional details). The finished chart should look like the following:

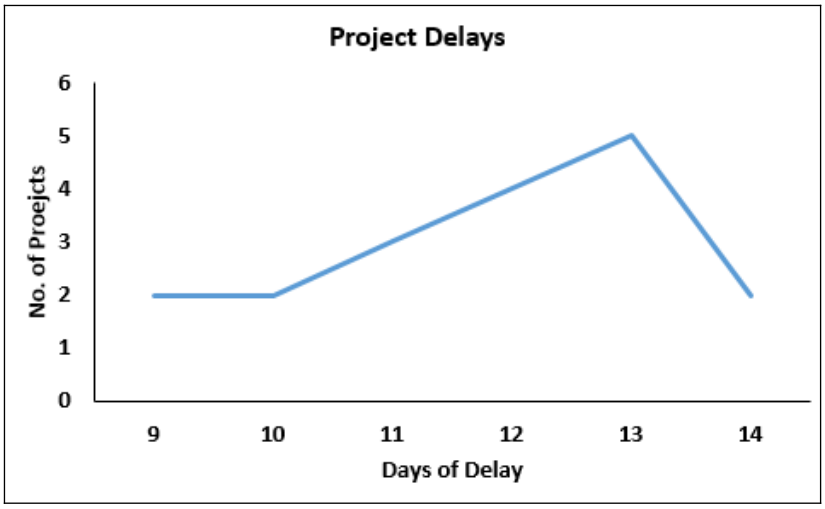



15. The **frequency polygon** is constructed by joining the midpoints of the columns with a line. For this example, we can use the given data for Days of Delay. In other types of data sets, we may need to create the midpoint of a bin. In order to create this chart in Excel, select the same set of data as you did for the column chart. Then click on the **Insert** tab, then click on the down arrow to the right of **Line** (graph) and then **More Line Charts....**




16. You will see the choices of line charts available. Select the chart that shows only one line with *Midpoints* on the *x*-axis and *Frequency* for the *y*-axis.
Click **OK**.
The resulting graph will need to be resized and labels and title added in the same method as the column chart.






**Demonstration Exercise 2.4: Descriptive Measures for Frequency Distributions**

The frequency table below shows ADC's 30-year fixed rate for home mortgages over the past 30 days:

| Rate(%) *x* | No. of days *f(x)* |
|---|---|
| 5.2 | 3 |
| 5.3 | 6 |
| 5.4 | 12 |
| 5.5 | 6 |
| 5.6 | 3 |

a) Compute the mean rate for this 30-day period.
b) What is the median mortgage rate for this 30-day period?
c) Compute the variance and the standard deviation for the rates. Treat the data as a population.

1. Input the data into Excel as shown.

2. When data are given to us in the format of a frequency table, we will use the following formula to compute the mean rate.

$$\mu = \frac{\sum x \cdot f(x)}{N}$$

3. In the column to the right of frequency, we are going to sum the product of *x* and *f(x)*. The first cell will be a simple formula starting with = , click on the first cell for *x*, type * to multiply, and click on the first cell for *f(x)*. The formula results will look like the following:

| | A | B | C |
|---|---|---|---|
| 1 | Rate(%) x | No. of days f(x) | x*f(x) |
| 2 | 5.2 | 3 | =A2*B2 |
| 3 | 5.3 | 6 | |
| 4 | 5.4 | 12 | |
| 5 | 5.5 | 6 | |
| 6 | 5.6 | 3 | |

4. Copy the formula down and use the Autosum button to add the column of calculations.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Rate(%) x | No. of days f(x) | x*f(x) | | |
| 2 | 5.2 | 3 | 15.6 | | |
| 3 | 5.3 | 6 | 31.8 | | |
| 4 | 5.4 | 12 | 64.8 | | |
| 5 | 5.5 | 6 | 33 | | |
| 6 | 5.6 | 3 | 16.8 | | |
| 7 | | | =SUM(C2:C6) | | |
| 8 | | | SUM(**number1**, [number2], ...) | | |

5. Find the sum of the number of days.

6. The last step to calculating the mean is to input a formula equal to the sum of $x*f(x)$ divided by N. The result is 5.4.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Rate(%) x | No. of days f(x) | x*f(x) | |
| 2 | 5.2 | 3 | 15.6 | |
| 3 | 5.3 | 6 | 31.8 | |
| 4 | 5.4 | 12 | 64.8 | |
| 5 | 5.5 | 6 | 33 | |
| 6 | 5.6 | 3 | 16.8 | |
| 7 | | 30 | 162 | Sum |
| 8 | | | =C7/B7 | Mean |

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Rate(%) x | No. of days f(x) | x*f(x) | |
| 2 | 5.2 | 3 | 15.6 | |
| 3 | 5.3 | 6 | 31.8 | |
| 4 | 5.4 | 12 | 64.8 | |
| 5 | 5.5 | 6 | 33 | |
| 6 | 5.6 | 3 | 16.8 | |
| 7 | | 30 | 162 | Sum |
| 8 | | | 5.4 | Mean |

7. In a set of 30 values, the median is halfway between the 15th and the 16th values [(30+1)/2 = 15.5]. This would mean that the median is 5.4 (Start by counting down the right hand (frequency) column until the frequency total is 15. At that point, you should be able to see from the table that the 15th and the 16th values are both 5.4.).

