z

MA 2611-DL01 Applied Statistics I
D-Term Spring 2023 - Homework 02

Due: R 03/30 by 11.59 p.m.

I. Show all work as described in class. Partial credits will be given. Submit your solutions in a pdf format to canvas. Please write your name.

II. For the questions that are based on R programming, please copy, and paste the R plots and codes into a word file and then submit or submit the R markdown file.

1. According to the U.S. Census Bureau (census.gov), in 2016, the median sales price of new houses was $315,500 and the mean sales price was $370,800.
   (a) Interpret the median sales price.
   (b) Interpret the mean sales price.
   (c) Discuss the shape of the distribution of the price of new houses.

a). The median price of the new houses sold was $315,000. The median is usually considered to be the middle value of a sorted data set. Therefore, from this information, we can interpret that 50% of the houses sold were higher than the median price range. Most importantly, the median's value does not influence based on the extreme values of the data set, giving us a measure of central tendency.

b). The mean price of the new houses sold was $375,000, which could also be termed as the average price or the average value in the data set, which could be written as the sum of all observations divided by the total number of observations in the data set. Unlike the median, these values can be heavily influenced by extreme values, which could skew the data average but give us a gist of approximately for much each house was sold for.

c). We cannot determine the exact form of the distribution of the price of new homes without knowing the standard deviation or other measures of dispersion. On the basis of the variance between the mean and the median, we can, nonetheless, draw certain conclusions. The right-skewed nature of the price distribution is suggested by the fact that the mean sales price is greater than the median sales price. This suggests that a small number of really costly homes may be increasing the average price. If there are various subpopulations of houses with varying prices, the distribution could be either bimodal or multimodal.

2. Facebook's stock price in 2015 increased by 34.15%, and in 2016, it increased by 9.93%.
   (a) Computer the geometric mean rate of return per year from the two-year period 2015-2016.
   (b) If you purchased $1000 of Facebook stock at the start of 2015, what was its value at the end of 2016.

a). The Geometric mean rate of return per year can be calculated by taking the percentage values them and adding them to 1, and using the geometric mean formula to find the outcome. Therefore the solution is as follows:

$$2015: 1 + 0.415 = 1.3415 \qquad 2016: 1 + 0093 = 1.0993$$

$$\sqrt[2]{(1.3415)(1.0993)} - 1$$

$$0.2134 \; or \; 21.34\%$$

b). If the stock was purchased at $1,000 in 2015 and was held until 2016, the rate of change in value can be calculated as follows:

$$Ending\ Value = Beginning\ Value * (1 + Rate\ of\ Return)^{Number\ of\ Years}$$

$$Ending\ Value = 1{,}000 * (1 + 0.2134)^2$$

$$Ending\ Value\ (2016) = \$1{,}479.79$$

3. Please do it **by hand** – you will need to do so on the exam also. Use of calculator OK – it will be permitted on the exam.

Customer	1	2	3	4	5	6	7	8	9
Spending	96	75	84	89	95	73	59	82	115

(a) Calculate the range, mean, and sample standard deviation of the spending.

(b) Compute the coefficient of variance (CV).

a). The Range, Mean and Standard Deviation can be calculated as follows:

$$Range: 115 - 59 = 56 \qquad Mean: \frac{96+75+84+89+95+73+59+82+115}{9} = 85$$

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

where Σ is the sum symbol, xi is each individual value, x̄ is the mean, and n is the total number of values. Using this formula, we can calculate the variance as follows:

$$Variance = \frac{(96-85)^2 + (75-85)^2 + (84-85)^2 + (89-85)^2 + (95-85)^2 + (73-85)^2 + (59-85)^2 + (82-85)^2 + (115-85)^2}{9-1}$$

$$Variance = \frac{2067}{8} \qquad Variance = 258.375$$

$$Standard\ Deviation = \sqrt{258.375} \qquad Standard\ Deviation = 16.074$$

b). Calculating the Co-efficient of Variance is as follows:

The coefficient of variation (CV) is a relative measure of variability and is calculated as the ratio of the standard deviation to the mean, expressed as a percentage:

$$Co-efficient\ Variance = \frac{Sample\ Standard\ Deviation}{Mean} \times 100\%$$

$$Co-efficient\ Variance = \frac{16.074}{85} \times 100\%$$

$$Co-efficient\ Variance = 18.833\ \%$$

4. You and some friends have decided to test the validity of an advertisement by a local pizza restaurant, which says it delivers to the dormitories faster than local branch of a national chain. Both the local pizza restaurant and national chain are located across the street from your college campus. You define the variable of interest as the delivery time, in minutes, from the time the pizza is ordered to when it is delivered. You collect the data by ordering 10 pizzas from the local pizza restaurant and 10 pizzas from the national chain at different time. The sample statistics are summarized in the following table.

Restaurant	Mean	Standard Deviation
Local	16.70	3.0955
Chain	18.88	3.0955

(a) Determine the coefficient of variation of both restaurants.

