

FREQUENCY DISTRIBUTIONS AND GRAPHS IN EXCEL

This lab exercise will show you how to use Microsoft Excel (2007) and a grouped frequency distribution to create a histogram and a frequency polygon from a set of raw data. The following list gives the resources in Bluman's Elementary Statistics, Eighth Edition to help you with the tasks outlined in this assignment:

- Sections 2-1 and 2-2
- Page 23 (Excel Step by Step) Excel's Analysis Tool-Pak Add-In
- Page 65 (Excel Step by Step) Constructing a Histogram
- Page 67 (Excel Step by Step) Constructing a Frequency Polygon

Note: If the instructions in the textbook differ from those outlined in this assignment, please follow the instructions in the assignment.

The Data

The Cy Young Award is given annually to the best pitchers in Major League Baseball, one from the American League and one from the National League. Winners are decided by votes of the members of the Baseball Writers Association of America. The award is named for the Hall of Fame pitcher Cy Young (1867 – 1955). The award was first given in 1956 and at that time was given to only one pitcher each year. Beginning in 1967, the current format of awarding one pitcher from each league was adopted. The data set includes the winners from 1967 – 2011. For each pitcher, the following information on his season is included: year won, team, league, ERA (earned run average), number of games won, number of games lost, number of strikeouts (SO), number of innings pitched (IP), height (in inches), age when he won the award, and throwing hand. The data were compiled from www.baseball-reference.com and www.baseball-almanac.com.

Getting Started: The Histogram

Follow the directions to create a “practice” histogram for the *Ages* of the Cy Young Award winners. Create the histogram and compare your results to the solution provided at the end. After that you will create a second histogram to turn in for credit.

1. After logging on to the computer, go to the course website and click on the link for “Lab 1 Data Set” (cy_young_data.xlsx). When you click on the link, the computer will select Microsoft Office Excel as the default program with which to open the file. Click **OK** to open the file. In the Cy Young spreadsheet you should see ninety-one rows of data, organized into twelve columns. Each column is labeled to indicate which piece of information about the winner is shown there.

2. In the ribbon at the top of the Excel screen, click on the **Data** tab and check that there is a command titled **Data Analysis** located in the **Analysis** group. This is an Excel Add-In. If you are working in the computer lab, the Analysis Tool-Pak should already be there. If you do not see the **Data Analysis** option, follow the directions on page 23 of your textbook to add the package.
3. You are going to create a histogram for the *Ages* of the award winners in a blank Excel worksheet. To open a new Excel workbook, press CTRL+N. In cell G1 of the blank worksheet, please type your name and your class meeting days and time (*example*: Bob Jones, TTh 9 – 10:30).
4. From the Cy Young spreadsheet, copy the *Age* data in column K, including the column title, and paste it into column A of your new worksheet, starting in cell A1. To do this: highlight the data (by left-clicking on cell K1 and dragging the mouse down to cell K92; or by clicking on cell K1, holding down the SHIFT key and then clicking on cell K92), press CTRL+C, place the cursor in cell A1 of the new worksheet and press ENTER or CTRL+V.
5. One way to create a histogram in Excel is to provide data and *bins*. You have the age data in Column A of your new worksheet; you will enter the bins by hand. *Bin* is the Excel name for *upper class limit*. Follow the procedure from Section 2-1 to determine the class width and class limits for a grouped frequency distribution of the *Age* data with eight classes.
 - **Tip:** Use built-in Excel functions to help calculate the range of the data set. Click on cell D1 and then click on the “insert function” icon f_x (next to the blank formula bar above the worksheet). In the *Insert Function* dialog box that appears, select the **Statistical** category from the menu to see a list of all the built-in statistical functions included in Excel. Scroll down and double-click on MAX(), the function that returns the maximum value from a list of values. In the *Function Arguments* dialog box that appears, enter A2:A92 (the first cell you want to include, then a colon, then the last cell you want to include) in the box labeled **Number1**. Click **OK**. In cell D1, you should see the maximum age in the list, 42. Label this value by typing *Max* in cell E1. Now find the minimum age a little faster: in cell D2, type =MIN(A2:A92) and hit ENTER to see the minimum age in the list, 20. Label this value by typing *Min* in cell E2. (This is the way to use a built-in function without having to scroll through a list to find it; always start by typing the equal sign when you want to use a function or formula in Excel.) To find the range, use Excel to compute the difference between the maximum and minimum values by typing = D1 - D2 in cell D3 and pressing ENTER. Label this value by typing *Range* in cell E3. *Note:* Now that you know how to use functions in EXCEL you could compute the age range in one step by typing =MAX(A2:A92) – MIN(A2:A92) in any blank cell.
6. Using the procedure from Section 2-1 (and starting the first class with a lower class limit of 20), you should have found the following upper class limits: 22, 25, 28, 31, 34,

37, 40, 43. Click cell B1 and type the name *BINS*. Enter the upper class limits into column B, beginning in cell B2.