8. The variance for frequency table is computed by the following formula:

$$\sigma^2 = \frac{\sum (x-\mu)^2 \cdot f(x)}{N}$$

9. In the column to the right of $x*f(x)$, we are going to sum $(x - \mu)^2 * f(x)$ for each row. The first cell will be a simple formula starting with = , click on the first cell for $x$, type - to subtract, and click on the calculated $\mu$ and make that reference absolute The formula results will look like the following:

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Rate(%) x | No. of days f(x) | x*f(x) | | $(x - \mu)^2 * f(x)$ |
| 2 | 5.2 | 3 | 15.6 | | =(A2-$C$8)^2*B2 |
| 3 | 5.3 | 6 | 31.8 | | |
| 4 | 5.4 | 12 | 64.8 | | |
| 5 | 5.5 | 6 | 33 | | |
| 6 | 5.6 | 3 | 16.8 | | |
| 7 | | 30 | 162 | Sum | |
| 8 | | | 5.4 | Mean | |

10. The resulting value for the variance is 0.012 and the square root of that value is the standard deviation which is =SQRT(0.012) = 0.11.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Rate(%) x | No. of days f(x) | x*f(x) | | $(x - \mu)^2 * f(x)$ |
| 2 | 5.2 | 3 | 15.6 | | 0.12 |
| 3 | 5.3 | 6 | 31.8 | | 0.06 |
| 4 | 5.4 | 12 | 64.8 | | 0 |
| 5 | 5.5 | 6 | 33 | | 0.06 |
| 6 | 5.6 | 3 | 16.8 | | 0.12 |
| 7 | | 30 | 162 | Sum | 0.36 |
| 8 | | | 5.4 | Mean | 0.012 |

## 2.4 Frequency Distributions

A *relative* frequency table offers an alternative to the frequency table as a way of presenting data in partially summarized form. Here, rather than reporting the number of data set members having the value 1 or 2 or 3, etc., we'll report the percentage or the *proportion* of members having each of the values. A relative count — we'll label it P($x$) is substituted for the absolute count, $f(x)$, to produce the **relative frequency distribution**.

**Demonstration Exercise 2.5: Relative Frequency Distributions**

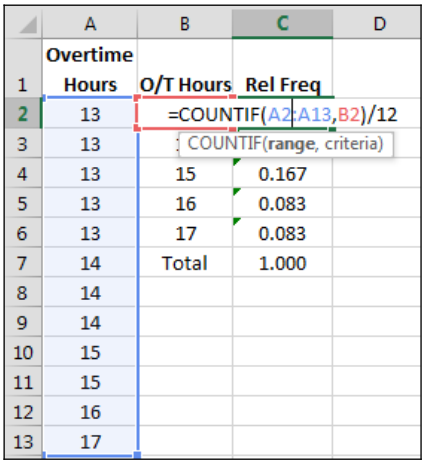
Overtime hours during the past week for the 12 staff members in the Human Resources Office at Palmer Software were:

13 14 13 13 15 14 15 16 13 17 13 14

a) Show the data in relative frequency table
b) Show the data in a relative frequency bar chart.

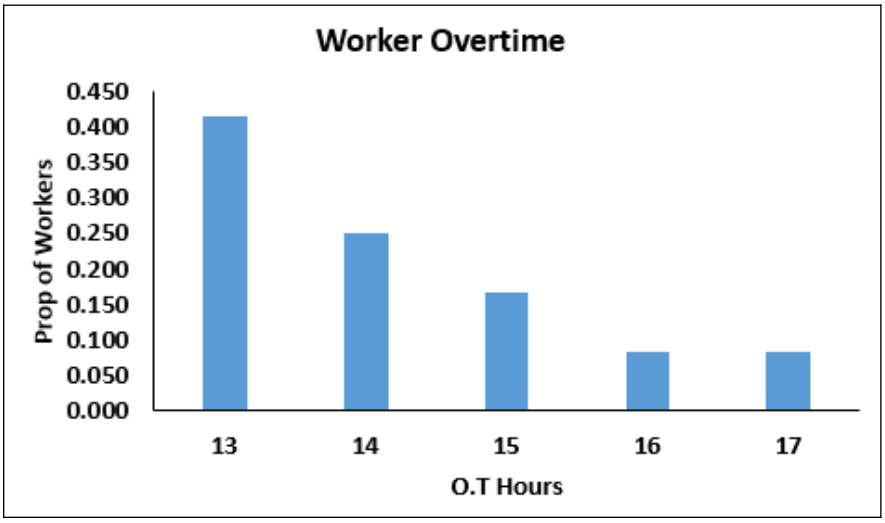
1. Input the data into a column in Excel.

