Applied Statistic with R

Fall 2019, ASDS, YSU

Homework No. 03

Due time/date: 9:28 PM, 12 October, 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, Boxplot

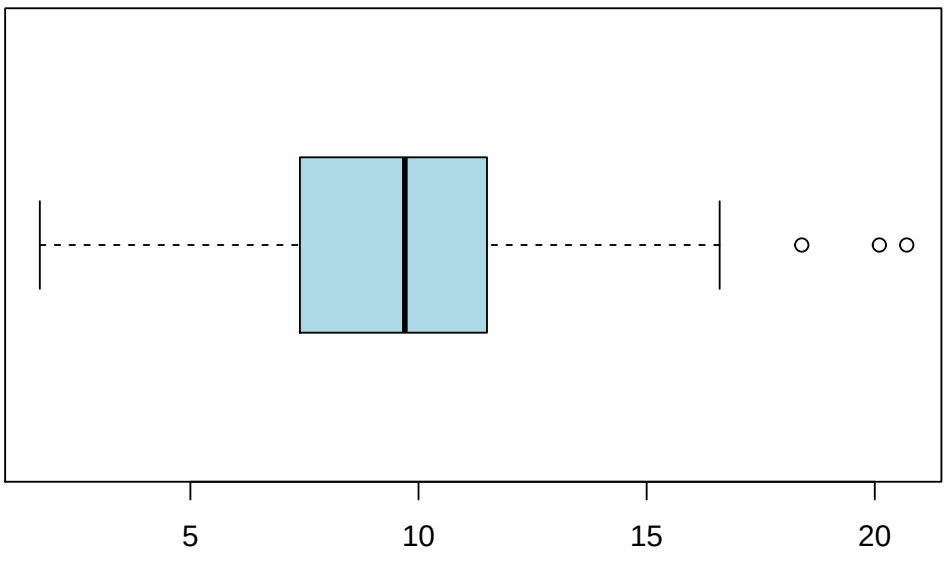
a.

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 2, Sample Quantiles

a.

Find the 15% quantile (using our lecture definition) of the following Dataset:

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

b. (R)

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

Problem 3, Theoretical Quantiles

a.

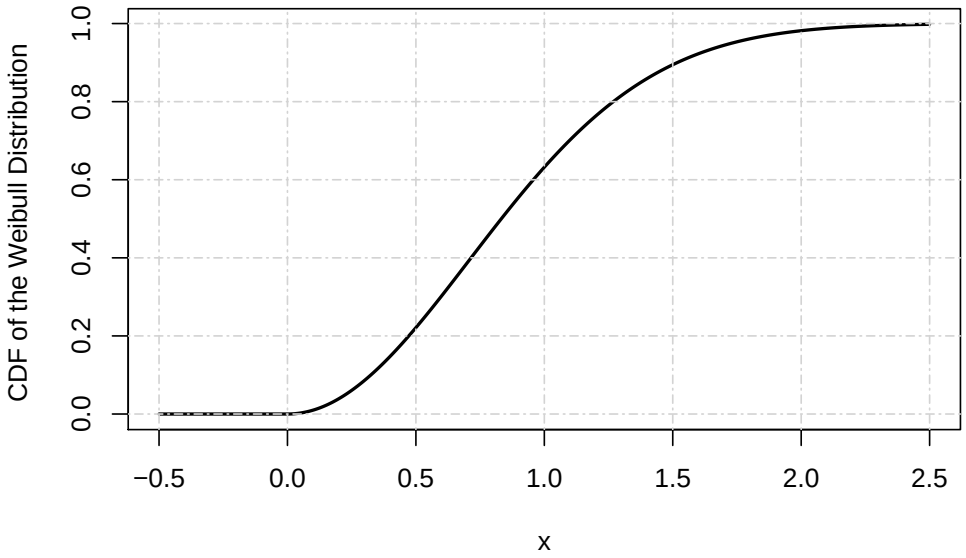
- Find, for any $\alpha \in (0, 1)$, the α-quantile of $Exp(1)$ Distribution.
- Find the 10% quantile of the Distribution with the PDF

$$f(x) = \begin{cases} 0.5 \cdot \sin(x), & x \in [0, \pi] \\ 0, & otherwise \end{cases}$$

b.

