

DESCRIPTIVE STATISTICS IN EXCEL

This lab exercise will show you how to use Microsoft Excel (2007) to compute numerical descriptive statistics for 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 3-1 and 3-2
- Page 123 (Excel Step by Step) Finding Measures of Central Tendency
- Page 141 (Excel Step by Step) Finding Measures of Variation
- Page 161 (Excel Step by Step) Descriptive Statistics in Excel

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

The Data

For this assignment, you will again work with the Lab 1 data set of Cy Young Award 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: Descriptive Statistics

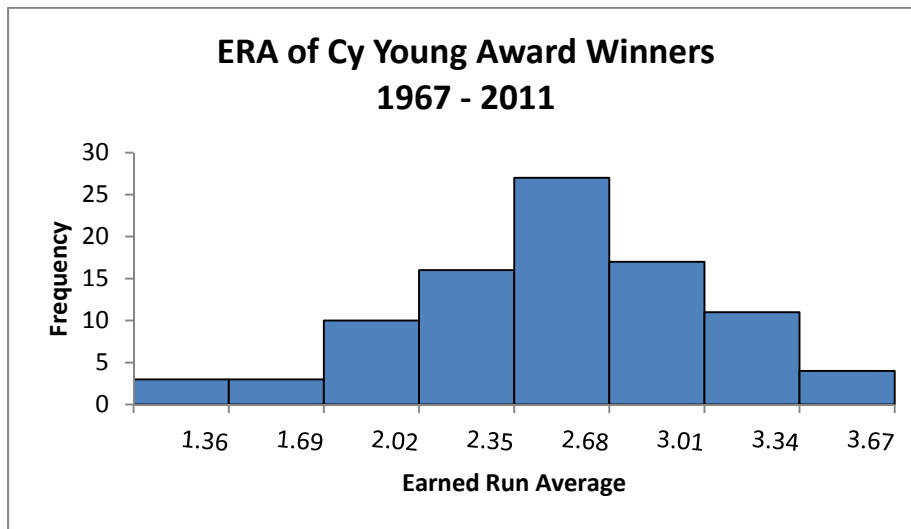
1. If you have not already done so, print pages 2, 3 and 5 of the lab instructions so that you can handwrite responses to portions of the questions directly on them. Now go to the course website and click on the Excel file "Lab 1 Data Set" (cy_young_data.xlsx). When you click on the link, the computer will select Microsoft Office Excel as the default program with which to open the file. Click **OK** to open the file.
2. Open a new Excel workbook by pressing 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).
3. One way to compute descriptive statistics in Excel is by using built-in functions. To get warmed up, try this: copy and paste the data column *THROW HAND* (including the column title) from the Cy Young spreadsheet into column A of your blank worksheet. Next to the data, in cells B2 and B3, type the labels *Right* and *Left*, respectively. You will use the function COUNTIF() to determine the number of right-handed pitchers in the list. Recall from Lab 1 that you can access built-in Excel functions either by clicking the "insert function" icon f_x and scrolling through the list of options (look in the **Statistical** category for COUNTIF()), or by typing = in a blank cell followed by the name of the

function and its appropriate inputs. Either way, in cell C2, you should have the function “=COUNTIF(range, criteria)” and now you need to enter the input range and criteria. The input range of cells is the *THROW HAND* data column and you can enter it by typing the appropriate first cell number: last cell number or by highlighting the block of cells you wish to include. After that, type a comma, then the criteria “R” (including the quotations) and press ENTER. This is telling Excel to count the cell only if it contains an R; the quotations are necessary in this case because you are dealing with a non-numeric entry. In cell C2, you should now see the number of right-handed pitchers in the list. To determine the number of left-handed pitchers in the list, you could subtract the number of right-handed pitchers from the total number of pitchers in the list. Instead, use the COUNTIF() function again in cell C3 and change the criteria to “L.” Use your results to fill in the blank below:

Approximately _____% of the Cy Young Award winners since 1967 have been right-handed pitchers. Round the answer to the nearest tenth.

