

## MATH 140 Lab 3: Descriptive Statistics, Boxplots, Standardization

### Problem 1. SAT scores again!

In your MINITAB software, and using **File**, go to **Open Worksheet ...** and open the file **Ga.MTW**. Alternatively, you can obtain this file from the course web page. Just go to:

<http://www.csub.edu/~sbehseta/.Labs140.htm>

and click on the link: **Ga.MTW**. This is a particularly interesting data set because it gives us the verbal and math scores of 100 Northeastern University students along with their GPAs (columns 1-3 respectively).

You should have a worksheet open with three columns named "Verbal", "Math", and "GPA" from left to right. First, let's plot these data.

- (a) Create a boxplot for "Verbal" scores. To generate dot plots, in your MINITAB **Graph** menu, select **Boxplot...** Next, select "Verbal" and place it under **Y** in the **Graphs** box. Hit "OK". Interpret the resulting boxplot. Comment on the shape of the distribution and any outliers or extreme values detected by MINITAB.
- (b) Repeat the same procedure this time for "Math" scores.
- (c) Repeat the same procedure for "GPA".
- (d) From the **Graph** menu, click on **Histogram ...**. Select "Verbal" and place it under "X" mark. Hit "OK". You should have a histogram for "Verbal".
- (e) Generate a histogram for "Math".
- (f) Generate a histogram for "GPA".
- (g) Compare your histograma with your boxplots.
- (h) Go to **Graph** followed by **Boxplot ...**. This time select "Verbal" variable and place it in the first row under **Y**. Next, click on the second row under the **Y** mark. From the variable menu, select "Math" and place it in the second row. Next, click on **Frame**. Choose **Multiple Graphs ...**. Within the **Multiple Graphs** menu, select **Overlay graphs on the same**

**page.** Hit "OK" twice. You should have a side-by-side boxplot representation of "Verbal" and "Math".

- (i) Do not close the plot but minimize it.
- (j) Go to **Stat** → **Basic statistics** → **Display Descriptive Statistics ...**. Select "Verbal" and "Math" variables and hit "OK". Compare the mean, the median,  $Q_1$  (Lower Quartile),  $Q_3$  (Upper Quartile), standard deviation of "Math" and "Verbal"
- (k) MINITAB detected two outliers for "Math" and one outlier for "Verbal". Verify this yourself using  $1.5 \times IQR$  rule of thumb.

## Problem 2. CEO salaries

Consider the CEO salaries data:

Forbes magazine published data on the best small firms in 1993. These were firms with annual sales of more than five and less than \$350 million. Firms were ranked by five-year average return on investment. The data extracted are the age and annual salary of the chief executive officer for the first 60 ranked firms. In question are the distribution patterns for the ages and the salaries (reported in thousands). The data (CEO.MTW) can be obtained from

<http://www.csub.edu/~sbehseta/Labs140.htm>

Note that one ceo has declined to report his/her salary.

- (a) Obtain the descriptive statistics for "Age" and "Salary" of those CEOs.
- (b) Obtain the histograms for the two columns of this data set. Describe the shape of the histograms.
- (c) Obtain the boxplots for the two variables.
- (d) Summarize your findings.
- (e) Note in particular that the "Age" variable has a symmetric distribution. Let's try to obtain the  $z$ -scores for this variable. Remember that  $z = \frac{x - \bar{x}}{s}$  where,  $\bar{x}$  is the sample mean and  $s$  is the sample standard deviation.

- (f) Go to **Calc** and open the **Calculator**. Store results in column "C3". In the **Expression** box, type:  $(C1-51.47)/8.92$  and hit "OK". Name the newly generated column "standardized-age". What would you suggest that this new column "C3" represents?
- (g) In fact, there is a easier way of obtaining  $z$ -scores. Go to **Calc** and choose **Standardize**. Choose "C1" as the input column, and store the results in "C4". The  $z$ -scores appear. Notice the values are very similar to the ones in "C3" but slightly different. Can you comment on the difference?
- (h) Histogram the new column.
- (i) Calculate the descriptive statistics if the new column.
- (j) Draw a boxplot for this new column.
- (k) Comment on the previous three parts and in particular compare your findings with the results relating to the original "Age" variable.