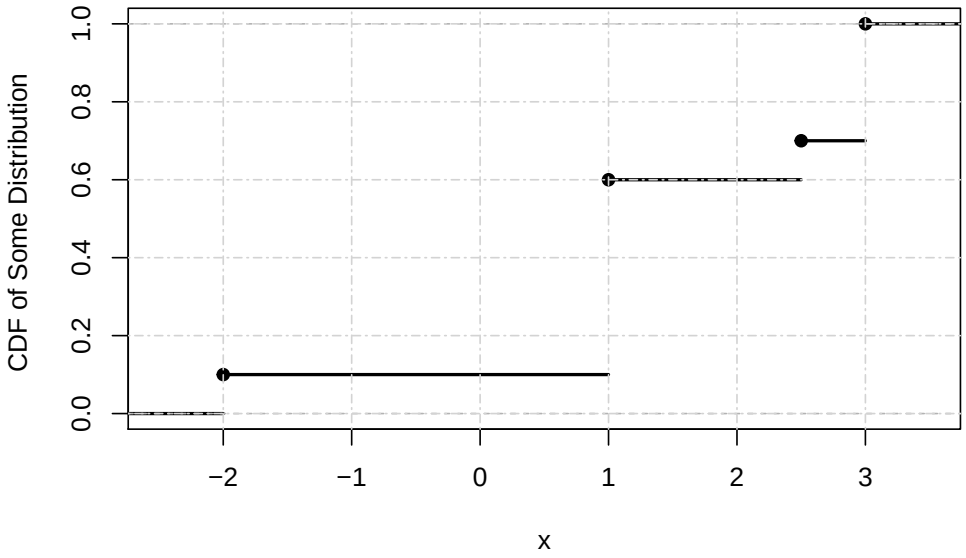
- Below you can find the graph of the CDF of the Weibull Distribution[1] with some parameters.

[1] See Wiki for Weibull Distrib



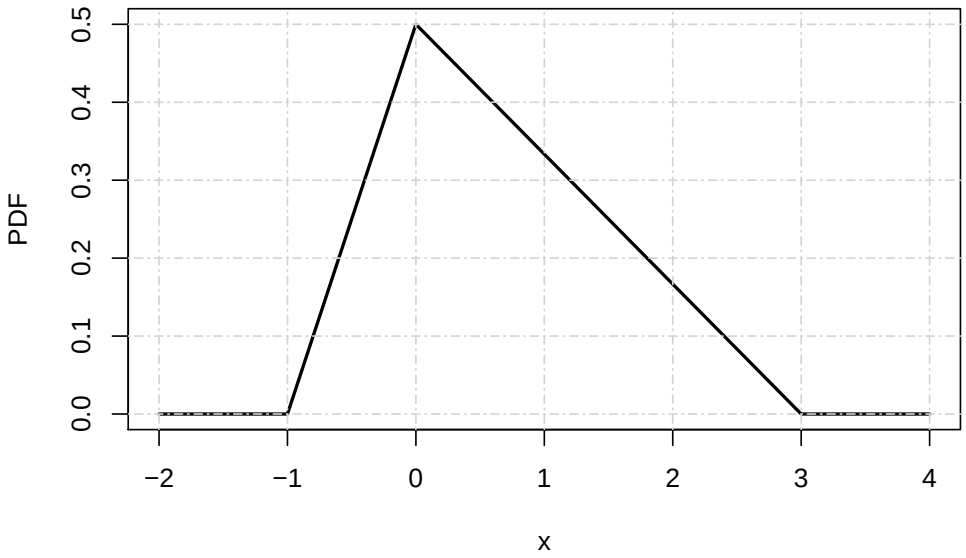
Find the approximate values of the Median and the 20%, 70% quantiles of that Distribution. Explain your reasoning.

- Below you can find the graph of the CDF of some Distribution.