2. Follow the method in Exercise 2.3 for sorting the data and counting the frequencies. First sort the data and then count the frequencies. Then divide by the total number 12 to calculate the relative frequency.



3. The relative frequency table and its 2D column chart:

| | A | B | C |
|---|---|---|---|
| 1 | Overtime Hours | O/T Hours | Rel Freq |
| 2 | 13 | 13 | 0.417 |
| 3 | 13 | 14 | 0.250 |
| 4 | 13 | 15 | 0.167 |
| 5 | 13 | 16 | 0.083 |
| 6 | 13 | 17 | 0.083 |
| 7 | 14 | Total | 1.000 |
| 8 | 14 | | |
| 9 | 14 | | |
| 10 | 15 | | |
| 11 | 15 | | |
| 12 | 16 | | |
| 13 | 17 | | |




**Demonstration Exercise 2.6: Descriptive Measures for Relative Frequency Distributions**

Daily absences for employees at GHT Inc. are reported below:

a) Compute the mean number of employees absent per day.
b) Determine the median number of absences.
c) Compute the variance and standard deviation for the daily absence data.

| *x* | P(x) |
|---|---|
| 0 | 0.12 |
| 1 | 0.18 |
| 2 | 0.26 |
| 3 | 0.24 |
| 4 | 0.13 |
| 5 | 0.07 |

1. Input the data into a column in Excel.

2. Input the following formula into the column to the right of $P(x)$. Use the Autosum function to sum the results: $\mu = 2.29$ absences per day.

| | A | B | C |
|---|---|---|---|
| 1 | *x* | P(x) | $\mu$ |
| 2 | 0 | 0.12 | =A2*B2 |
| 3 | 1 | 0.18 | |
| 4 | 2 | 0.26 | |
| 5 | 3 | 0.24 | |
| 6 | 4 | 0.13 | |
| 7 | 5 | 0.07 | |

| | A | B | C |
|---|---|---|---|
| 1 | *x* | P(x) | $\mu$ |
| 2 | 0 | 0.12 | 0 |
| 3 | 1 | 0.18 | 0.18 |
| 4 | 2 | 0.26 | 0.52 |
| 5 | 3 | 0.24 | 0.72 |
| 6 | 4 | 0.13 | 0.52 |
| 7 | 5 | 0.07 | 0.35 |
| 8 | | | 2.29 |

3. Input the following formula into the column to the right of $P(x)$.

4. The median is found by estimating where the halfway point is in the data at $x = 2$.

5. Compute the variance by inputting its formula in the next column to the right. The mean has to be referenced by using an absolute reference. The summed result is 1.966.

$$\sigma^2 = \Sigma (x - \mu)^2 \cdot P(x)$$

| | A | B | C | D |
|---|---|---|---|---|
| 1 | *x* | P(x) | $\mu$ | $\sigma^2$ |
| 2 | 0 | 0.12 | | =(A2-$C$8)^2*(B2) |
| 3 | 1 | 0.18 | 0.18 | 0.300 |
| 4 | 2 | 0.26 | 0.52 | 0.022 |
| 5 | 3 | 0.24 | 0.72 | 0.121 |
| 6 | 4 | 0.13 | 0.52 | 0.380 |
| 7 | 5 | 0.07 | 0.35 | 0.514 |

6. The standard deviation is the square root of the variance which = 1.4 absences.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | *x* | P(x) | $\mu$ | $\sigma^2$ | | |
| 2 | 0 | 0.12 | 0 | 0.629 | | |
| 3 | 1 | 0.18 | 0.18 | 0.300 | | |
| 4 | 2 | 0.26 | 0.52 | 0.022 | | |
| 5 | 3 | 0.24 | 0.72 | 0.121 | | |
| 6 | 4 | 0.13 | 0.52 | 0.380 | | |
| 7 | 5 | 0.07 | 0.35 | 0.514 | | |
| 8 | | | 2.290 | 1.966 | =SQRT(D8) | |
| 9 | | | | | SQRT(number) | |

## 2.5 Cumulative Distributions

It's sometimes useful to construct *cumulative* versions of frequency or relative frequency tables to display data. In a **cumulative frequency distribution**, we can show directly the number of data set members *at or below* any specified value.

