Applied Statistic with R

Fall 2019, ASDS, YSU

Homework No. 02

Due time/date: 9:28 PM, 25 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, ECDF

a.

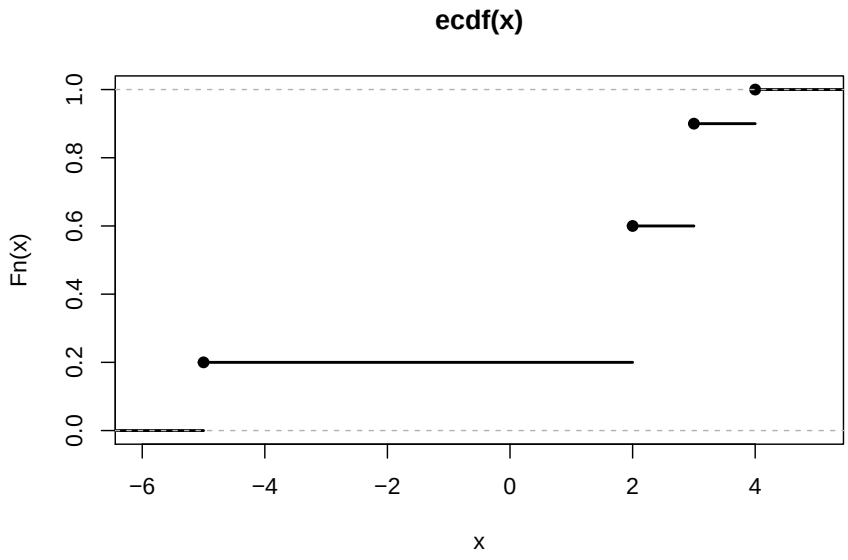
We are given the dataset

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

Construct the Empirical CDF (ECDF) for this Dataset.

b.

Below is the graph of the ECDF of some Dataset with 20 elements:



Reconstruct the Dataset. Is the information about the number of elements in the Dataset necessary? Why?

c. (R)

We want to simulate 100 Die Rolls in **R**. To that end, we can use the **R** command `sample`. Say, `sample(1:6, 3, replace = T)` will randomly choose 3 times an element from 1:6 (i.e., from the set {1,2,3,4,5,6}), with replacements (i.e., the same element can be chosen again). Here is an example:

```
sample(1:6, 3, replace = T)
```

```
## [1] 1 2 2
```

Now, simulate 100 Die Rolls in **R**. Let `res` be the result.

- Print the Frequency Table of `res`
- plot the ECDF of `res`
- plot the histogram or barplot of `res` (the one which is more appropriate for this case)

d. (R)

Now, we want to check that ECDF approximates well the CDF behind the data. To that end,

- generate 1000 samples from the $Exp(0.3)$ distribution;
- plot the ECDF of the result;
- plot over the previous graph the theoretical CDF of the distribution, with green color and linewidth 2 (use the `lwd=2` parameter value in `plot`)

Note: You need to adjust the axis scales for both graphs.

Problem 2, Histograms and KDE

a.

Consider the following Dataset x:

```
##  [1] -4.7  5.7  5.4  4.3  6.1  5.9 -1.8 -3.8  2.0 -3.7  7.6  8.6 -0.8  3.3
## [15]  8.9  5.5  8.9  2.4  4.6  4.5
```

Break the range of x (or some interval containing the range) into 5 equal-length bins and construct 3 Histograms: Frequency, Relative Frequency and Density.

b. (R)

Consider one of the standard Datasets in **R**, `islands`.

- call the help page for this Dataset to see the description
- print the structure of the Dataset
- print the head of this Dataset
- plot the Frequency Histogram for the islands with the area less than 200 sq miles
- plot the Density Histogram for the islands with the area less than 200 sq miles
- add to the previous plot the KDE (in red, with linewidth 3) for the islands with the area less than 200 sq miles
- add also Datapoints to the graph

c. (R)

Here we want to check that the Density Histogram is approximating well the PDF behind the data. To that end, we consider the *Weibull* distribution (see Wiki).

- Take $n = 1000$
- generate a sample of size n from the Weibull distribution with the shape parameter 2 (see `rweibull` and its parameters in **R**)
- plot the Density Histogram of that sample, in cyan color
- plot the theoretical PDF (use `dweibull` in **R**) over the previous graph, in red, and with linewidth 3.

