CS 108 - Statistics, Sections B

Fall 2019, AUA

Homework No. 03

Due time/date: Section B: 10:32 AM, 20 September, 2019

Note: Please use **R** only in the case the statement of the problem contains (R) at the beginning. Otherwise, show your calculations on the paper. Supplementary Problems will not be graded, but you are very advised to solve them and to discuss later with TA or Instructor.

Problem 1, Measures of the Central Tendency

a.

We are given the dataset

$$2, 2, 2, 5, 3, 2, 0, 0, 3, 5.$$

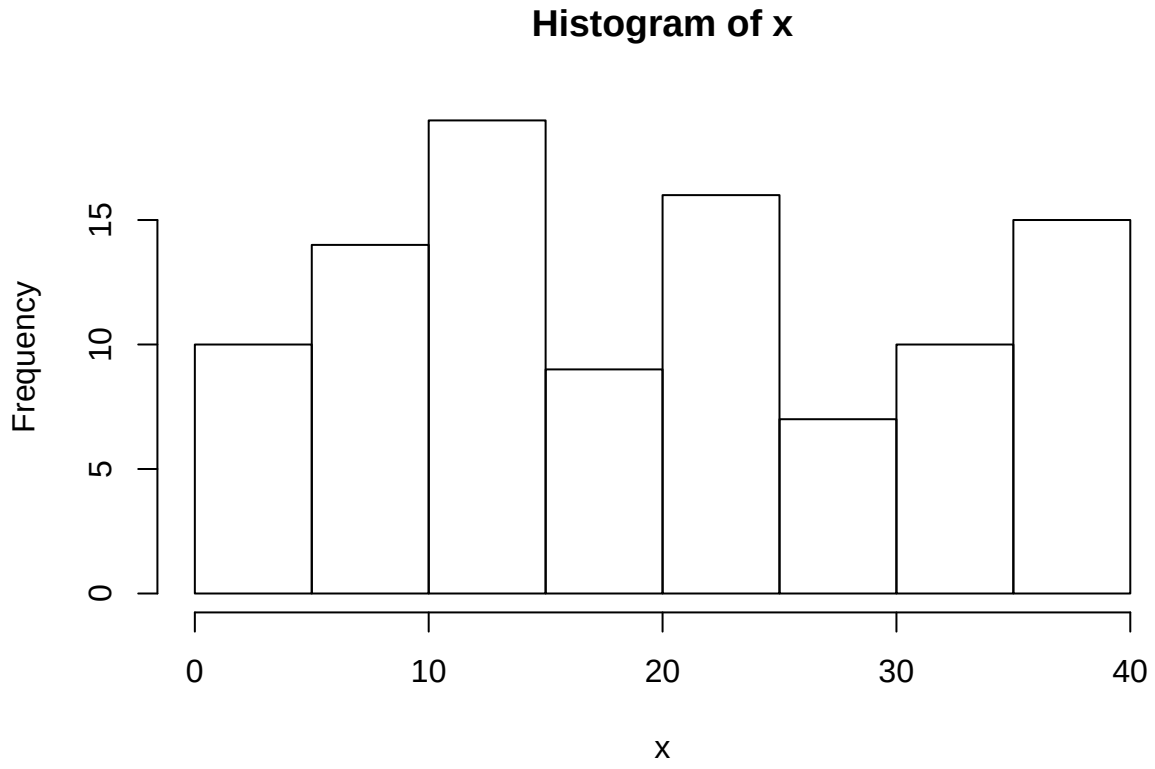
- Calculate the Sample Mean, Median and Mode of this Dataset.
- Find the 25% Trimmed and Winsorized Sample Means of this Dataset.

b.

- Construct a Dataset x of size 6 with $\bar{x} = -3$ and $median(x) = 5$.
- (Supplementary) Construct a Dataset x of size 10 with $\bar{x} = -3$, $median(x) = 5$ and a $mode(x) = 4$.

c.

Here is a histogram of some Dataset x:



Calculate, approximately,

- $mean(x)$
- $median(x)$

Explain your calculations.

d. (R)

Write an **R** code to calculate the Winsorized Mean of given vector.

- Your function need to take 2 inputs - the Dataset x and the number of elements to be replaced from the both ends of the sorted array, p. You need to check if p is appropriately chosen, otherwise your code need to give an error. The output need to be the Winsorized Mean of x.

Note: I suggest to use named variables, say the call of your function can be $winmean(data = ..., drop = ...)$

- (Supplementary) Your function need to be of 3 arguments - the Dataset x, the number of elements to be replaced from the both ends of the sorted array, p, and the ratio r of elements to be replaced from the both ends of the sorted array. Your function need to work if x and p or x and r are given (or if all three are given). If p is given, your code need to calculate the Winsorized Mean as above (so even if r is given, your code need to ignore it). If only x and r are given, then your code need to calculate p and then do the Winsorized Mean Calculation.

e. (R)

We again consider the `ChickWeight` Dataset from **R**.