**Demonstration Exercise 2.7 Cumulative Distributions**

The number of years of seniority for each of your company's 24 employees is shown in the frequency table below:

| Years of Seniority *x* | No. of employees *f(x)* |
|---|---|
| 0 | 9 |
| 1 | 6 |
| 2 | 4 |
| 3 | 3 |
| 4 | 2 |

1. Input the data into a column in Excel.

2. The cumulative frequency starts with the first frequency value.

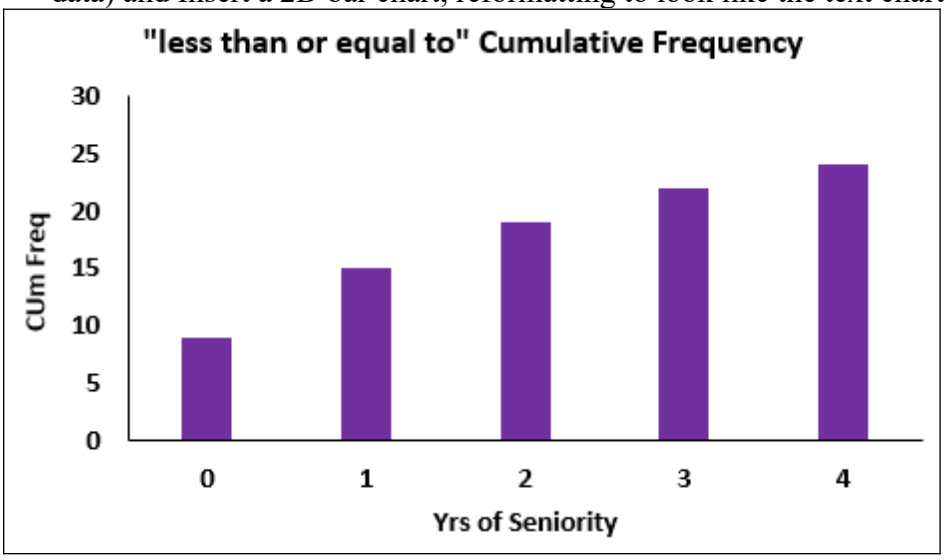
| | A | B | C |
|---|---|---|---|
| 1 | *x* | f(x) | Cum Freq |
| 2 | 0 | 9 | =B2 |
| 3 | 1 | 6 | |
| 4 | 2 | 4 | |
| 5 | 3 | 3 | |
| 6 | 4 | 2 | |

3. The second cell contains the cumulative frequency formula. Copy the formula down.

| | A | B | C |
|---|---|---|---|
| 1 | *x* | f(x) | Cum Freq |
| 2 | 0 | 9 | 9 |
| 3 | 1 | 6 | =C2+B3 |
| 4 | 2 | 4 | |
| 5 | 3 | 3 | |
| 6 | 4 | 2 | |

| | A | B | C |
|---|---|---|---|
| 1 | *x* | f(x) | Cum Freq |
| 2 | 0 | 9 | 9 |
| 3 | 1 | 6 | 15 |
| 4 | 2 | 4 | 19 |
| 5 | 3 | 3 | 22 |
| 6 | 4 | 2 | 24 |

4. Highlight the *x* values and the Cum Freq values (you can highlight separated columns by highlighting the first set of data, then hold the Ctrl key down and highlight the second set of data) and Insert a 2D bar chart, reformatting to look like the text chart.




## 2.6 Grouped Data

When a data set involves a large number of distinct values, effective data presentation may require putting data points together in manageable groups. Grouped data can be effectively displayed in a frequency **histogram**— a kind of bar chart for the grouped data case. A histogram has intervals charted on the *x* axis that are of equal width. Data values are placed into appropriate bins or groupings much like sorting coins. The count of how many data points are in each bin is graphed on the *y* axis.

When constructing a histogram with Excel, you need to define your own bin sizes. Otherwise, if you let Excel do this step for you, you could end up with some odd looking bin sizes with lots of decimal places. The recommendation is to create the bin intervals the same way that is outlined in the text and in our first example on frequency distributions. Key elements: use the range of data divided by the number of intervals desired. This gives you an idea of the bin sizes and you can round the values from there.

**Demonstration Exercise 2.8 Grouped Data.**

Below is a list of dividends paid recently by 60 of the largest firms in the telecommunications industry:

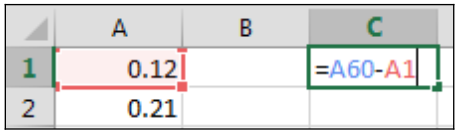
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $1.23 | .56 | .97 | 3.65 | 5.16 | 4.02 | 5.06 | 6.83 | 6.51 | 8.45 |
| .66 | .12 | .80 | 2.54 | 4.12 | 5.17 | 5.45 | 6.02 | 7.94 | 9.66 |
| .21 | 1.31 | .43 | 3.50 | 4.89 | 4.33 | 4.80 | 7.35 | 6.56 | 9.07 |
| 1.09 | .56 | 2.13 | 2.98 | 4.36 | 5.78 | 5.67 | 7.92 | 6.41 | 9.54 |
| 1.45 | 1.43 | 3.21 | 4.78 | 5.66 | 4.21 | 4.39 | 7.14 | 6.83 | 8.22 |
| 1.87 | 1.22 | 2.09 | 5.43 | 4.91 | 5.67 | 6.12 | 6.77 | 7.62 | 8.49 |

a) Show the values in a grouped data frequency table, using the intervals 0 to under $2, $2 to under $4, $4 to under $6, and so on.
b) Draw the histogram for the table that you produced in part a).
c) Using the grouped data table, approximate the mean, the variance and the standard deviation of the dividend amounts, and compare your results to the actual mean, variance and standard deviation of the raw data. (The mean of the raw data is 4.52; the variance is 7.08; the standard deviation is 2.66.)

1. Input the data into a column in Excel of open **Demo Exercise 2-8** accessed from the student companion site.

   Click on the first cell of data, then select the Data tab and click on the lower right arrow of the **Sort & Filter** ribbon button. Select **Sort Smallest to Largest**.


2. Find the **Range**: the range is defined as the difference between the largest and smallest numbers. It would be helpful to sort the data to make it easier to locate those values.
   You can input a simple formula in Excel to subtract the smallest value from the largest value.
   The range in this data set is 9.54.




3. Find the bin size: determine what size interval or how many classes or intervals that you want to use for this particular data set. In this case, if 5 bins are used, the bin width is calculated by dividing the range by the desired number of classes (9.54 divided by 5). The answer is 1.9 and gives us an idea of what we could use for an interval. A rounded off number close to this value would be 2.0 and might work better in a table or graphic.

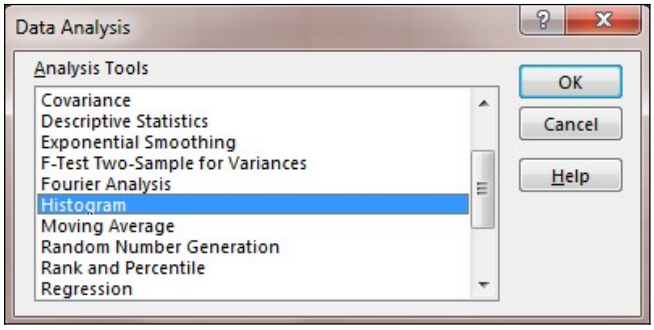
4. The first class endpoint must be 2.0 or lower to include the smallest value. It often does not work to select the first number as the first bin. Instead, use the next larger rounded number. In addition, the last endpoint must be 9.66 or higher to include the largest value. We can start with the interval 0 to 2 and end with 8 to 10. In Excel, data are input in an interval up to and including each interval value. For example, the interval 8 to 10 includes all values up to and including 10. The output we generate in Excel could differ slightly from output generated in another statistical program that may be shown in your text that interprets the intervals differently. Input the bins as numbers in a column to the right of the data set.

| **Bins** |
|---|
| 2 |
| 4 |
| 6 |
| 8 |
| 10 |

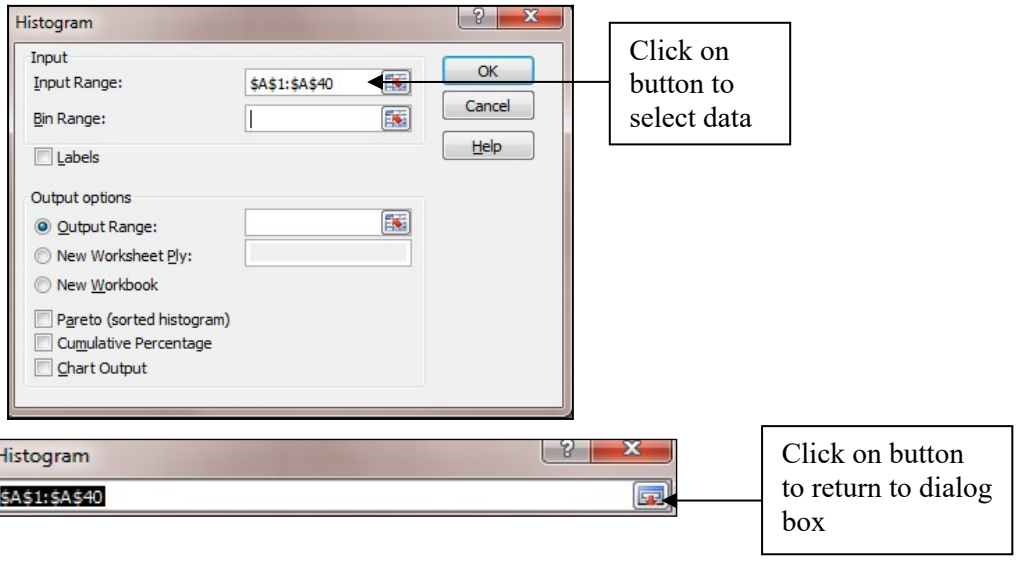
5. Determine how many data points go into each interval by using the **Data Analysis** capability in the Analysis Toolpak in Excel.