Note: Adjust the scales of axes for both graphs!

d. (R)

Now let's plot compatarive Histograms. We will work with the **R**-s default `ChickWeight` Dataset.

- Explore the Dataset: read the description and print the first 5 rows of that Dataset;
- Separate in x the `Weight` variable for all Chicken with the `Diet` 1;
- Separate in y the `Weight` variable for all Chicken with the `Diet` 2;
- Plot the Frequency Histograms of x and y one over another. You can use transparent colors to make your graphs nicer. For that, you can use the `scales` library's `alpha` command:

```
library(scales)
hist(x, col = alpha("magenta", 0.2))
```

This will draw a histogram of x with transparent magenta color.

- What can be deduced from the Histograms?

Problem 3: Steam and Leaf Plot

a.

Consider the following Dataset:

```
##  [1] 1.2 4.6 0.3 2.9 2.3 0.8 0.6 1.1 2.4 3.8 5.0 2.3
```

- make the S-n-L Plot of this Dataset, and give the key (i.e., explain the position of period wrt `|`)
- make the S-n-L Plot of this Dataset with the follwing smaller "bins": $[0, 0.5)$, $[0.5, 1)$, $\ldots$

b.

Here is a S-n-L Plot drawn by **R** (no roundings were made):

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 | 26609
##   2 | 221224
##   4 | 1635
```

Reconstruct the Dataset.

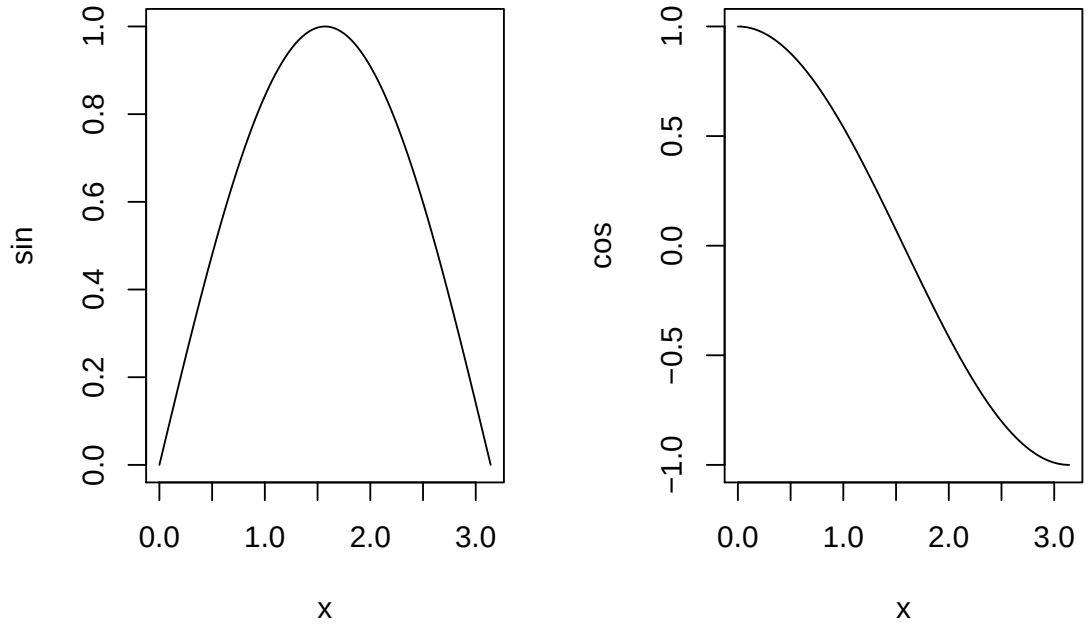
c. (R)

Consider the `iris` Dataset.

- Choose the `Petal.Length` variable and make its S-n-L Plot
- Now do the same variable S-n-L Plot with the `scale` parameters equal to 0.5, 2 and 4
- (Supplementary) Now plot the S-n-L Plot and Histogram of our Dataset side-by-side.

Note: To plot 2 figures side by side, you can use `par(mfcol = c(1,2))` parameter value befor doing the plotting. The command says: draw Multiple Figures Columnwise, 1 rows and 2 columns. For example,

```
par(mfcol=c(1,2))
plot(sin, 0, pi)
plot(cos, 0, pi)
```



See more at DataCamp.

Note: R-s `stem` output is not a graph. To make a graph, you can use the following code[1]:

```
x <- rnorm(10) # Just a random Sample
plot.new()
out <- capture.output(stem(x))
text(0,1, paste(out, collapse='\n'), adj=c(0,1), family='mono' )
```

Problem 4, ScatterPlot

a. (R)

Plot the following points:

$$(0,2), \quad (3,-1), \quad (4,2), \quad (5,5), \quad (-1,2)$$

b. (R)

R-s `pressure` Dataset consists of 2 Variables. Give the ScatterPlot of these Variables.

[1] Found from StackOverflow

Problem 5, Apple Stock Weekly Returns Histogram (R)

Go to Yahoo Finance page, navigate to the Apple Stock page (Apple's symbol (ticker) is AAPL, make a search for it), then choose Historical Data, 5 years time period, and weekly frequency. Download that Data. It will be in .csv format.

- Using the **R** `read.csv` command, extract the Adjusted Close Prices ("Adj Close" column), calculate weekly returns of the Apple stock [2].

Note: To read a `.csv` file into a DataFrame, you can use the following:

```
aapl <- read.csv(file.choose())
```

Instead of `file.choose()` you can give the exact path of your downloaded `.csv` fiel. But I prefer to have an `Open` dialog instead.

- Plot the histogram of weekly returns;
- Describe the results.

Note: To be able to run your code, please attach also the `.csv` file.

Problem 6, 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 $mode(x) = 4$.

[2] A return for some time period is the ratio

$$\text{Return} = \frac{\text{Last Price - First Price}}{\text{First Price}},$$

where the Last Price is the price at the end of the period, and the First Price is the price at the beginning of the period. So the return shows the percentage change during that period:

$$\text{Last Price} = \text{First Price} \cdot (1 + \text{Return}).$$

c.

Prove that for any 1D numerical Dataset $x: \ x_1, x_2, ..., x_n$, and for any real numbers α, β,

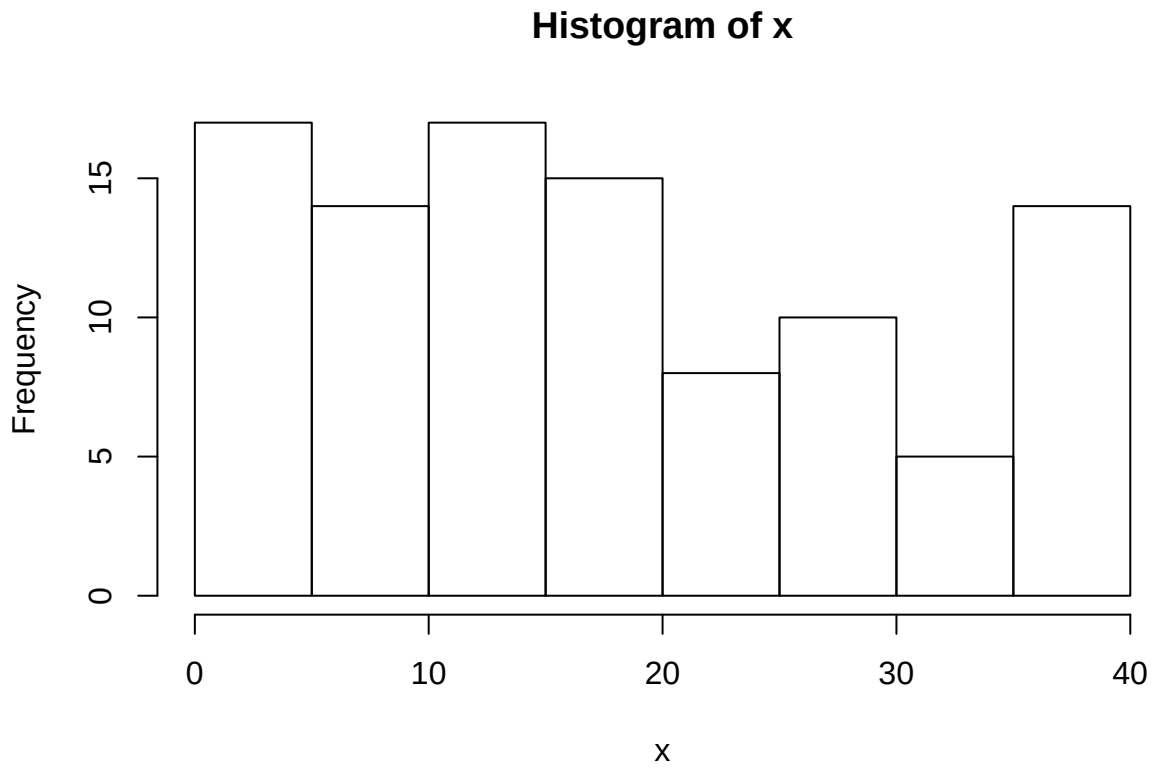
- $x_{(1)} \leq \bar{x} \leq x_{(n)}$;
- for any weights $w_1, ..., w_n$,

$$x_{(1)} \leq weighted\ mean(x; w) \leq x_{(n)};$$

- $mean(\alpha \cdot x + \beta) = \alpha \cdot mean(x) + \beta$;
- Does the previous property hold true for the *median*? Prove or give an example.

d.

Here is a histogram of some Dataset x:



Calculate, approximately,

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

Explain your calculations.

e. (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.