(b) Are you agree with the advertisement? Explain your answer

a). The Co-efficient of variance for each restaurant can be calculated using the formula and can be calculated as follows:

$$CV = \frac{Standard\ Deviation}{Mean} \times 100\%$$

$$CV_{Local} = \frac{3.0955}{16.70} \times 100\% \qquad CV_{Local} = 18.51\%$$

$$CV_{Chain} = \frac{3.0955}{18.88} \times 100\% \qquad CV_{Chain} = 16.39\%$$

b). The local pizzeria appears to deliver pizzas more quickly than the national chain, with an average delivery time of 16.70 minutes compared to 18.88 minutes for the national chain, according to the mean delivery timings. The coefficient of variation indicates that there is a reasonably high level of variability in the delivery times for both restaurants, despite the fact that the difference between the means is not very significant. As there is a lot of overlap in the delivery timings for both restaurants, it is likely that the national chain will occasionally deliver faster even if the data points to the local pizza shop being faster on average. A higher sample size might produce different results because the sample size of 10 pizzas from each restaurant is rather limited.

Overall, it is challenging to draw firm conclusions from this data alone, but the results do point to the possibility that the neighborhood pizzeria may serve orders more quickly on average, even though there is considerable variation between the establishments' delivery times.

5. The following contains the average room price (in US$) paid by various nationalities while travelling abroad (away from their home country) in 2016:

124 101 115 126 114 112 138 85 138 96 130 116

(a) Compute the first quartile (Q_1), second quartile (Q_2), third quartile Q_3, and the interquartile range.

(b) List the five-number summary.

(c) Construct a boxplot and describe its shape.

a). Computing the First Quartile (Q1), Second Quartile (Q2), Third Quartile (Q3) and the Interquartile Range is as follows:

$$Sorted\ Data\text{: } 85, 96, 101, 112, 114, 115, 116, 124, 126, 130, 138, 138$$

$$Q_2 = 115.5$$

$$Q_1 = Median\ (85, 96, 101, 112, 114) \quad Q_1 = 106.5$$

$$Q_3 = Median\ (124, 126, 130, 138) \quad Q_3 = 128$$

$$IQR = Q_3 - Q_1 \qquad IQR = 130 - 102 \qquad IQR = 21.5$$

b). The five-number summary consists of the minimum value, the first quartile (Q1), the median (Q2), the third quartile (Q3), and the maximum value:

$$Minimum = 85$$

$$Q1 = 106.5$$

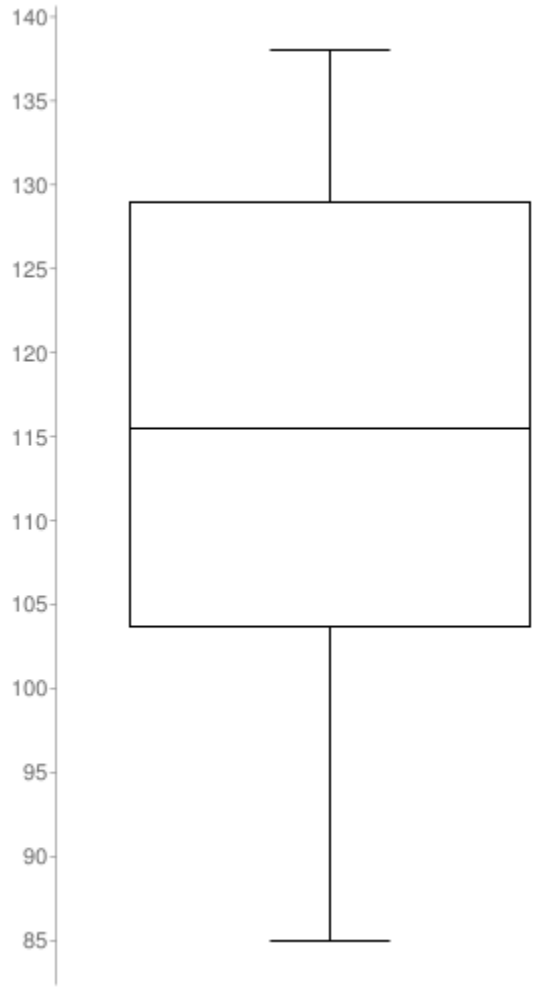
$$Q2 = 115.5$$

$$Q3 = 128$$

$$Maximum = 138$$

c). The five-number summary is represented graphically in the boxplot. The median (Q2) is depicted as a line inside the box, which reflects the middle 50% of the data (between Q1 and Q3). The

minimum and highest values which aren't considered outliers are covered by the whiskers that extend from the box. Outliers are values that deviate from the mean by more than 1.5 times the IQR and are shown as separate points. The boxplot in this instance would have a box spanning from 101 to 130 and a line at 115 designating the median. There wouldn't be any outliers, and the whiskers would range from 85 to 138. The box and whiskers of the boxplot would extend approximately equally in both directions from the median, giving it a roughly symmetrical shape. Extreme values or outliers that would significantly alter the boxplot's shape don't appear to exist.