**Note on Analysis Toolpak:**
Check to see if the **Analysis Toolpak** has been installed by selecting the **Data** tab to see if there is a Data Analysis button on the right side of the ribbon. If you do not see it, select **File** ⇨ **Options** ⇨ **Add-ins**. At the bottom of the dialog box, you will see **Manage Excel Add-ins**. Click on the **Go...** button. Check the boxes for the **Analysis Toolpak** and the **Solver Add-ins** and **OK.** You should now see the Data Analysis and Solver options.

6. Select the **Data** tab and then **Data Analysis** on the right side of the ribbon. Select **Histogram** and **OK**. Select the data for the **Input Range** and the interval sizes for the **Bin Range**. Select an output cell that is to the right of the data. Check **Cumulative Percentage** and **OK**. The result should display the frequency or count of data points in each interval and then the cumulative percentage. Remember that this display could be slightly different than your text because of the way Excel interprets the bin intervals by using the rule – "up to and including."




**How to Select Data:** Click on the first cell of the data set, hold down the left mouse button and drag the cursor down to the last cell and release the mouse button. If the dialog box is obstructing the data, there are two methods for accessing the data. You can move the dialog box by clicking and dragging the title bar, and moving the dialog box to another location. You can also click on the button to the right of the box for **Input Range** or other input boxes like this. The dialog box will shrink to the size of the input box. When you are done selecting the data range, click on the box to the right of the input box to return to the dialog box and complete the selections.

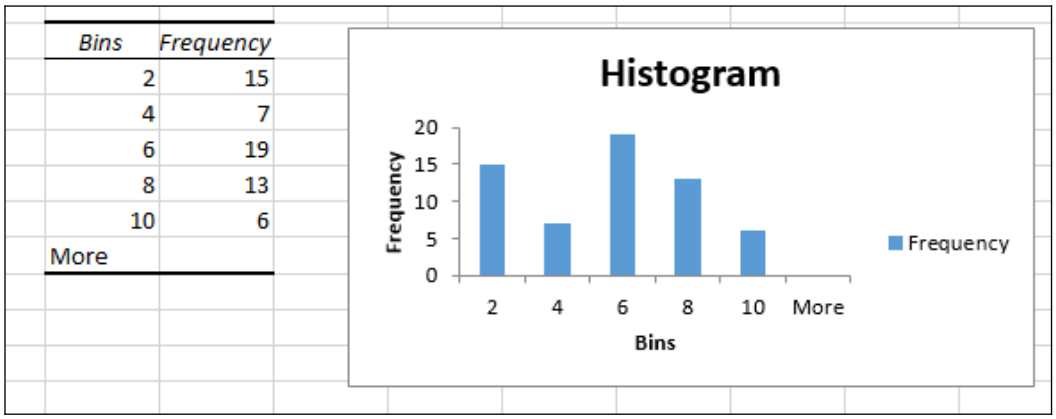


7. Select the values for the Bin Range in the same way, not including the label.

8. Click inside the **Output Range** box under the **Output Options**. Verify that the cursor is inside the box. Otherwise, a previously selected range could be changed. Select a cell to the right of the data set such as cell B1. This cell will be the upper left cell of the output placed on the worksheet.

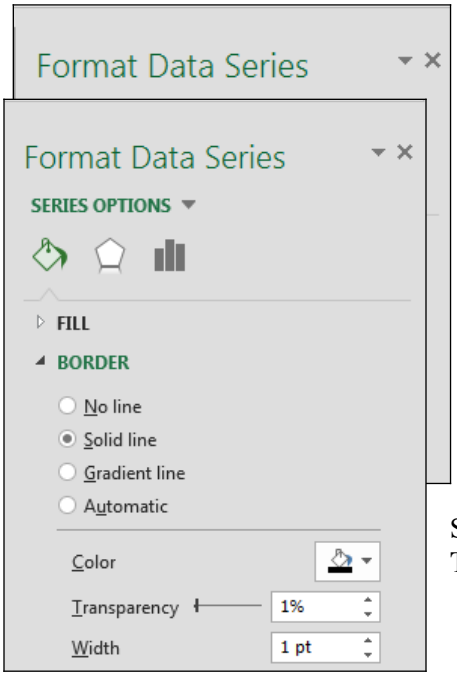
9. Select **Chart Output** to create the histogram. Click **OK**.

