Your Name: ____________________________

Section: ____________________________

36-201 INTRODUCTION TO STATISTICAL REASONING
Computer Lab #3
Interpreting the Standard Deviation and Exploring Transformations

Objectives:

1. To review stem-and-leaf plots and their interpretations.
2. To develop an understanding of the standard deviation as a measure of spread.
3. To learn how to transform skewed data so that it looks more nearly normally-distributed.
4. To learn to use the file server to read data files.

Part I. Interpreting The Standard Deviation

*** * * Do Not Use Minitab to Answer These Questions * * ***

<u>Background</u>

After an examination teachers often report as summaries of the class performance the mean and the standard deviation of the distribution of exam scores. You may have wondered why. In this section we will suggest an answer. Specifically, we will see for symmetric, unimodal distribution how the mean and standard deviation tell you practically everything you want to know about the distribution. In addition, we will review creating stem-and-leaf plots and finding numerical summary measures.

The following are hypothetical scores for 25 students who took an examination. Assume there are a maximum of 100 points on this exam.

<u>Scores:</u> **80 71 85 70 93 78 62 80 94 71 71 78 74 48 86 77 68 70 79 79 64 61 82 58 76**

Question #1 Make a stem-and-leaf plot of these data. You should order the leaves.

♣**Question #2 Describe the features of the distribution of exam scores**.

Question #3 Find the median, Q_1, Q_3, and the IQR for this distribution. Show your work.

<u>THE 68–95–99.7 RULE</u>

There is a rule called *The 68-95-99.7 Rule* , also known as *The Empirical Rule* and *The Standard Deviation Rule*. The rule states that that: "In any bell-shaped distribution with mean, $\bar{x}$, and standard deviation, s, approximately:

- 68% of the observations fall in the interval between the values $\bar{x} - s$ and $\bar{x} + s$;
- 95% of the observations fall in the interval between the values $\bar{x} - (2 \cdot s)$ and $\bar{x} + (2 \cdot s)$; and
- 99.7% of the observations fall in the interval between the values $\bar{x} - (3 \cdot s)$ and $\bar{x} + (3 \cdot s)$."

For the distribution of exam scores, the mean is $\bar{x} = 74.2$ and the standard deviation is $s = 10.6$. For the following calculations you can use the **Calculator** which can be found in the menu, if you want.

Question #4 Find the values $\bar{x} - s$ and $\bar{x} + s$. Using your stem-and-leaf plot in question #1, how many observations fall between these two values? What percent of the distribution is this?

Question #5 Find the values $\bar{x} - (2 \cdot s)$ and $\bar{x} + (2 \cdot s)$. How many observations fall between these two values? What percent of the distribution is this?

Question #6 Find the values $\bar{x} - (3 \cdot s)$ and $\bar{x} + (3 \cdot s)$. How many observations fall between these two values? What percent of the distribution is this?

Question #7 Does *The 68–95–99.7 Rule* seem to be approximately correct for these data? Explain.

The 68–95–99.7 rule enables us to use the mean and standard deviation to make relatively simple calculations about a distribution. For example,

♣ **Question #8. Since the distribution of exam scores is bell-shaped, use the standard deviation rule to find the proportion of exam scores below 64. (Hint: Draw a unimodal, symmetric histogram for this distribution and show where $\bar{x} - s$ and $\bar{x} + s$ are located.) From your stem-and-leaf plot, how many observations are actually below 64?**

Part II. Transformations of Data

Getting Started

COPYING FILES

1. Open the **Student** folder. To do this, double click on the "Cluster HD" icon. Then double click the "student" folder. You will eventually move the data into this folder.
2. In the following order: From the menu in the upper left hand corner of the titlebar, choose **Chooser**. In the dialog box that appears *click* **AppleShare**. In the bottom left box called "AppleTalk Zones" *select* **BH**. In the top right box where it says "Select a file server" select **HSSHELIOS**. *Click* **Ok**.
3. In the dialog box that appears login using "36201" as your Name and "36201" as your password. Click **Ok**. In the dialog box that appears put a check next to "Class." Click **Ok**.
4. A "Class" icon will appear to the right. Double click on the icon. Then double click on 36201. You will see many files in a folder.
5. Copy all of the files that you will need to the **Student** folder. Do this by dragging the files into the Student folder. For today, the data file you will need is *population.dat*.
6. Close all windows.
7. **IMPORTANT:** Drag the *class* icon into the trash. This is necessary to let others in the lab get access to the data.

To start Minitab:

1. *Double-click* the "Cluster HD" icon.
2. *Double-click* the Applications icon.
3. *Double-click* the Minitab 10.5 folder.
4. *Double-click* Minitab 10.5.

CITY	DENSITY
Berkeley, CA	9480
Aurora, CO	7652
Stamford, CT	2689
Atlanta, GA	3244
Warren, MI	4684
Springfield, MO	2051
Albuquerque, NM	3481
Buffalo, NY	8561
Winston-Salem, NC	2173
Philadelphia, PA	12413
Irving, TX	1634
Lubbock, TX	1920
Pasadena, TX	2962
Arlington, TX	5894
Chesapeake, VA	637
Takoma, WA	3323

Table 1: Population densities (people per square mile) for 16 cities (Source: US Census, 1980).