- Calculate the Mean of Wights for chicken fed with the first diet;
- Calculate the Mean of Wights for chicken fed with the second diet;
- Compare the results: can the difference between the means bea result of just randomness, or we can state that one of the diets is better than the other one?

f. (Supplementary)

Assume x is a 1D numerical Dataset. Assume also that x has a uniques mode, and

$$mean(x) = median(x) = mode(x).$$

Is it true that x is symmetric[1]? Prove or give a counterexample.

g. (Supplementary)

Assume for the Dataset x we have only its Frequency or Relative Frequency Table (say, x_k are unique values and f_k are the corresponding frequencies/relative frequencies). Express $\bar{x}$ in terms of that unique values and frequencies/relative frequencies.

Problem 2, Measures for the Spread/Variability

a.

For the Dataset

$$x: \ -1, 3, 4,$$

calculate

- The $range(x)$
- The Sample Variance $var(x)$, using n in the denominator;
- The Sample Standard Deviation $sd(x)$, using $n-1$ in the denominator;
- The Mean Absolute Deviation $MAD(x)$ from the Mean.

b.

If I will generate two samples of size 150 from the $Unif[-1, 3]$ and $Exp(0.5)$, for which case I will (mostly) get larger Sample Variance? Why?

[1] btw, what is a symmetric Dataset?

c. (R)

Calculate and compare the Sample Standard Deviations and Variances for the `mpg` variable from the Dataset `mtcars` for different cylinder type cars. For example, compare 6 cylinder cars mpg-s SD with the 4 cylinder cars mpg-s SD.

d. (R)

We consider the `iris` Dataset. For which type of the flower (for shich Species) the variability in `Petal.Width` is maximal, and for which is minimal?

e. (R)

The **R** function `mad` computes **The Median Absolute Deviation** from the Median.

- Calculate the Median Absolute Deviation from the Median for the `dist` variable of the `cars` Dataset;
- Calculate the Median Absolute Deviation from the Mean for the `dist` variable of the `cars` Dataset (see the documentation of the `mad` function, you can change the `center` parameter);
- Write a function `mad1` which will calculate the Mean Absolute Deviation from the Mean. Test it on the same Dataset as above;
- Write a function `mad2` which will calculate the Mean Absolute Deviation from the Median. Test it on the same Dataset as above;
- (Supplementary) Join the previous functions into one, so that the user will be able to choose the Center Measure

Problem 3: Quartiles

a.

For the Datasets

$x:\ -6,\ 15,\ 0,\ 5,\ 17,\ -4,\ 1,\ -9,\ -9,\ 13, \qquad y:\ 0.0,\ 3.6,\ 2.7,\ -1.5,\ 5.7,\ 1.5,\ -3.0,\ 4.5,\ 6.0$

- Calculate all three Quartiles;
- Calculate the IQR;
- Check if Datasets have outliers (in the BoxPlot sense)

b.

- Is it possible to have a Dataset of size 4 with 2 outliers? Prove that it is not possible or give an example.
- I have a Dataset x with 80 elements, and the result of the `summary` command is the following:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.9035  0.1951  2.6917  3.0553  5.8178  9.7739
```

Assess, approximately, how many Datapoints in which interval we have.

- Our Dataset x is of size 120. Is it possible that 67 elements of x are outliers (in the BoxPlot sense)? Explain.

c. (R)

- Calculate the Quartiles of x and y form part **a.** by using the `quantile` function of **R**;
- Write an **R** function `quartile(x)` which will return the Quartiles of the input vector x just like we have defined. Test it on the Datasets x and y from the part **a.** of this Problem.

Problem 4, Boxplot

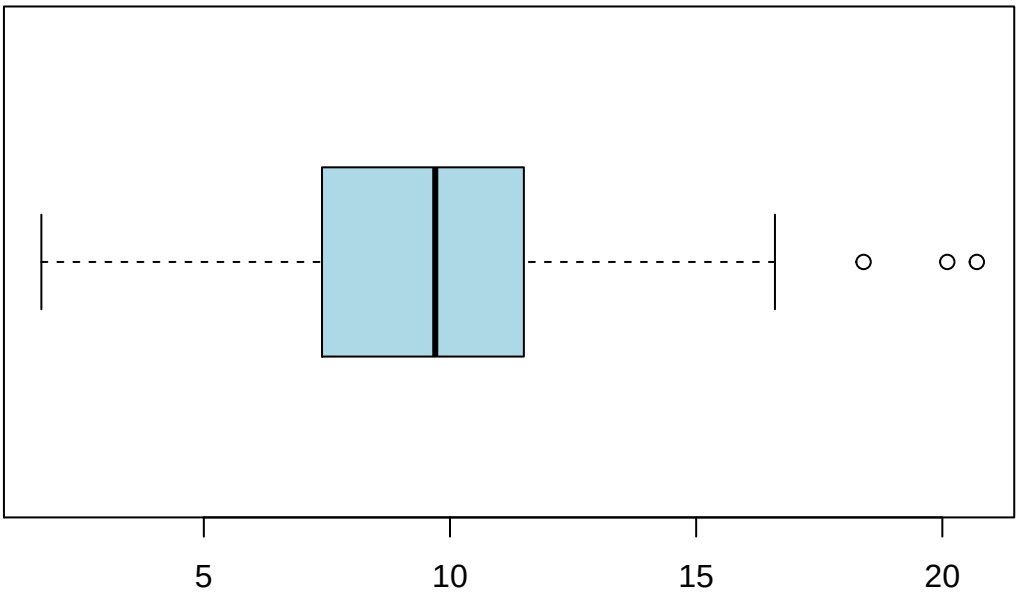
a.

Construct (with calculations) the BoxPlot for the Dataset