10. The resulting histogram still requires several changes. First, eliminate the **More** category by selecting the histogram chart by clicking on it once, then drag the lower-right blue box (at the bottom of the Frequency column) up one line.



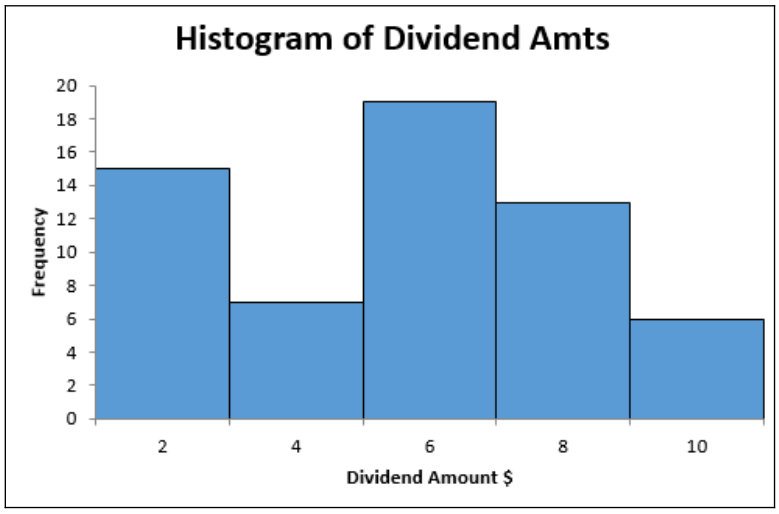

11. Delete the legend by clicking on it once and **Delete** using the keyboard.

12. Eliminate the gaps between the bars by right-clicking on any bar and select **Format Data Series**, change the **Gap width** to 0% and **Close X**.

13. 

Select **Border Color**, **Solid line**, black **Color** and **Transparency** at 100%.

14. Resize the chart and change the title and labels.




15. Using the grouped data table, we can approximate the mean, using the formula:

$$\mu = \frac{\Sigma m_i \cdot f_i}{N}$$

16. Insert a column between the Bins and Frequency columns and input the bin midpoints.

| *Bins* | *Midpoints* | *Frequency* |
|---|---|---|
| 2 | 1 | 15 |
| 4 | 3 | 7 |
| 6 | 5 | 19 |
| 8 | 7 | 13 |
| 10 | 9 | 6 |

17. In a column to the right, input a formula for the midpoint*frequency and copy it down. Sum the result and divide by N or 60.

| *Bins* | *Midpoints* | *Frequency* | *Mean* |
|---|---|---|---|
| 2 | 1 | 15 | =G2*H2 |
| 4 | 3 | 7 | |

| Bins | Midpoints | Frequency | Mean | |
|---|---|---|---|---|
| 2 | 1 | 15 | 15 | |
| 4 | 3 | 7 | 21 | |
| 6 | 5 | 19 | 95 | |
| 8 | 7 | 13 | 91 | |
| 10 | 9 | 6 | 54 | |
| | | 60 | 276 | Sum |
| | | | =I7/H7 | |

18. The approximated mean is $4.60 compared to the raw data mean of $4.52.

19. Using the grouped data table, we can approximate the mean, using the formula:

$$\sigma^2 = \frac{\Sigma (m_i - \mu)^2 \cdot f_i}{N}$$

20. Input the numerator into the first cell of the variance column and copy the formula down. Sum the resulting values and divide by N. The result is 6.773.

| Bins | Midpoints | Frequency | Mean | Variance |
|---|---|---|---|---|
| 2 | 1 | 15 | 15 | =(G2-$I$8)^2*H2 |
| 4 | 3 | 7 | 21 | |

| Bins | Midpoints | Frequency | Mean | Variance | |
|---|---|---|---|---|---|
| 2 | 1 | 15 | 15 | 194.4 | |
| 4 | 3 | 7 | 21 | 17.92 | |
| 6 | 5 | 19 | 95 | 3.04 | |
| 8 | 7 | 13 | 91 | 74.88 | |
| 10 | 9 | 6 | 54 | 116.16 | |
| | | 60 | 276 | 406.4 | Sum |
| | | | 4.6 | =J7/60 | |

| Variance | |
|---|---|
| 194.4 | |
| 17.92 | |
| 3.04 | |
| 74.88 | |
| 116.16 | |
| 406.4 | Sum |
| 6.773333 | |
| =sqrt(J8 | |

SQRT(**number**)

21. The approximated variance is $6.773 compared to the raw data variance of $7.08.

22. The approximated standard deviation is the square root of the variance which is $2.60 compared to the raw data variance of $2.66.

## SUMMARY OF EXCEL COMMANDS USED IN CHAPTER 2

**Creating Charts & Graphs (General)**

- Click on the **Insert** tab found along the top of an Excel worksheet. You can construct many different types of charts, including column charts, line charts, pie charts, bar charts, area charts, and XY (scatter) charts.
- Excel can generate frequency distributions and histograms using the **Data Analysis** feature.