7. In the ribbon at the top of the screen, click on the **Data** tab and in the **Analysis** group click on the **Data Analysis** command. On the menu that appears, double-click on **Histogram**. This should produce a *Histogram* dialog box where you will enter the Input Range and Bin Range. This can be done in one of two ways:

(a) In the **Input Range** box enter the first cell you want to include, then a colon, then the last cell you want to include (just as you did when working with the MAX() and MIN() functions in Step 5). This is the method described on page 65 of the textbook. For the *Age* histogram, the **Input Range** is A2:A92 and the **Bin Range** is B2:B9.

(b) Alternatively, click on the icon with the red arrow to the right of the **Input Range** box, highlight the block of cells containing the data, and then click the icon again. The same procedure can then be used to enter the **Bin Range**.

8. The next step in the *Histogram* dialog box is to specify where you would like the output to be displayed. You can click on **Output Range** and specify a cell in the spreadsheet where you want to see the histogram. Alternatively, you can click on **New Worksheet Ply** and Excel will output the histogram on a separate worksheet from the data. For this practice histogram, select **Output Range** as the output option and enter a cell such as C11.

9. Finally, before closing the *Histogram* dialog box, check the box next to **Chart Output**. Click **OK** to close the dialog box. A frequency table and histogram should appear on your worksheet with the upper left corner of the table in the cell you entered as the **Output Range**.

10. You will now make a few minor modifications to the histogram.

(a) In the *Bin* and *Frequency* columns of the frequency table, highlight the cells that read "More" and "0." Right-click, select **Delete**, click the bubble next to **Shift cells up** and click **OK**. This modifies the histogram so that it covers only the range of values included in the *Age* data set.

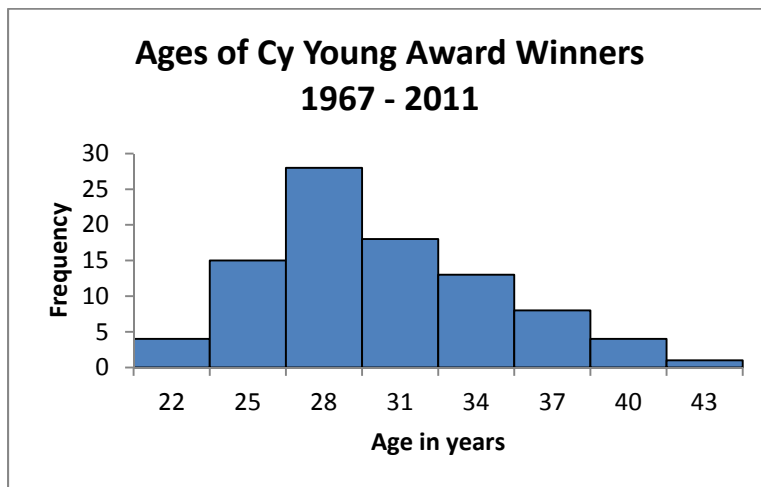
(b) Next you will remove the gaps between the bars on the histogram. To do this, right-click on any of the bars in the histogram and select **Format Data Series** from the menu that appears. In the *Format Data Point* dialog box, under "Gap Width," slide the tab over to "No Gap" (or enter 0 as the "Gap Width" value). Click on **Close** and you will see the histogram no longer has gaps between its bars.

(c) Next you will edit the title and axes labels for the histogram. Click on the current title, "Histogram," and type a more descriptive title, such as *Ages of Cy Young Award*

Winners 1967 - 2011 and hit ENTER. You can change the font, size and style of the title using the options in the **Font** group of the **Home** tab. Reduce the title font size to 14. Click on the horizontal axis label, "Bins," and type a more descriptive label, such as *Age in years*. The vertical axis label, "Frequency," is appropriate, so do not change it.

(d) Finally, you will remove the legend from the graph and show borders for the histogram bars. Click on any of the histogram bars and a group of tabs labeled **Chart Tools** should appear on the ribbon at the top of the screen. Click on the **Layout** tab and from the **Labels** group select **Legend**. From the menu that appears, select **None** (Turn off Legend). Now click on the **Format** tab in **Chart Tools**. From the **Shape Styles** group, select **Shape Outline** and choose the color black. Each of the histogram bars should now be outlined.

Your histogram for the ages of Cy Young Award winners since 1967 should look like this:



Note: If your histogram does not show the scale of the *Frequency* axis in increments of five, left-click either the lower left or right hand corner of the box containing the graph and you will see the ↙ symbol. Drag the corner of the box straight down and you should see the y-axis scale adjust. (If you hold down SHIFT while you drag the corner, both the height and width of the box will adjust uniformly.)