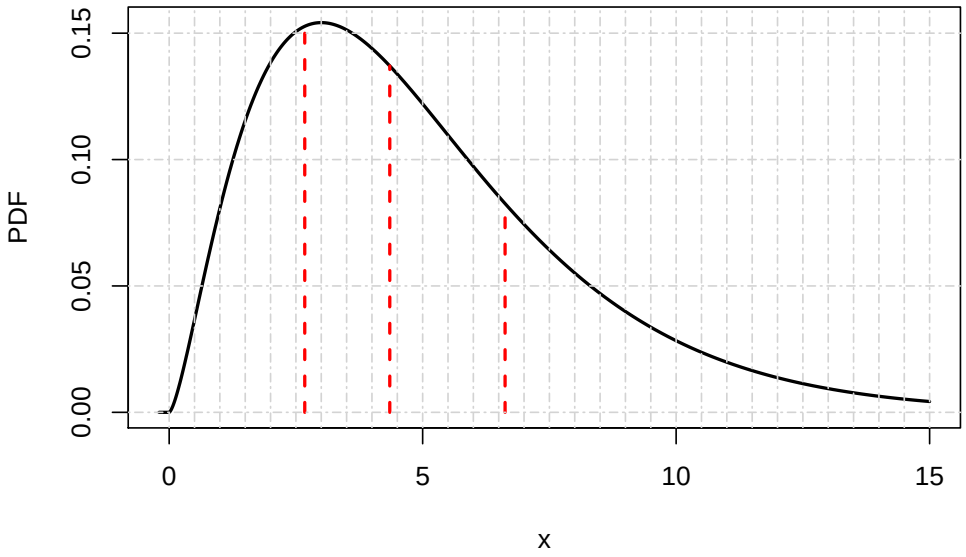
Find the approximate values of the quantiles of order 20%, 40%, 60%, 90%.

- Below you can find the graph of the PDF of some Distribution:



Find the exact values of the Median and the 20%, 70% quantiles of that Distribution. Show your calculations.

- Below you can find the graph of the PDF of some Distribution (x-gridline stepsize is 0.5):



The red dashed lines divide the area under the curve into 4 equal parts. Find the approximate values of the quantiles of order 10%, 25%, 50% and 75%. Explain your reasoning.

c. (R)

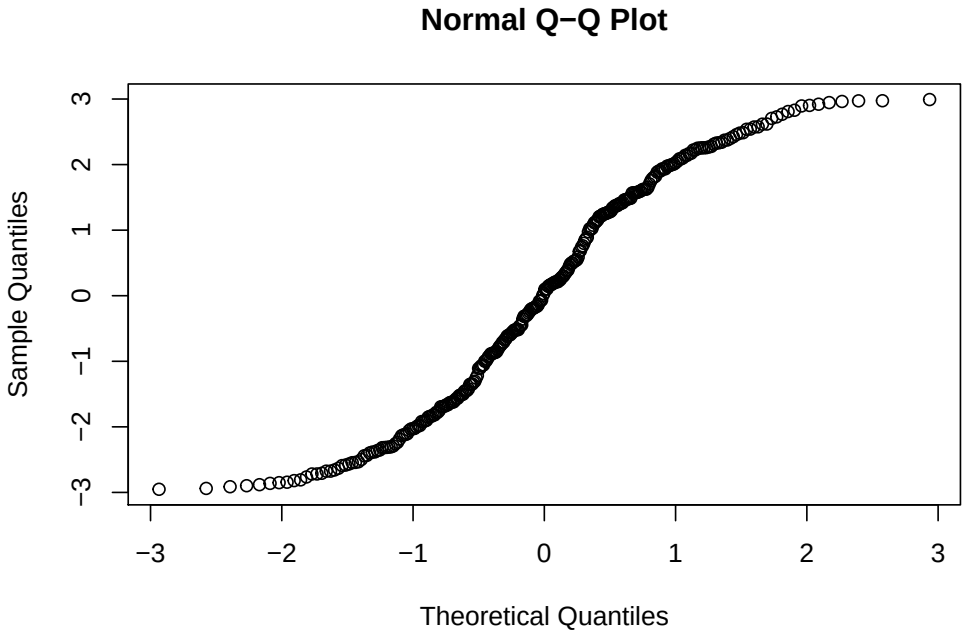
- Calculate all Deciles (quantiles of order $10\%, 20\%, 30\%, ..., 90\%$) of the Standard Normal Distribution;
- Construct a sequence of quantile levels $\alpha = 0.01, 0.02, ..., 0.99$. Plot the $\alpha-$Quantiles z_α of the Standard Normal Distribution vs α. Then, on the same graph, and using another color, plot the $(1-\alpha)$-level Quantiles of the same Distribution. Explain the symmetry (if you have plotted correclty, of course ☺).
- Find a symmetric interval $[a, b]$ such that for $X \sim \mathcal{N}(0, 1)$,