Transformations of Data

Table 1 gives the population densities (people per square mile) as reported in the 1980 U.S. Census for 16 cities in the U.S. If you have already copied the data from the **H&SS File Server** you may now import it into column C10 of Minitab, using the directions below. Otherwise, just type it into column C10 directly from Table 1, since it is such a short data set.

1. In Minitab, pull down the **File** menu, and select **Import Text** under the **Other Files** option. You must identify the columns into which the data are to be read, by double-clicking on column names on the left, or by typing column names in at the top. There is just one column, so type C10. Then click on **OK**.
2. A new dialog box will appear with the heading **Minitab** followed by a list of files. Change the folder from **Minitab** to **Desktop**. A new set of folders, as well as the file `population.dat` that you copied from the H&SS server at the beginning of the lab, will appear.
3. Select the file `population.dat` and click on **Open** to import the data into Minitab. (You may have to tell the Mac to open the file in "Minitab 10.5 Xtra Power" at this point.) After a brief period, the data will appear in the Minitab spreadsheet.

Statisticians often transform data to try to change the shape of a distribution. Specifically, when a distribution is skewed we might try a transformation to make the shape of the distribution more symmetric. Recall, for example, that the 68%, 95%, 99.7% rule applies to a symmetric, unimodal distribution, but not to a skewed distribution. In this part of the lab, you will look at some transformations of data to try to get a feeling for the effect different transformations have on shape.

To begin with, sort column C10 and store the sorted data in C1 (**Sort** is under the **Manip** menu. Type C10 in the "Sort Column" block, type C1 for the "Store Sorted Columns in" block, and finally type C10 again in the first "Sort by Column" block. Then of course click **Ok**).

Use **Stat**, **Basic Statistics**, **Descriptive Statistics** to find basic descriptive statistics for column C1. Enter C1 in the "Variable" block, *and today select "Graphical Form" instead of "Tabular Form" near the bottom of the dialog box.* Click **Ok** and wait for a window summarizing the population data to pop up.

Question #9. Fill out the table of descriptive statistics below from the information in the "Descriptive Statistics" window for C1. (You will have to do the calculations indicated below.)

Measures of Location or Center	*Measures of Variability or Spread*
Mean=	StDev (SD)=
Min=	
Q1=	
Median=	IQR = Q3−Q1=
Q3=	
Max=	Range = Max−Min=

Questions #10. In the "Descriptive Statistics" window, Minitab has plotted both a histogram and a boxplot. Copy the boxplot from the "Descriptive Statistics" window into the space below.

♣ **Question #11. Based on the histogram and the boxplot, describe the distribution of populations per square mile, using as many words like "unimodal", "bimodal", "symmetric", "skewed left", "skewed right", "outliers" as you can in your description. (Don't forget to mention whether or not there are outliers!).**

Question #12. Notice that Minitab has also overlaid a normal curve on the histogram. Does it look like the histogram follows the normal curve well? Explain why or why not.

Question #13. In the space below, give evidence *for* or *against* using the the 68%, 95%, 99.7% rule with these data, using the evidence Minitab has given you.

Now let's investigate several transformations of the data to see if we can make it more normal. Generally, when the data are skewed to the right, one might consider a power transformation of the form $y = x^p$ with $p < 1$. Type the following in the **Session** window (or use the **Calc**, **Mathematical Expressions...** menus if you prefer):

- LET C2 = C1**(.667)
- LET C3 = C1**(.500)
- LET C4 = C1**(.333)
- LET C5 = LOGE(C1)

C2, C3, C4 and C5 contain transformations of the data using successively smaller powers[1]: $x^{2/3}$, $x^{1/2}$, $x^{1/3}$ and $\log_e(x)$.

[1]Notice that each tranformation is a power, x^p, for p between 0 and 1. The last transformation, $\log_e(x)$, corresponds to $(x^p - 1)/p$ as p get closer and closer to 0.

♣ **Question #14. Produce separate boxplots for C1, C2, C3, C4 and C5 and put rough sketches into the space below. Comment on what is happening to the shape of the distribution as the power p in the transformation x^p gets smaller and smaller. Which column—C1, C2, C3, C4, or C5—produces the most nearly-symmetric boxplot?**

You may have to select **Window**, **Close all graphs** *to free up enough memory for the next part of this data analysis.*

For the column you chose as best in Question 14, make graphical descriptive statistics as you did in Question #9. Then answer the following questions.

Question #15. Based on the histogram and the boxplot, describe the distribution of populations per square mile, using as many words like "unimodal", "bimodal", "symmetric", "skewed left", "skewed right", "outliers" as you can in your description.

Question #16. Notice that Minitab has again overlaid a normal curve on the histogram. Does it look like the histogram follows the normal curve well? Explain why or why not.

SUMMARY

Generally the power transformations are used to make positively skewed data more symmetric. These transformations take the relatively large values in the positive tail of a distribution and make them relatively smaller. However, if the original data has a skewed left tail too, power transformation will worsen this skewness. With mild positive skewness, a power of .5 or more should suffice, and a lower power may skew the left tail. For more severely right-skewed data, a smaller power will be needed, all the way down to the logarithm transformation. Using the computer, one can often find a good power transformation that will balance both tails and make the resulting distribution much closer to normal.

To quit Minitab, from the **File** *menu, choose* **Quit***. Do not save any files. Remember to delete any files and folders that you might have created.*
Turn in your cover sheet.