6. Indicate each of the following as **TRUE or FALSE**:
   (a) A histogram can replace a bar graph wherever it appears.
   (b) The first quartile is the first value of data.
   (c) An outlier is a part of the overall pattern of a data set.
   (d) Mean and median are the same for any data.
   (e) Standard deviation measures the spread of data.

(a) FALSE. A histogram and a bar graph have different purposes and display different types of data. A histogram is used to display the distribution of a continuous variable, while a bar graph is used to display the values of categorical or discrete variables.

(b) FALSE. The first quartile is not the first value of data. It is the value that separates the lowest 25% of the data from the highest 75% of the data when it is arranged in order.

(c) FALSE. An outlier is a value that is significantly different from the other values in a data set. It is not part of the overall pattern of the data, but rather an extreme value that can skew the results.

(d) FALSE. The mean and median can be different for a data set, especially when there are outliers or the distribution is skewed.

(e) TRUE. The standard deviation is a measure of the spread or variability of data around the mean. It tells us how much the data deviates from the mean on average.

7. **Use R to do this question.**

Two teams compete in a robotics tournament. Their scores are:

Team 1 :27, 28, 18, 29, 30, 29, 33, 20, 25, 35, 25, 23, 26, 29, 33, 22, 47, 18, 20, 20
Team 2 :07, 08, 21, 28, 19, 17, 09, 18, 07, 05, 05, 05, 11, 33, 36, 19, 09, 13, 07, 10

(a) Enter the above data into R.
(b) Make a back-to-back boxplot. **(Hint: Use " boxplot" function in R)**
(c) Describe the "overall patterns" of both teams.
(d) Who won the tournament? Why?

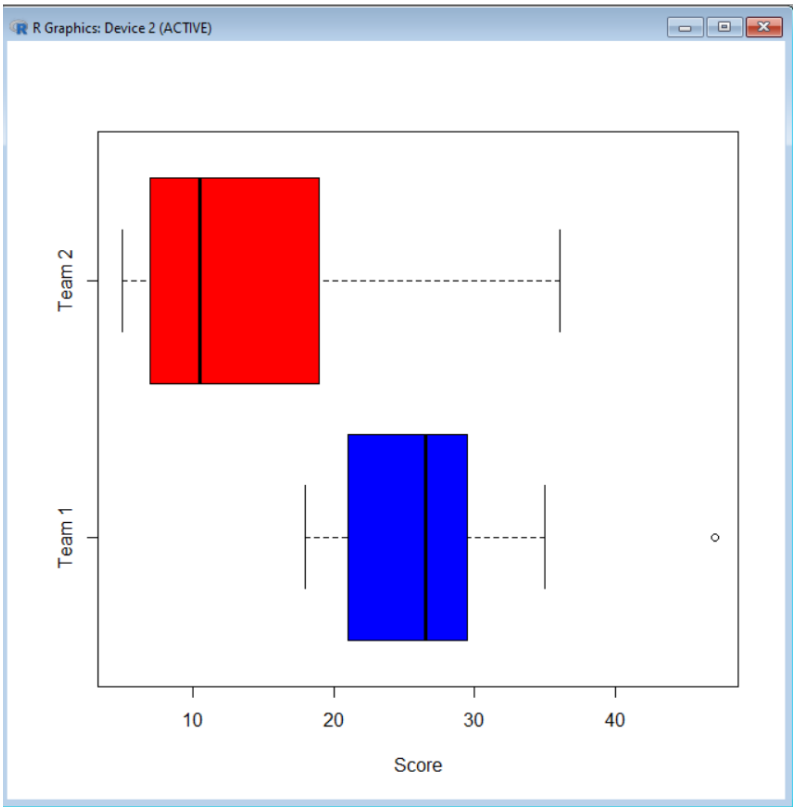
a). Enter the data into R

```
> team1 <- c(27, 28, 18, 29, 30, 29, 33, 20, 25, 35, 25, 23, 26, 29, 33, 22, 47, 18, 20, 20)
```

```
> team2 <- c(7, 8, 21, 28, 19, 17, 9, 18, 7, 5, 5, 5, 11, 33, 36, 19, 9, 13, 7, 10)
```

b). Make a back-to-back boxplot

```
> boxplot(team1, team2, horizontal = TRUE, xlab = "Score", names = c("Team 1", "Team 2"), col = c("blue", "red"))
```



c). Describe the “overall pattern” of both teams.

With a few outliers on the high end, the box for Team 1 is rather narrow and centered around the median score. This implies that Team 1's performance was largely constant, with a few standout results. With no outliers, the box for Team 2 is broader and tilted toward lower scores. This shows that Team 2's performance was less consistent, as shown by the broader range of their scores and the lack of any remarkable ones. In terms of scores and consistency, Team 1 seems to have fared better than Team 2 overall.

d). Who won the tournament why?

The data presented indicates that Team 1 won the competition. They outperformed Team 2 in terms of points and consistency, indicating that they were the overall better squad. It's crucial to remember that the data presented only contain results from one tournament and might not accurately represent the teams' overall performance of the team.