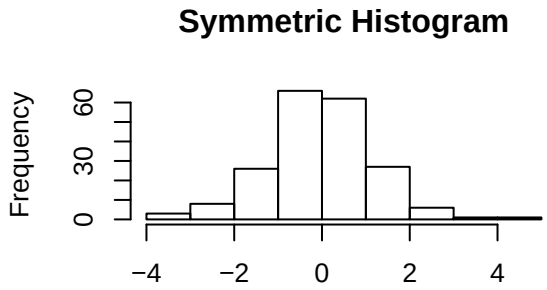
f. (R)

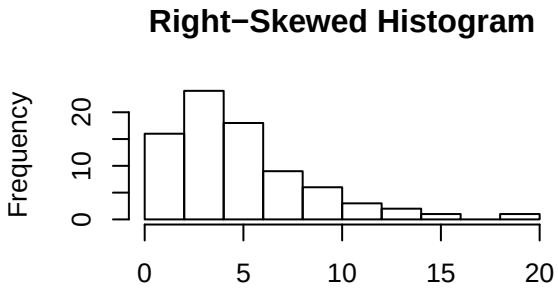
We again consider the `ChickWeight` Dataset from **R**.

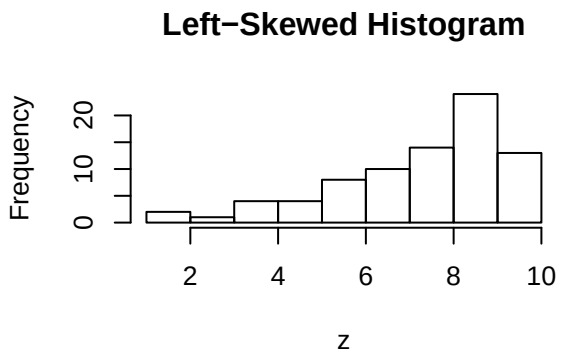
- Calculate the Mean of Wights for chicken fed with the first diet;
- Calculate the Mean of Wights for chicken fed with the second diet;
- Compare the results: can the difference between the means bea result of just randomness, or we can state that one of the diets is better than the other one?

g. (Supplementary)

We have three histograms for Datasets x, y, z and a







For each case, approximately, deduce if $median < mean$, $median = mean$ or $median > mean$. Explain you choice.

h. (Supplementary)

Is the following true for any 1D numerical Dataset x? Prove or give an example:

- $trimmed\ mean(x) \leq mean(x)$;
- $winsorized\ mean(x) \leq mean(x)$;
- $mean(x) \leq median(x)$

i. (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[3]? Prove or give a counterexample.

[3] btw, what is a symmetric Dataset?

j. (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 7, 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?

c.

Assume x is a Dataset of size n, and β is a real number. Here we denote by $var(x)$ the Sample Variance of x with n in the denominator, by $sd(x)$ the Sample Standard Deviation with n in the denominator, and let $mad(x)$ be the Mean Absolute Deviation of x from the Mean.

- Prove that $var(x + \beta) = var(x)$
- Assume $n = 2$. Compare $sd(x)$ and $mad(x)$ - can we state that one of these measures is always larger than (or equal to) the other one? Prove your statement.
- (Supplementary) Solve the previous one with general n.
- (Supplementary) Is it true for any Datasets x and y of the same size that $var(x + y) = var(x) + Var(y)$?
- (Supplementary) Is it true for any Datasets x and y of the same size that $mad(x + y) \leq mad(x) + mad(y)$? Here mad is the Mean Absolute Deviation from the Mean.

d. (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.

e. (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?

f. (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 8: 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.
##      -4       0       3       3       6      10
```

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)?

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.