### Problem 3. More on Standardization

Download the file Randomdata.MTW from: <http://www.csub.edu/~sbehseta/Labs140.htm>

This is a randomly generated set of 10000 data points. Later in this course, you will learn how to generate random data yourself. But for now answer the following questions:

- (a) Draw a histogram and boxplot for this data set.
- (b) Note in particular that data is symmetric! Obtain the descriptive statistics for this data set.
- (c) Now, we are going to verify 68-95-99.7 rule! To do this we need to refine the data for a while.
- (d) Now, go to **Manip** → **Code** → **Numeric to Numeric**. In the box **Code data from columns** select "C1". In the box **Into columns** select "C2". In the first row of **Original values** type -1:1 and under **New:** in the same row type 1. In the second row under **Original values** type 1:5 and under **New:** type 0. In the third row under **Original values** type -5:-1

and under **New:** type 0. Go to **Calc** → **Column Statistics** and select **Sum**. Hit "OK". Report your finding.

Note that in this exercise, we coded all the values that are within 1 standard deviation of the mean to be 1, and everything else to be 0. Hence we are simply counting the number of values that are within 1 standard deviation of the mean by summing up all values in "C2". What does the result of the sum tells you about the proportion of values that are within 1 standard deviation of the mean?

(e) Same process but slightly different:

In the box **Code data from columns** select "C1". In the box **Into columns** select "C3". In the first row of **Original values** type -2:2 and under **New:** in the same row type 1. In the second row under **Original values** type 2:5 and under **New:** type 0. In the third row under **Original values** type -2:-5 and under **New:** type 0. Go to **Calc** → **Column Statistics** and select **Sum**. Hit "OK". Report your finding.

(f) Same process but slightly different:

In the box **Code data from columns** select "C1". In the box **Into columns** select "C4". In the first row of **Original values** type -3:3 and under **New:** in the same row type 1. In the second row under **Original values** type 3:5 and under **New:** type 0. In the third row under **Original values** type -3:-5 and under **New:** type 0. Go to **Calc** → **Column Statistics** and select **Sum**. Hit "OK". Report your finding.

(g) We have just accomplished something very interesting! That is, we have verified the 68-95-99.7 rule!! Elaborate on this claim.

## Problem 4. Hurricane strengths

Download the file Hurricane.MTW. You can obtain the file from <http://www.csub.edu/~clam/math140f03.html>

This is a file of tropical storm strengths (in knots) and their minimum central pressure (in millibars) for all the storms that appeared in the Atlantic Ocean in the past 9 years.

- (a) Create boxplots for the wind reading for the tropical storms in 1996, 1997, and 1999 and put all of them in one single graph as in Problem 1.
- (b) It is conjectured that hurricane strengths are affected by El Niño and La Niña, which are ocean current changes in the Pacific Ocean. Given that 1997 is an El Niño year while 1999 is a La Niña year, can you say anything about the hurricane strength of those two years as compared to a regular year, 1996?
- (c) It is also conjectured that the strength of a hurricane is related to its central minimum pressure. In order to see this phenomenon, draw a scatter plot for the year 1996. Go to **Graph** and choose **Plot**. In Graph Variables, Choose “1996 Wind” under “Y” and “1996 Pressure” under “X”. Click ok.
- (d) Judging from the graph, do you see any relationship between the two variables? If there is a correlation, what kind of correlation is it?
- (e) Open the file Hurricane96.MTW. This is the data for the year 96, from the previous worksheet. Find out the  $z$ -scores for Wind and Pressure, store the results in “C3” and “C4”.
- (f) Calculate the values of  $z_x z_y$  and store the results in “C5”.
- (g) Now calculate the Pearson’s Correlation Coefficient from the values you obtained in “C5”. Use the calculator function provided in MINITAB.
- (h) Verify the results by using **Stat, Basic Statistics, Correlation...**, put ‘1996 Wind’, ‘1996 Pressure’ under “variables”. Click ok.