$$\mathbb{P}(X \notin [a, b]) = 0.99$$

Problem 4, Q - Q Plot

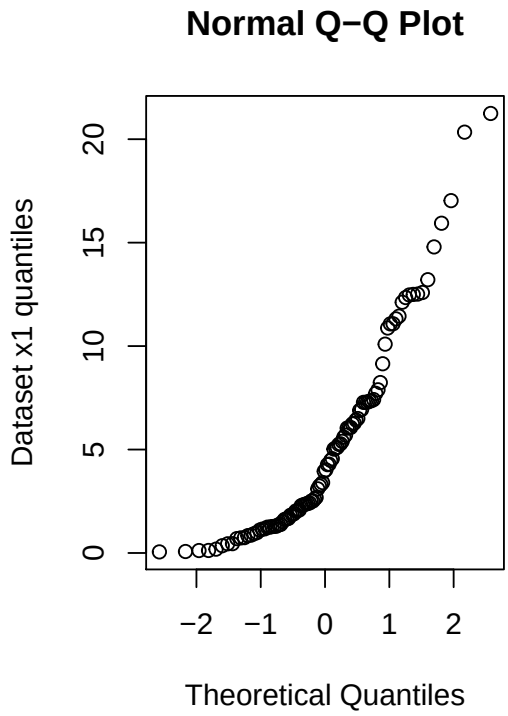
a.

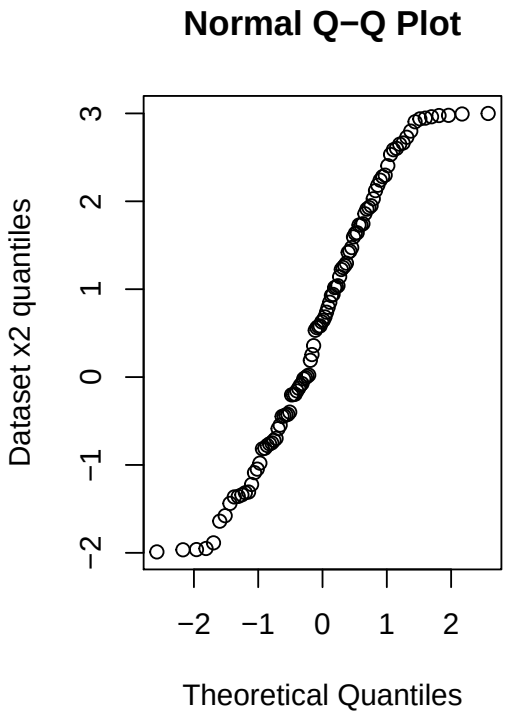
- I have generated a random sample from one of the Distributions $Unif[-3, 3]$ or $Exp(3)$, but forgot from which one. But I have the Q-Q Plot of my sample vs the Standard Normal Distribution:

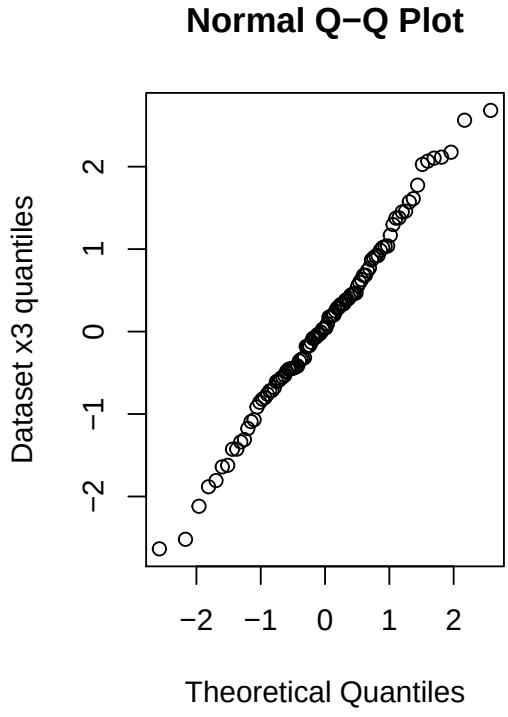


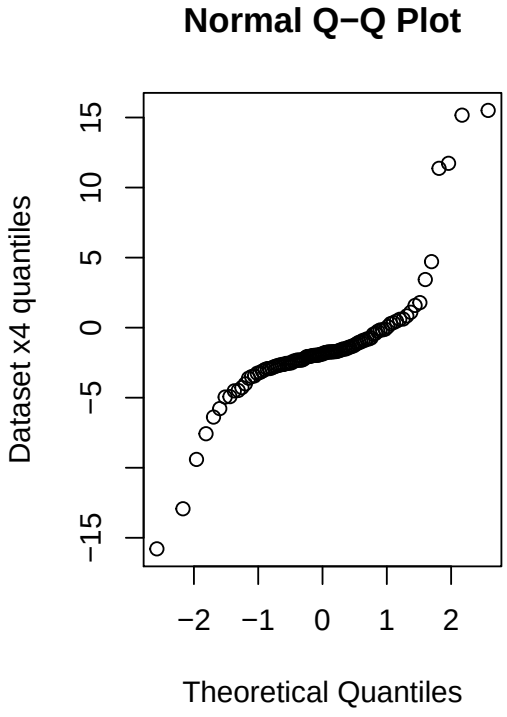
Help me to identify from which Distribution was my sample. Give your reasoning (so that next time I will be able to identify by myself ☺).

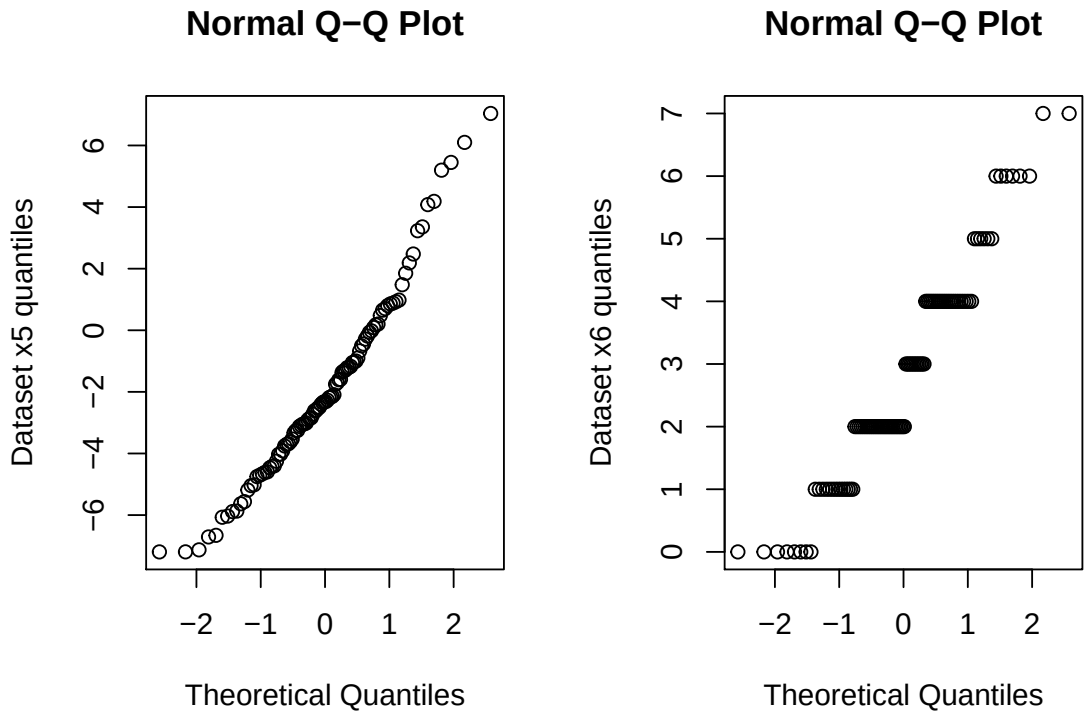
- Here are the Q-Q Plots of some random samples vs Standard Normal Distribution.


