CS 108 - Statistics, Sections B

Fall 2019, AUA

Homework No. 02

Due time/date: Section B: 10:32 AM, 13 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.

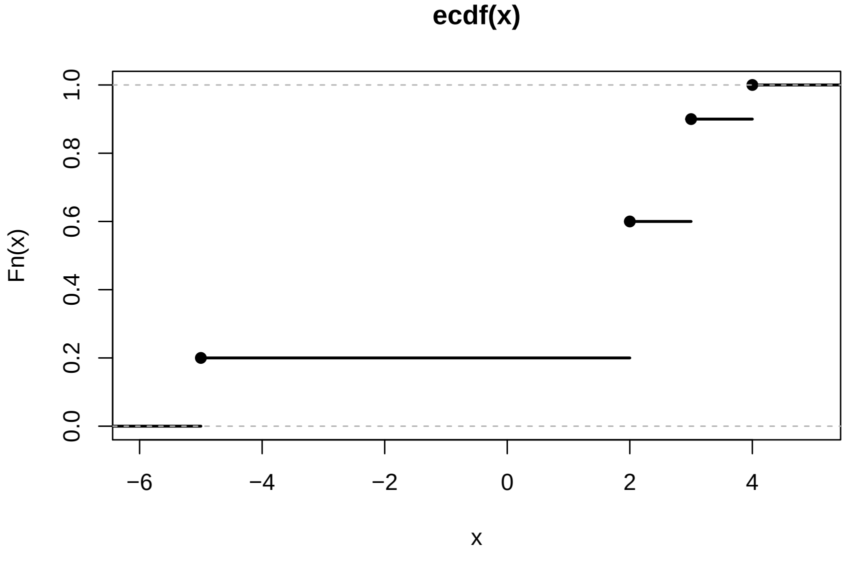
We are given the dataset

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

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] 3 3 6
```

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]  9.103  3.291 -3.829  4.382  3.118  4.140  1.103  4.745 -0.007  1.226
## [11]  3.550  0.159 -0.177  0.819  9.962  3.507  3.715  3.953  1.803  8.446
```

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:

```r
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] 4.6 0.8 3.4 1.3 4.8 4.5 4.0 3.3 2.8 3.7 4.1 2.8
```

- 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": $[0, 0.5)$, $[0.5, 1)$, ...

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.

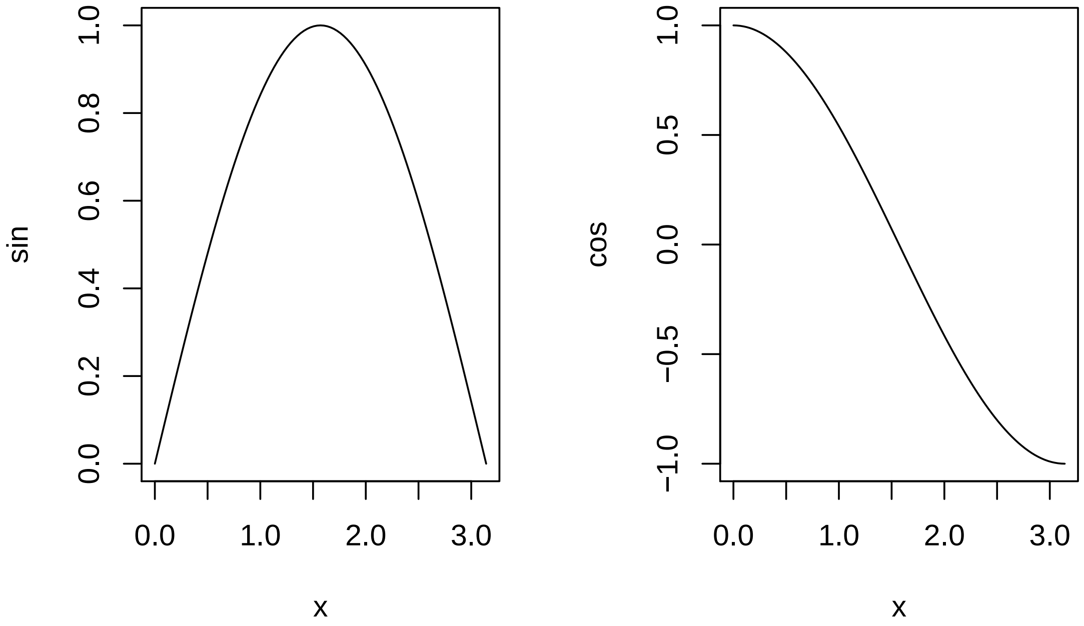
c. (R)

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,

```r
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.

[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:

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