4. Next you will calculate measures of central tendency for the *ERA* data from the Cy Young spreadsheet. Copy and paste the data column *ERA* (including the column title) from the spreadsheet into column D of your worksheet (use the same worksheet as for the *THROW HAND* data). Next to the data, in cells E2, E3 and E4, type the labels *Mean*, *Median*, and *Mode*, respectively. Next to the corresponding label, calculate these values using the Excel built-in functions AVERAGE(range), MEDIAN(range) and MODE(range). These functions can be accessed from the **Statistical** category of the f_x menu or by typing them in yourself (just remember to type the equal sign first if you choose to type them). The input “range” refers to the range of cells you want to include in the calculation, in this case make sure to include all of the data cells in the *ERA* column.

5. In Lab 1 you created a histogram for the *ERA* of Cy Young Award winners since 1967:



(a) On the histogram shown above, please label the mean (rounded appropriately), median and mode that you calculated in Step 4 above. *Note:* If Excel does not display the mean to the number of decimal places that you need, right-click on the cell containing the mean and select **Format Cells** from the menu that appears. In the *Format Cells* dialog box, click on the **Number** tab and enter the desired number of decimal places in the space provided.

(b) List the names of the three measures of center for the *ERA* data in ascending order (from left to right as they appear on the graph).

(c) Are the three measures of center for the *ERA* data close in value? _____

(d) Does the *ERA* distribution appear to be uniform, right-skewed, left-skewed or approximately symmetric? _____

6. Now you will calculate measures of variation for the *ERA* data. In column E of your worksheet, type the labels *Range*, *Variance* and *Std Dev* under the labels for the measures of center. Next to the label for range, in column F, use the functions MAX() and MIN() introduced in Lab 1 to calculate the range of the *ERA* data set. Excel has built-in functions for calculating the variance and standard deviation (VAR(range) and STDEV(range), respectively); use them to calculate these values for the *ERA* data set in the cells next to their corresponding labels. As before, the input “range” refers to the range of cells you want to include in the calculation. If necessary, adjust the cells containing the variance and standard deviation to display the correct number of decimal places as you previously did for the mean.

7. Now you know how to use built-in Excel functions to calculate descriptive statistics. These values (and more) can also be found using Excel’s Data Analysis Add-In.

- *Note:* In Lab 1 you gained some experience working with Excel, so you know that menu options are organized into tabs, groups and commands. In the lab instructions from now on, the notation for this will be shortened to: **Tab→Group→Command**. For example, the instruction “Click on the **Data** tab and in the **Analysis** group click on the **Data Analysis** command” will now be written “go to **Data→Analysis→Data Analysis**.”

Go to **Data→Analysis→Data Analysis**. In the menu that appears, double-click on **Descriptive Statistics**. In the *Descriptive Statistics* dialog box that appears, enter the cells for the *ERA* data as the input range, enter E9 as the output range (otherwise the output will be displayed on a new sheet), and check the box next to “Summary statistics.” Click **OK** and the output should appear in your worksheet. You probably will not be able to read everything in the table, so with column E highlighted, go to

Home→Cells→Format and select **AutoFit Column Width**. This should widen the column enough so that you can read each of the entry labels. First, check the values for the three measures of center and the three measures of variation -- do they match up with the values you calculated above using built-in functions? They should! Second, notice there are additional values included in the table: standard error, kurtosis, skewness, minimum, maximum, sum and count. Later in the course we will discuss *standard error* (more specifically, standard error of the mean). We will not cover kurtosis and skewness in this course, but for your reference *kurtosis* measures the peak of the distribution and *skewness* measures the asymmetry of the data set (both are used as measures of *normality*). *Minimum* and *maximum* refer to the smallest and largest values in the data set, respectively; *sum* gives the sum of all the values in the data set and *count* refers to how many values are included in the data set.