Normal Q−Q Plot
Dataset x1 quantiles
Theoretical Quantiles
0
5
10
15
20
−2
−1
0
1
2


Normal Q−Q Plot
Dataset x2 quantiles
Theoretical Quantiles
−2
−1
0
1
2
3
−2
−1
0
1
2


Normal Q−Q Plot
Dataset x3 quantiles
Theoretical Quantiles
−2
−1
0
1
2
−2
−1
0
1
2


Normal Q−Q Plot
Dataset x4 quantiles
Theoretical Quantiles
−15
−5
0
5
10
15
−2
−1
0
1
2



Which of these Dataset is likely to be from the Normal Distribution?

(Supplementary) Why is the last (bottom-right) Q-Q Plot different from the others? Explain!

b. (R)

- Generate a sample of size 200 from $Exp(3)$ and another one of size 400 from $Exp(0.2)$, and draw their Q-Q Plot.
- Write a function `qqexp(x)` and `qqunif(x)` so that they will draw the Q-Q Plot of the Dataset x vs the theoretical Quantiles from $Exp(1)$ and $Unif[0,1]$, respectively.

c. Q-Q Plot of AMZN daily returns (R, Supplementary)

Here we want to see if the weakly returns of the Amazon Stock can be modelled using a Normal Distribution. Daily returns are usually close to zero, sometimes positive (when the price increases), and sometimes negative (if the price decreases).

- Navigate to finance.yahoo.com and search for the Amazon ticker `AMZN`. Navigate to `Historical Data`, change the time period to `1Y` (1 year), choose daily frequency, hit `Apply`, and then `Download Data`. You will have the file of daily prices `AMZN.csv`.
- Read, using the `read.csv(file.choose())` command that `.csv` file. Separate in a new variable the `Adj.Close` (Adjusted Close Prices) variable.
- Calculate daily returns using the Adjusted Close Prices.
- Plot the Histogram of that daily returns.

- Draw the Q-Q Plot of that daily returns vs Standard Normal Distribution, using the `qqnorm` command. Add the `qqline` to the graph.
- Explain and make conclusions - will it be reasonable to model daily returns by using a Normal Distribution?
- I want to know the possible price for Amazon Stock for tomorrow. Suggest me a method to generate the possible value of the tomorrow's return, and I will calculate tomorrow's possible price.

