## Lesson 2 - Homework

* The following numbered homework assignments refer to the problems in the textbook*: An Introduction to Statistical Methods and Data Analysis*, e.g., 1.2 is problem 1.2 (found on p.14 in chapter 1 for the fifth edition and p.13 in chapter 1 for the sixth edition).
* The answers to the problems will be posted on the **Sunday after the due date**. To view the answers, click on the **Homework Solutions** link in the menu bar to the left.
* You may use minitab whenever appropriate and include the minitab output in your homework. To place Minitab output into your homework, highlight the portion of the Minitab printout you want, copy that to the clipboard, and then paste it into the document

**1.** (15 pts) Problem 3.21 of 6th edition (Problem 3.32 of 5th edition)

**2.** (10 pts) We know that trimmed mean is a better measure of central location than mean when there are extreme outliers. When the data has some extreme outliers, can we trim certain % of the data and then use the trimmed data set to carry on analysis to make inference about the central location? What about to make inference for the dispersion ? Answer yes or no to each question and explain briefly. (hint: when making inference, standard deviation of the estimate is important)

**3.** (15 pts) Problem 3.40 of 6th edition (Problem 3.55 of 5th edition) When answering part b, boxplot is not the best choice to plot the data since it shows median and IQR, not corresponding to mean and standard deviation asked in part a. Use another plot to plot the sample data to get a more complete illustration of the data. When answering part d, use a numerical measure to justify your answer.

**4.** (20 pts) Problem 3.43 of 6th edition (Problem 3.58 of 5th edition) When answering f, provide a relevant graph for the data to support your explanation.

**5.** (15 pts) For the data given in problem 4 (the previous problem), use minitab to draw the boxplot and answer the following questions:

1. Is the data left skewed, symmetric or right skewed?
2. What are the outliers?
3. Is the median closer to the lower quartile or the upper quartile? Does that indicate that the density of data between first quartile and the median is higher than the density of data between the median and the third quartile?

**6.** (10 pts) Problem 3.45 of 6th edition. (Problem 3.60 of 5th edition)

7. (15 pts) Get data from <http://lib.stat.cmu.edu/DASL/Datafiles/Singers.html>

1. Use minitab to find the descriptive statistics for each type of singers. For each case, does the approximate value of s give a good estimate of s?
2. Use minitab to draw the boxplots for these cases side by side. Comment on the central tendency and the dispersion for the four types of singers.

**indexReading:** *An Introduction to Statistical Methods and Data Analysis*, chapters 3.4, 3.5, and 3.6.

**Sample homework problem**:

**Sample problem 1:** In a packing plant, a machine packs carton with jars. The times it takes each machine to pack 10 cartons are recorded. The results, in seconds, are show in the following table:

|  |  |
| --- | --- |
| New machine | Old machine |
| 42.1 41.3 42.4 43.2 41.8  41.0 41.8 42.8 42.3 42.7 | 42.7 43.8 42.5 43.1 44.0  43.6 43.3 43.5 41.7 44.1 |

a. Compute the mean and standard deviation for the time to pack a carton for each machine.

b. Plot the data for each machine.

c. Describe the data for the two machines.

**Solution to the Sample problem 1:**

a. Enter the data into two columns called “New machine” and “Old machine”. Using minitab>Stat>Basic Statistics>Display Descriptive Statistics, one can get the following output:

**Descriptive Statistics: New machine, Old machine**

Variable N N\* Mean SE Mean StDev Minimum Q1 Median Q3

New machine 10 0 42.140 0.216 0.683 41.000 41.675 42.200 42.725

Old machine 10 0 43.230 0.237 0.750 41.700 42.650 43.400 43.850

Variable Maximum

New machine 43.200

Old machine 44.100

Thus, mean and standard deviation for new machine is 42.140 and 0.683. The mean and standard deviation for old machine is 43.230 and 0.750.

b. One good graph to plot the new and old machine data is a dotplot, which shows the complete data set from which one can get some idea about the central location and the dispersion.



c. We can see that new machine has smaller mean value and slightly smaller standard deviation than the old machine.

**Sample problem 2**: Problem 3.44 of 6th edition ( *problem 3.59 of 5th edition) and also an additional part e*. Construct the intervals and check whether the empirical rule applies to this data set.

**Solution to the Sample problem 2:**

a. Enter the data into a column called time and use minitab>Stat>Basic Statistics>Display Descriptive Statistics, one can get the following output:

Variable N N\* Mean SE Mean StDev Minimum Q1 Median Q3 Maximum

time 50 0 15.96 1.20 8.50 4.00 9.00 15.00 22.25 34.00

Thus, median is 15.00, mean is 15.96

To find the mode, one can get the stem-and-leaf diagram and find what value appears most often. According to the following stem-and-leaf diagram, 5 appears five times and is the mode.

**Stem-and-Leaf Display: time**

Stem-and-leaf of time N = 50

Leaf Unit = 1.0

7 0 4455555

10 0 667

14 0 8999

17 1 011

21 1 2223

(5) 1 44555

24 1 666777

18 1 89

16 2 111

13 2 23

11 2 555

8 2 677

5 2 99

3 3

3 3 33

1 3 4

b. To guess the value of s, that is to find the approximate value of s by range/4 which is (maximum – minimum)/4= (34-4)/4 =7.5

c. Compute s. Minitab has computed s for us and give that as st.dev. in the output. s= st.dev. = 8.5

d. No. To explain the reason, it will be useful to draw a boxplot of the data:



Note that the median is closer to the first quartile than to the third quartile and the upper whisker is longer than the lower whisker, the distribution is right skewed and not symmetric.

e. Construct the intervals and check whether the empirical rule applies to this data set.

The total number of observation = 50. Now, the stem-and-leaf diagram will enable us to find the following information easily:

|  |  |  |  |
| --- | --- | --- | --- |
| Interval | Number in Interval | Percentage | Empirical |
| = (7.46, 24.46) | 29 | 58% | 68% |
| = (- 1.04, 32.96) | 47 | 94% | 95% |
| = (-9.54, 41.46) | 50 | 100% | 99.7% |

These percentages do not match the Empirical Rule very well but it is not too bad.