(e) There are other (optional) modifications you can make. Here are a few examples: click on the **Design** tab to change the histogram bar color using **Chart Styles**; click on the **Layout** tab and use the **Labels** and **Axes** groups to change the positions of the title, axes labels and data labels; or click on the **Format** tab and use the **Size** group to resize the histogram (this is an alternative to the method described in the note above). You can also copy and paste the histogram into a Word document by right-clicking on the box containing histogram and selecting **Copy** from the menu that appears. Feel free to experiment with these menus to see what other changes can be made.

11. WHAT YOU WILL TURN IN FROM THIS SECTION:

At the end of the assignment, you will submit one Excel file containing all of your work for the lab. **Sheet 1** of your Excel workbook should now contain:

- (a) The two columns of data (*Ages* and *Bins*).
- (b) The *Ages* frequency table and the *Ages* histogram, properly labeled.

You will return to the *Ages* frequency distribution later in this assignment, so make sure to save the worksheet now.

Now It's Your Turn: The Histogram

Now you will create a histogram for the *ERA* (earned run average) of the award winners to submit for credit.

Note: Earned run average refers to the **average** number of earned runs a pitcher gives up in a full game (9 innings); runs that result from defensive errors are not counted as earned runs. The *ERA* values you see listed were calculated using this formula:

$$ERA = \frac{\text{number of earned runs}}{\text{number of innings pitched}} \times 9$$

1. Look at the bottom of the screen and you will see that your work with the *Age* data is located in **Sheet 1**. Click the tab for **Sheet 2** and you will see a blank worksheet. Use **Sheet 2** for your work with the *ERA* data. In cell G1 of this sheet, please type your name and your class meeting days and time (*example:* Bob Jones, TTh 9 – 10:30).
2. Carefully follow the process outlined in steps 4 – 10(d) of **Getting Started** to create a histogram for the *ERA* data. To receive full credit for this assignment, you must edit and properly label the histogram as demonstrated in the *Ages* example. **Hint:** Since the *ERA* data set consists of values less than four (given to the hundredths place), the class width should be less than one. Try using a two-digit decimal as the class width.

3. WHAT YOU WILL TURN IN FROM THIS SECTION:

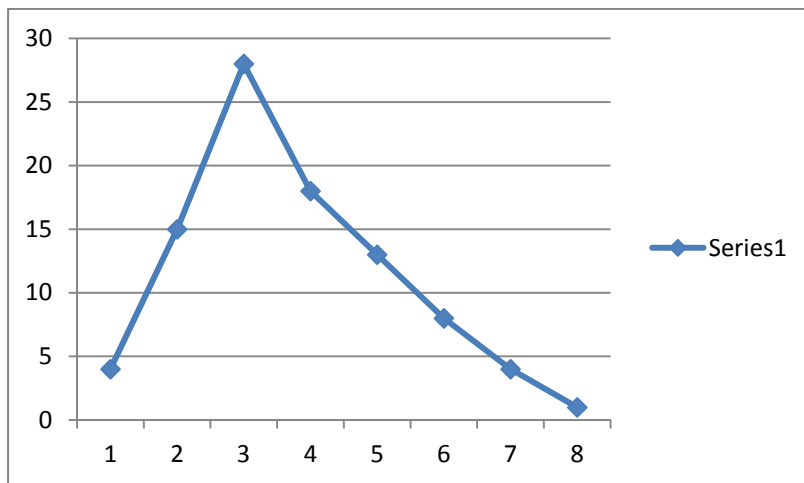
At the end of the assignment, you will submit one Excel file containing all of your work for the lab. **Sheet 2** of your Excel worksheet should now contain:

- (a) The two columns of data (*ERA* and *Bins*).
- (b) The *ERA* frequency table and the *ERA* histogram, properly labeled.

Now It's Your Turn: The Frequency Polygon

Follow the directions to create a frequency polygon for the *Ages of the Cy Young Award winners*. There will be no “practice” polygon; just create the polygon according to the steps below to submit for credit.

1. Open **Sheet 3** (a blank Excel worksheet) by clicking on its tab at the bottom of the screen. In cell G1 of the worksheet, please type your name and your class meeting days and time (*example*: Bob Jones, TTh 9 – 10:30).
2. One way to create a frequency polygon in Excel is to provide frequencies and midpoints for a grouped frequency distribution. You have the frequencies and bins in **Sheet 1** from your previous work with the *Age* data. Copy and paste the frequency column (including the column title) from the *Age* frequency table in **Sheet 1** into column A of **Sheet 3**, starting in cell A1. In cell B1, type *Midpoints*. You can calculate the midpoints from the bins (upper class limits) using the procedure from Section 2-1. Recall that you started the first class with a lower class limit of 20. Enter the midpoints into column B, beginning in cell B2.
3. Highlight the *Frequency* data in column A (you do not need to highlight the title). Click on the **Insert** tab, and in the **Charts** group click on the **Line** command. From the menu that appears, select **2-D Line with Markers**. This should produce a line graph that looks like the one below:



Note: Line graphs in Excel can also be created using the **Scatter** command in the **Charts** group. You will work this feature in a later lab assignment.