8. WHAT YOU WILL TURN IN FROM THIS SECTION:

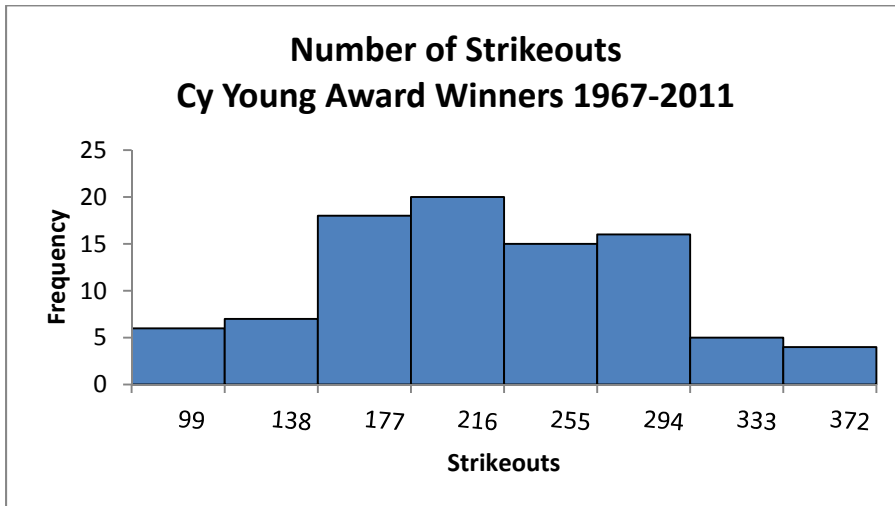
(a) At the end of the assignment, you will submit one Excel file containing all of your work for the lab. **Sheet 1** of your workbook should now contain all of the Excel work you completed in Steps 1 – 7 above, including the *THROW HAND* and *ERA* data. You will continue the assignment in this workbook, so save it now and keep it open.

(b) The filled-in answers to questions in Steps 3 and 5 above. You only need to turn in the pages where you wrote something. Please make sure **YOUR NAME** is at the top of each of these pages.

Now It's Your Turn: Descriptive Statistics

1. Open **Sheet 2** (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. In column A of your blank worksheet, copy and paste the data column *SO* (including the column title) from the Cy Young spreadsheet. "*SO*" stands for the number of strikeouts the pitcher had during the season.
3. Calculate the mean, median, mode, range, variance and standard deviation for the *Strikeouts* data set. You may use the built-in Excel functions or the Data Analysis Add-In to do so. If you use built-in functions, please type in cell labels next to your computations as demonstrated in **Getting Started**; if you use the *Descriptive statistics* option, please display the output chart in cell C3 and widen the labels column so that each cell is readable.

4. Here is a histogram for the *Strikeouts* data. Use it to answer the questions below.



(a) On the histogram shown here, please label the mean (rounded appropriately), median and mode that you calculated in Step 3 above.

(b) List the names of the three measures of center for the *Strikeouts* data in ascending order (from left to right as they appear on the graph).

(c) Does the *Strikeouts* distribution appear to be uniform, right-skewed, left-skewed, approximately symmetric or none of the above? _____

(d) Comparing the two distributions, which is more “spread out”: *ERA* or *Strikeouts*? _____ Explain.

(e) Give a possible reason why there might be more variation in the number of *Strikeouts* thrown by Cy Young Award winners than in their *ERAs*. (Feel free to discuss this with your classmates or with me if you are feeling lost!)

(f) Which measure do you think is a better basis for comparing two pitchers’ performances in a season – their *ERAs* or their number of strikeouts thrown? Explain.

Interesting Note: A baseball statistician named Bill James, together with ESPN.com's Rob Neyer, has developed a formula for predicting the winner of the Cy Young Award. The formula is based on various season statistics for the pitcher, including innings pitched, earned runs, strikeouts, saves, wins and losses. If you are interested, you can read more about the "Cy Young predictor" at <http://espn.go.com/mlb/features/cyyoung>.

5. **TURN IN:**

(a) One Excel file containing two worksheets. The contents of **Sheet 1** are described above. **Sheet 2** of your Excel worksheet should now contain the *Strikeouts* data and its summary statistics. Each sheet should contain your name in cell G1. Please submit your Lab 2 Excel file via e-mail.

(b) The filled-in answers to questions in Step 4 above. You only need to turn in the page where you wrote something. Please make sure **YOUR NAME** is at the top of the page. (Don't forget to include the handwritten answers from the **Getting Started** section too.)

PLEASE NOTE: As mentioned in Lab 1, the "formula view" of your Excel work will be graded (in addition to the computed values). To check 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.