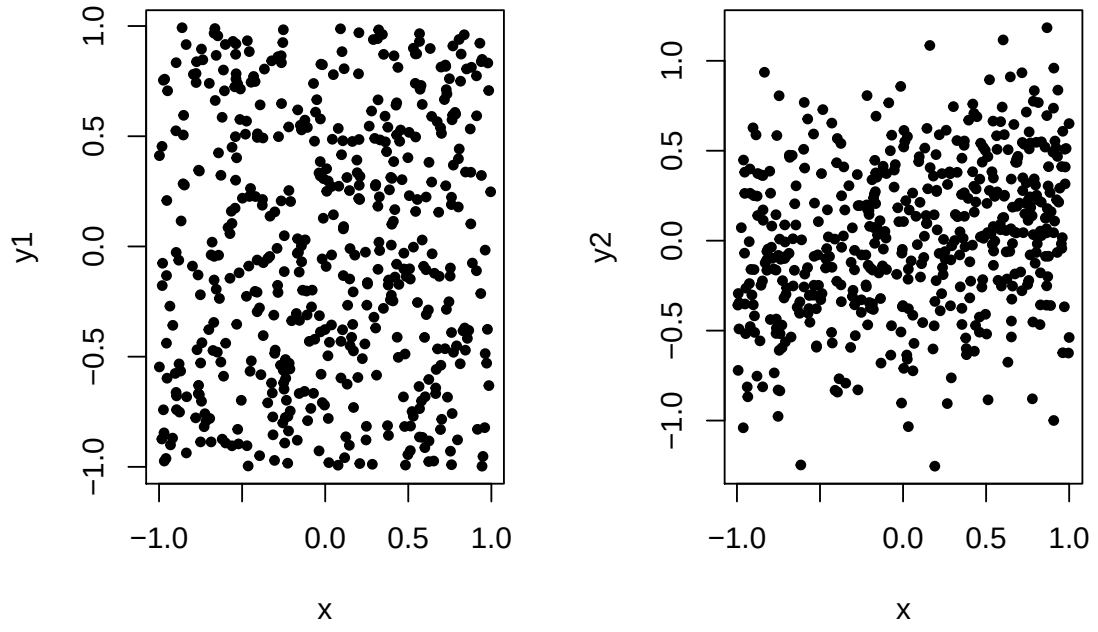
Problem 5, Covariance and Correlation

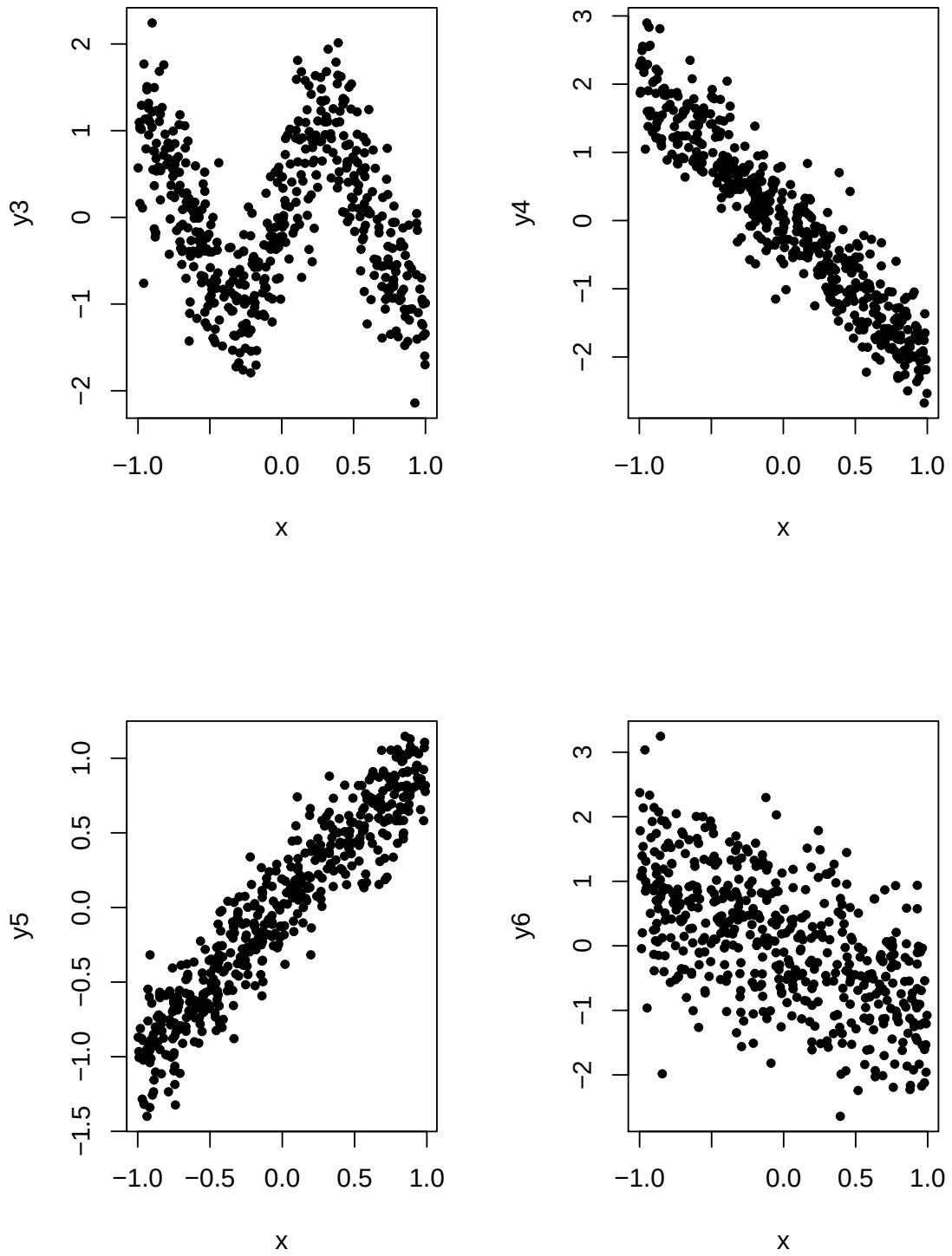
a.