4. Looking at the horizontal axis of the graph, you should notice that the values shown do not reflect the midpoints for the *Age* distribution. You will fix this now. Right-click on any portion of the graph and choose **Select Data** from the menu that appears. In the *Select Data Source* dialog box that appears, click **Edit** under the **Horizontal (Category)**


Axis Labels. This prompts the *Axis Labels* dialog box to appear. In the **Axis label range** space provided, enter the range of cells containing the midpoints, using one of the two methods outlined in Step 7 of the histogram exercise. Click **OK** in the *Axis Labels* box, then click **OK** in the *Select Data Source Box*. The graph should now have the midpoint values displayed along the horizontal axis.

5. You will now make a few more minor modifications to the frequency polygon.

(a) First you will edit the chart and axes titles for the graph. Click on any portion of the graph to make the **Chart Tools** group of tabs appear at the top of the screen. Click on the **Layout** tab and from the **Labels** group, select **Chart Title**. The menu that appears gives options for title placement, from it select **Above Chart**. The words “Chart Title” should appear above the graph; click on them and type a more descriptive title, such as *Ages of Cy Young Award Winners 1967 - 2011*, and hit ENTER. Reduce the title font size to 14. To edit the axes titles, once again click on the **Layout** tab in **Chart Tools**, but this time select **Axis Titles** from the **Labels** group. From the menu that appears, select **Primary Horizontal Axis**, then **Title Below Axis**. The words “Axis Title” should appear below the horizontal axis; click on them and type a more descriptive label, such as *Age in years*. Follow similar steps to label the vertical axis *Frequency*. *Recommendation:* choose the **Rotated Title** option under **Primary Vertical Axis**.

(b) Next, remove the legend from the graph. Under **Chart Tools**, click on the **Layout** tab and from the **Labels** group select **Legend**. From the menu that appears, select **None** (Turn off Legend).

(c) Now you will change the type of “markers” used to show points on the graph. The default marker is a diamond shape. Right-click on any one of the markers and from the menu that appears, select **Format Data Series**. The *Format Data Series* dialog box should open; in it click on **Marker Options**. Click the bubble next to **Built-in** and select the circle from the drop-down menu. You can also resize the points; in this case reduce them to size 5 and click **Close**. You will see the results on your graph.

(d) Finally, you will edit the frequency polygon at the “beginning” and “end” as discussed in class. Excel allows you to add data to a graph even after you have formatted it. In your worksheet, click on row 2, then click on the **Home** tab. In the **Cells** group click on **Insert**, then **Insert Sheet Rows** to insert a blank row 2. In the frequency polygon graph, you need to add two points on the horizontal axis (where the midpoints *before* the first one and *after* the last one would fall). In row 2, enter 0 for the frequency (to place the point on the horizontal axis) and enter 18 for the midpoint. In row 11, enter the appropriate values for the second new point. Now click on the line graph and you will see two separate colored borders appear around the columns of data. To include the two new points on the graph, resize the borders around the data columns to include rows 2 and 11. This can be done by placing the cursor over any corner of the border (until you see the  symbol), left-clicking and dragging the

border up or down as appropriate. Make sure to do this for both the *Frequency* and *Midpoint* columns.

Your frequency polygon for the *Ages of Cy Young Award* winners is now complete. As with the histogram, there are many other optional modifications you can make, but if you have finished through step 5(d), you are now ready for Step 6.

6. **TURN IN:**

One Excel file containing three worksheets. The contents of **Sheet 1** and **Sheet 2** are described above. **Sheet 3** of your Excel worksheet should now contain:

- (a) The two columns of data (*Frequency* and *Midpoints*).
- (b) The *Ages* frequency polygon, properly labeled.

Each sheet should contain your name in cell G1. Please submit your Lab 1 Excel file via e-mail.

PLEASE NOTE: To verify that you are practicing the use of formulas in Excel and not just hand-computing values, the “formula view” of your work will be graded (in addition to the values themselves). To see the “formula view” of your worksheet, hold down the CTRL key and type ~ (tilde). (Do not hold down SHIFT, even if it looks like ~ is the “uppercase” option on the keyboard.) To return to the previous version, just hit CTRL~ again. Remember this so that for each lab you can check your “formula view” before submitting it for credit.