```
##  [1]  25 -10   3   1   2   8   4   0  -1   7   7   2  -1   2  -6   5   0
```

b.

Here is a Boxplot of some Dataset:



Give all possible information about the DataSet you can read from this BoxPlot.

c. (R)

Construct the Boxplot of the part **a.** Dataset using **R**, in a horizontal position, with the green color.

d. (R)

Construct, on the same graph, the Boxplots for the `Petal.Width` variable for each type of the Species in the `iris` DataSet. Give some information you can read from this comparative plot.

Note: You can use the following code:

```
boxplot(Petal.Width~Species, data=iris, horizontal = T)
```

Problem 5, Quantiles

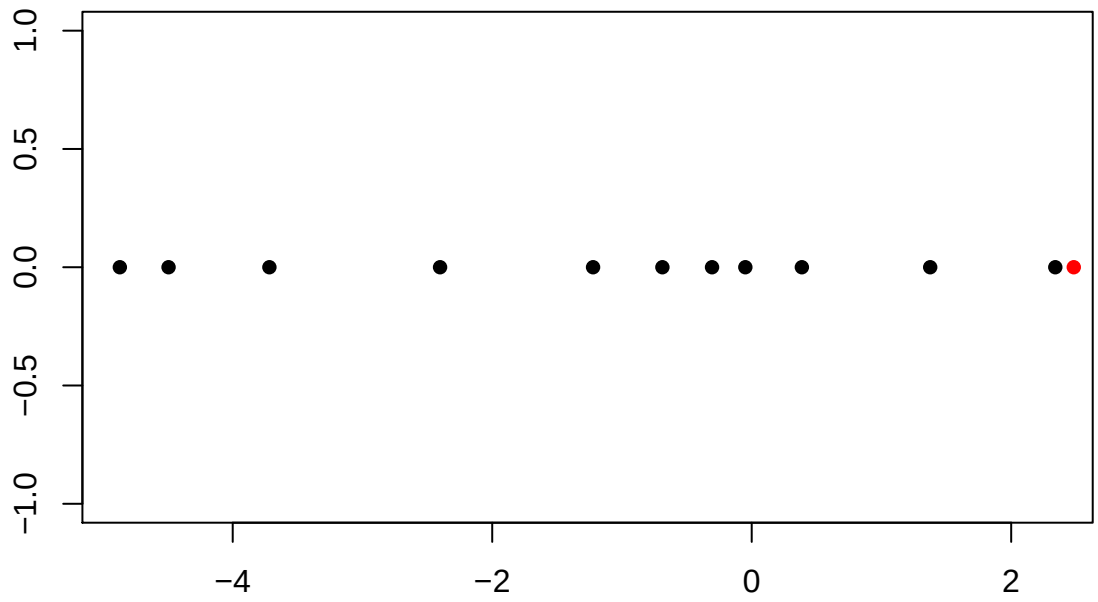
a.

Find the 10%, 15%, 80% quantiles (using our lecture definition) of the following Dataset:

$$x: \ -1, 2, 3, 2, 0, 2, 1, -1, 1, 5, 4$$

b.

I have a DataSet x, and some part of the DataPoints is shown in the graph below. The red point is the 28% quantile, calculated using **R**.



- What is the approximate size of my DataSet?
- Is it possible to approximately recover the rest of my Dataset?

c. (R)

Write an **R** function which will be calculate the Quantiles of a vector as we have defined during the lecture.