MATH 11: Discussion Week1

Apr. 2019

Displaying and Summarizing Quantitative Data

1. Mean and variance

Goal: Consider there are some midterm scores of students,

$$59, 67, 78, 75, 64, 89, 72$$

(1) Compute the **mean** of this data set(you can use calculator).
(Hint: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$*.)*
Solution:

$$\bar{x} = \frac{(59 + 67 + 78 + 75 + 64 + 89 + 72)}{7} = \frac{504}{7} = 72$$

(2) Compute the **variance** of this data set(you can use calculator). Could you also show the **standard deviation**? *(Hint:* $Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$*,* $SD(x) = \sqrt{Var(x)}$ *.)*
Solution:

$$\text{var}(x) = \frac{(59-72)^2 + (67-72)^2 + (78-72)^2 + (75-72)^2 + (64-72)^2 + (89-72)^2 + (72-72)^2}{7-1} = 98.6667$$

$$\text{SD}(x) = \sqrt{\text{Var}(x)} = 9.9331$$

2. Median and inter-quartile range(IQR)

Consider the data set given above
(1) Compute the **median** of your data set.
Solution: First sort the data as

$$59, 64, 67, 72, 75, 78, 89$$

median = 72
(2) Compute the **lower quartile** Q_1 and **upper quartile** Q_3 of your data set.
Solution: Since the total number of data is 7, which is odd. Add the median to the both halves of the data.
lower half: 59, 64, 67, 72. $Q_1 = \frac{64+67}{2} = 65.5$
upper half: 72, 75, 78, 89. $Q_3 = \frac{75+78}{2} = 76.5$.
(3) What is the **IQR** of this data set? *(Hint:* $IQR = Q_3 - Q_1$*)*
Solution: IQR = Q_3 - Q_1 = 76.5-65.6=11

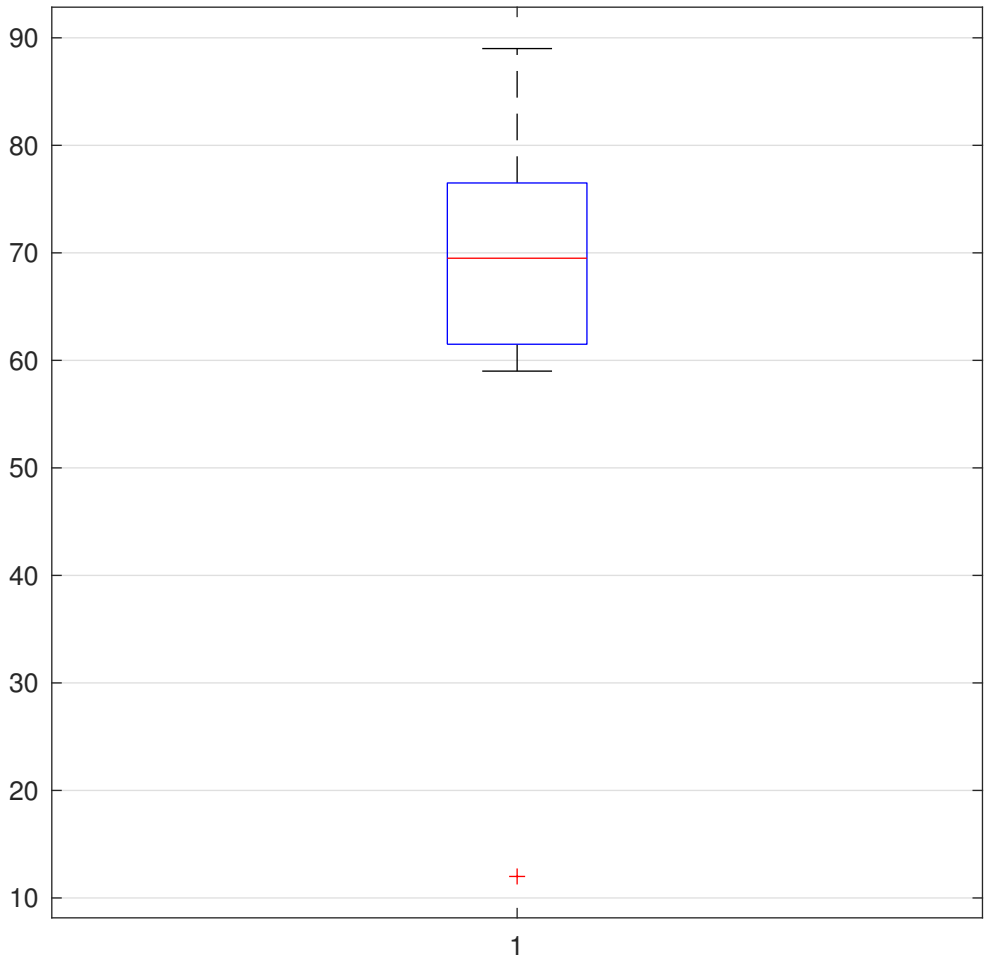

Figure 1: Boxplot

Consider a new data set given below

$$59, 67, 78, 75, 64, 89, 72, 12$$

(1) Recompute median, Q_1 and Q_3 again.
Solution: First sort the data set as

$$12, 59, 64, 67, 72, 75, 78, 89$$

median $= \frac{67+72}{2} = 69.5$. since it's even number data set:
lower half: 12, 59, 64, 67. $Q_1 = \frac{59+64}{2} = 61.5$
upper half: 72, 75, 78, 89. $Q_3 = \frac{75+78}{2} = 76.5$.
(2) Draw the boxplot (notice the outlier). (you need to recompute the IQR and the upper/lower fence.)
Solution: lower fence $= Q_1 - 1.5IQR = 39$.
upper fence $= Q_3 + 1.5IQR = 99$.
Boxplot is shown in Figure 1.

(3) Consider this new data set, given that the new mean and variance are 64.5 and 534.5714. Could you compare the result when you are using mean and variance, median (69.5) and IQR (15)

to describe the data set?
Solution: Since there is an outlier, the computation of mean and variance will be significantly affected by the outlier. Median and IQR could give a better representation of the data set.

3. Correlation

(1) Determine that the following statements about the correlation is **true** or **false**:

$$Hint: r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

- **True**: The range of correlation r is $-1 \leq r \leq 1$.
- **True**: The + sign of correlation r indicates the positive associations.
- **True**: The stronger association occurs when r is closer to 1 or -1.
- **True**: The choice of predictor/response doesn't matter, i.e. $\text{cor}(X, Y) = \text{cor}(Y, X)$.
- **True**: The correlation is unaffected by linear scale changes, i.e. $\text{cor}(X, Y) = \text{cor}(X, 2Y) = \text{cor}(X, Y + 5)$.

(2) Could you provide two examples to illustrate that correlation is not causation?

- Weight and height are correlated (positive trend), but increasing weight doesn't mean that people can get taller.
- In the summer, the sales amount of ice-cream and the number of sunburns reported are both increasing, but doesn't mean that they have causal relation.