**Data Analysis Tool**

Select the **Data** tab along the top of an Excel worksheet. If the **Data Analysis** feature has been uploaded into your Excel package, it will be found in the **Analysis** section at the top right of the **Data** tab page. If **Data Analysis** does not appear in the **Analysis** section, it must be added in.

To add in **Data Analysis:**
1.) Click on the **File** tab.
2.) Click on **options** in the menu.
3.) In the **Excel options** dialog box, click on **Add-Ins** next to the bottom of the left menu. A screen of add-ins will appear.
4.) Click on **Analysis ToolPak** and then click on **Go**… at the bottom of the page.
5.) In the dialog box **Add Ins**, check the box to the left of **Analysis ToolPak** and click **OK**. Your **Data Analysis** feature is now uploaded onto your computer, and you won't need to add it in again. You can bring up the **Analysis ToolPak** feature at any time by going to the **Data** tab at the top of the Excel worksheet and clicking on **Data Analysis**.

**Constructing Frequency Distributions (Histograms)**

In Excel, frequency distributions are referred to as histograms, and the classes of a frequency distribution are referred to as bins. If you do not specify bins (classes), Excel will automatically determine the number of bins and assign class endpoints based on a formula. If you want to specify bins, load the class endpoints that you want to use into a column.

- Select the **Data** tab in the Excel worksheet and then select the **Data Analysis** feature (upper right). If this feature does not appear, you may need to add it (see above).
- Click on **Data Analysis**, the dialog box features a pulldown menu of many of the statistical analysis tools presented and used in this text. From this list, select **Histogram**.
- In the **Histogram** dialog box, click in the space beside **Input Range and select the raw data values**.
- Place the location place the location of the raw data values of the class endpoints (optional) in the space beside **Bin Range**. Leave this blank if you want Excel to determine the bins (classes).
- If you have labels, check **Labels**. If you want a histogram graph, check **Chart Output**. If you want an ogive, select **Cumulative Percentage** along with **Chart Output**. If you opt for this, Excel will yield a histogram graph with an ogive overlaid on it.

### Creating Charts

- Select the **Insert** tab from the top of the Excel worksheet.
- In the **Charts** section, which is the middle section shown at the top of the **Insert** worksheet, there are icons for column, line, pie, bar, area, scatter, and other charts. Click on the icon representing the desired chart to begin construction. Each of these types of charts allow for several versions of the chart shown in the dropdown menu. For example, the pie chart menu contains four types of two-dimensional pie charts and two types of three-dimensional pie charts. To select a particular version of a type of chart, click on the type of chart and then the version of that chart that is desired.

### Frequency Polygons

- **Frequency polygons** can be constructed by using the **Histogram** feature. Follow the directions shown above to construct a histogram.
- Once the histogram is constructed, right-click on one of the "bars" of the histogram. From the dropdown menu, select **Change Series Chart Type**. Next select a line chart type. The result will be a frequency polygon.

### Ogive Chart

An ogive can be constructed at least two ways.

- One way is to cumulate the data manually. Enter the cumulated data in one column and the class endpoints in another column. Click and drag over both columns. Go to the **Insert** tab at the top of the Excel worksheet. Select **Scatter** as the type of chart. Under the **Scatter** options, select the option with the solid lines. The result is an ogive.
- A second way is to construct a frequency distribution first using the **Histogram** feature in the **Data Analysis** tool. In the **Histogram** dialog box, enter the location of the data and enter the location of the class endpoints as bin numbers. Check **Cumulative Percentage** and **Chart Output** in the **Histogram** dialog box. Once the chart is constructed, right-click on one of the bars and select the **Delete** option. The result will be an ogive chart with just the ogive line graph (and bars eliminated).

### Bar Charts & Column Charts

Bar charts and column charts are constructed in a manner similar to that of a pie chart. Begin by entering the categories in one column and the data values of each category in another column in the Excel worksheet. Categories and data values could also be entered in rows instead of columns. Click and drag over the data and categories for which the chart is to be constructed.

- Go to the **Insert** tab at the top of the worksheet.

- Select **Column** or **Bar** from the **Charts** section and the select the version of the chart to be constructed. The result is a chart from the data.
- Once the bar chart or column chart has been constructed, there are many options available. By right-clicking on the bars or columns, a menu appears that allows you, among other things, to label the columns or bars. This command is **Add Data Labels**. Once data labels are added, clicking on the bars or columns will allow you to modify the labels and the characteristics of the bars or columns by selecting **Format Data Labels...** or **Format Data Series...**.
- Usage of these commands is the same as when constructing or modifying pie charts (see above).Various options are also available under **Chart Tools** (see pie charts above).