## Lesson 2 – Homework (Daljeet Maken – dmm6393)

* 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)

1. Mean: 8.04, Median: 1.54
2. The values are:

Terrestrial Feeders: Mean: 15.01, Median: 6.03

Aquatic Feeders: Mean: 0.38, Median: 0.375

1. We can clearly see from the data and the above results that the mean is more sensitive to outliers than the median. The Terrestrial Feeders data set has most of its elements in single digit but there are two outliers in 76.50 and 41.70 that have an amplifying effect on the mean. The median of the Terrestrial Feeders at 6.03 is better representative of the underlying data. This is further reinforced by the Aquatic Feeders data set where the outliers are absent and we can clearly see that the median and mean are close to each other.
2. Following from the explanation above, we should recommend the median as a better representation of the central tendency of this data. The data for Terrestial Feeders indicates that the mean gets impacted by the outliers and takes on a value that is not representing the remaining data (excluding the outliers) very well. Therefore median because of its insensitivity to outliers is a better representation.

**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)

1. Yes, when the data has extreme outliers we can trim certain % of the data and then use the set to carry our analysis. This is an effective strategy to handle true outliers that are not representative of the underlying data set. There are different types of trimming strategies that are in use like: Trimmed Mean, Modified Mean (discard max and min value), Interquartile Mean etc. This is different as compared to truncation where the outliers are omitted from the data set for all types of analysis.
2. No, with the inference of dispersion we should try to maintain the dataset to have a good idea of its spread. Removal of a higher % of data will lead to loss of the spread information leading to bad inferences. The extreme example is that 50% trimmed data (which is the median) has 0 standard deviation since only one reading is left.

**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.

1. 
2. We will use two types of graphs
   1. Dot plots



* 1. Side by Side boxplots



1. The dotplot and boxplots above show that the supplier 1 and 3 have higher mean deviations as compared to supplier 2. Supplier 3 has the highest standard deviation too. The standard deviations of the lenses provided by S1 and S2 are comparable.
2. Supplier 2 provides material closest to the target value. The mean of the deviations of Supplier 2 are the least and therefore closest to the target. While supplier 2’s standard deviation is higher than the S1 it is only by a very small margin. The 95% data for supplier 2 is much below the 95% range for the other two suppliers:



**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.

1. Mean: 57.5, Median: 34
2. Median, since there are a few outliers at both ends (more at the higher end). Outliers have an impact on the mean while the median is resistant to outliers and therefore the median will be a better measure for central tendency.
3. Range = 273, s = 70.2
4. Approximate s as -> s ≈ range / 4 = 68.25. The approximation of 68.25 is fairly close to the actual 70.2 figure.
5. The table below (extracted from the excel workbook – also attached) provides the figures:



1. One reason for this might be evident from looking at the histogram of this data. The histogram looks like:



While the empirical rule holds for a normal distribution (with a single mound), this distribution doesn’t follow that distribution pattern. Therefore it is not in line with the empirical rule.

**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?



1. The boxplot indicates that the data is right skewed
2. The boxplot shows 3 outlier points corresponding to: 193, 226, 273
3. The median is closer to the lower quartile. Yes, that indicates 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)

1. Price per roll: s = 0.4233, Cost per sheet: s = 0.00552 (before the errata: 0.0059)
2. With simple comparison of s, we find that price per roll is more variable. However a better comparison requires coefficient of variation computed as:



The CV for the two is:



Here it actually becomes clear that the **cost per sheet is more variable.**

1. As it was shown in the above computation, the simple comparison of s leads to an incorrect inference. We should use CV since we are comparing considerably different entities and the magnitude of the population mean of the respective measure should be kept in context.

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.
3. The descriptive stats are:

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

Soprano 36 0 64.250 0.312 1.873 60.000 62.250 65.000 65.000 68.000

Alto 35 0 64.886 0.472 2.795 60.000 63.000 65.000 67.000 72.000

Tenor 20 0 69.150 0.719 3.216 64.000 66.250 68.500 71.750 76.000

Bass 39 0 70.718 0.378 2.361 66.000 69.000 71.000 72.000 75.000

The calculation of approximate s is below (attached spreadsheet)



The approximate s does provide a good estimate of s.

1. The boxplots are below:



* The central tendency of Soprano and Alto is at the same level with the median of both at 65. This is less than the median of Tenor (68.5) which is again less than the median of Bass (71)
* The median and Q3 of Soprano’s is overlapping indicating that the entire third quartile is made of a single value (65) that is also its median. The Alto has similar density in Q1 and Q3. The Tenor have higher density of data in Q1 (median closer to Q1) while the Bass have a higher density in Q3 (median closer to Q3).
* Alto and Tenor are slightly right skewed.

**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.