- Prove that for any 2D dataset (x, y), $cov(\alpha \cdot x, \beta \cdot y) = \alpha\beta \cdot cov(x, y)$;

b.

- Below you can find Scatterplots for some Bivariate Datasets:





Here are the correlation coefficients for that Datasets, in some order:

```
## [1] -0.94097993  0.27655574 -0.12114875 -0.58353359  0.05916389  0.94223689
```

Which one corresponds to which Dataset?

c. (R)

Here we want to plot the Correlation Matrix and Heatmap for the Correlation between several variables.

We will work again with the `mtcars` Dataset.

- Print the first 3 observations of the `mtcars` Dataset
- Choose only numerical Variables (say, the Variable `cyl` is not numerical, it is categorical) of that Dataset and make a new Dataset (DataFrame) with the name `mtcars.new` consisting only of that numerical Variables.

Hint: Say, to choose the first and 4th Variables, you can use `mtcars[,c(1,4)]`

- Print the first 3 observations in your new Dataset `mtcars.new`
- Calculate the pairwise Correlations Matrix for the Dataset `mtcars.new`, and keep it in the **R** variable `cor.mat`

Hint: The function `cor` can calculate also the pairwise correlations, if the argument is a matrix or a DataFrame (see the help page for the `cor` function). So just use `cor(mtcars.new)`.

- Which variables are strongly (highly) positively/negatively correlated?
- Plot the Heatmap for your Correlation Matrix

Hint: You can use the `heatmap(cor.mat)` command. I am suggesting to use the `symm=TRUE` to have a symmetric map.

- (Supplementary) Change the Color Pallette in the Correlation HeatMap. Add also the color labels. Explore Heatmaps in `ggplot2` and `corrplot` packages (see An Introduction to corrplot Package). Read about Dendrograms and Clustering.
- (Supplementary) Here is an example of the usage of some Statistical Plots: an article. No need to go into the details.

d.

1. Calculate the Spearman's ρ for

$$x: \ -2, 0, 4 \qquad \textit{and} \qquad y: \ 2, 0, 100.$$

2. **(R)** Calculate the above ρ using **R**.

Hint: use `cor(x,y,method="spearman")`.

3. Prove that if x and y are in perfect increasing relationship (i.e., the scatterplot of x and y is an increasing graph), then for these Datasets $\rho = 1$.
4. **(R)** We want to see some comparisons between the Spearman's and Pearson's Correlation Coefficients. To that end, do the following experiments:

- Define x to be the vector $(1, 2, ..., 50)$;
- Define y to be the vector $(1^4, 2^4, ..., 50^4)$;
- Calculate the Pearson's Correlation Coefficient between x and y;
- Calculate the Spearman's Correlation Coefficient between x and y.

5. **(R)** We want to see the effect (sensitiveness) of outliers on Correlation Coefficients. To that end,

- Define x to be the vector $(1, 2, 3, 4, ..., 50)$;
- Take $ol = 10$ (ol is for *OutLier*);
- Define y to be the vector $(1, ol, 3, 4, ..., 50)$ (so the second element is our outlier);
- Do the y vs x Scatterplot;
- Print both Pearson's and Spearman's Correlation Coefficients side by side, in one row **Hint:** To print 2 elements in a row, you can make a vector out of that 2 elements, and then print that vector
- Now change ol to be $ol = 100$, and then run the code again
- Now change ol to be $ol = 1000$, and then run the code again
- Explain

6. **(R)** Here we use the `Animals` Dataset from the `MASS` package. If you do not have that package, use `install.packages("MASS")` to install.

- Read the help page for the `Animals` Dataset and describe its Variables
- Print the first 3 and last 3 observations of this Dataset
- Calculate the Pearson's and Spearman's Correlation Coefficients between this Dataset Variables;
- Explain the difference between the Correlation Coefficients.