CS 108 - Statistics, Sections B

Fall 2019, AUA

Homework No. 05

Due time/date: Section B: 10:32 AM, 04 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, Covariance and Correlation

a.

For a r.v. X, the Standardization of X is defined as

$$Z = Standardize(X) = \frac{X - \mathbb{E}(X)}{SD(X)}.$$

Likewise, for a Dataset x, the Standardization of x is defined as

$$z = Standardize(x) = \frac{x - \bar{x}}{sd(x)}.$$

Prove that

1. if the r.v. Z is the Standardization of a r.v. X, then

$$\mathbb{E}(Z) = 0 \qquad and \qquad Var(Z) = 1.$$

2. if the Dataset z is the Standardization of a DataSet x, then

$$\bar{z} = 0 \qquad and \qquad var(z) = 1.$$

3. If r.v.s Z_X and Z_Y are the Standardizations of r.v. X and Y, respectively, then

$$Cor(X, Y) = Cov(Z_X, Z_y);$$

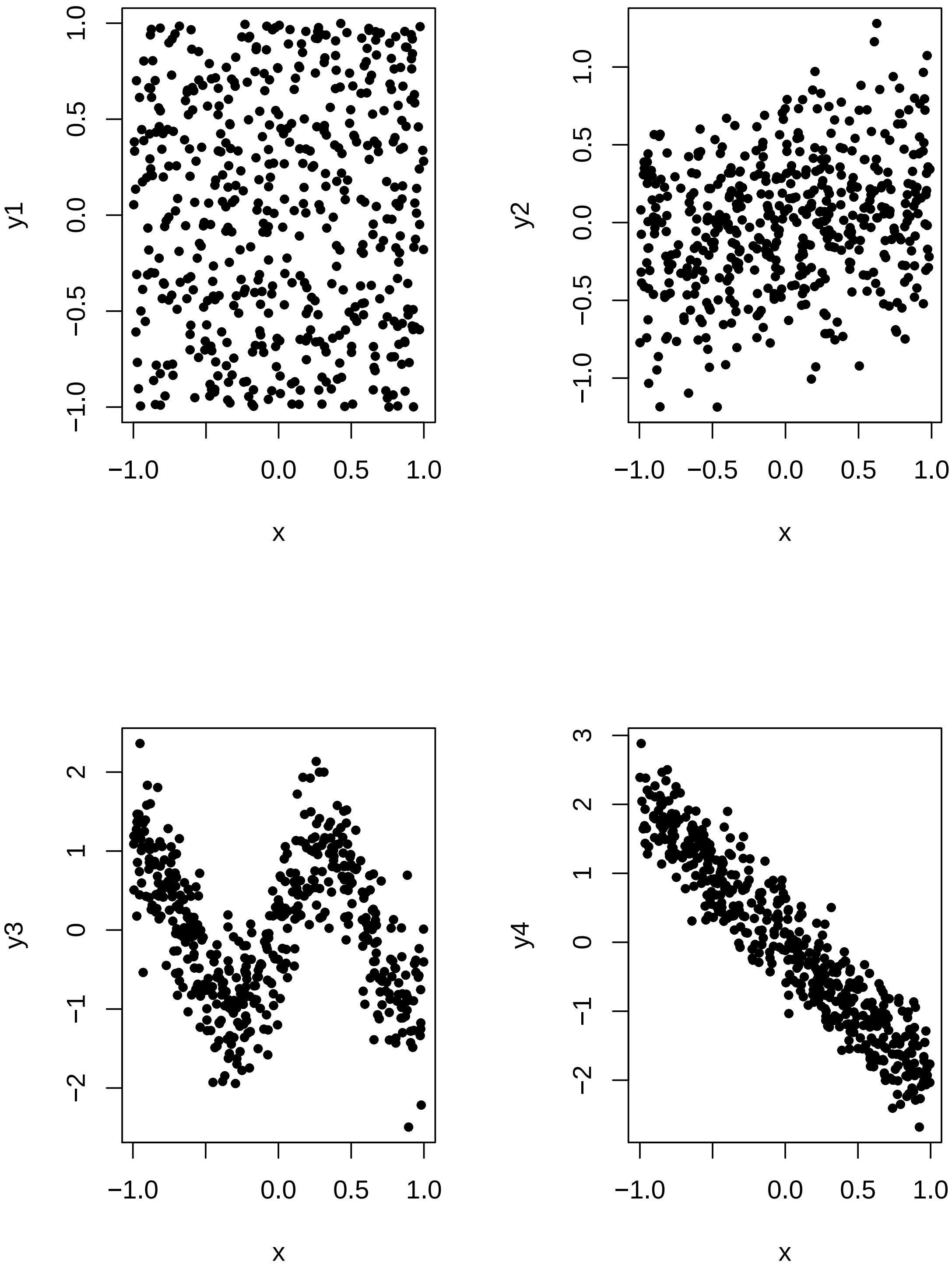
4. If Datasets z_x and z_y are the Standardizations of Datasets x and y, respectively, then

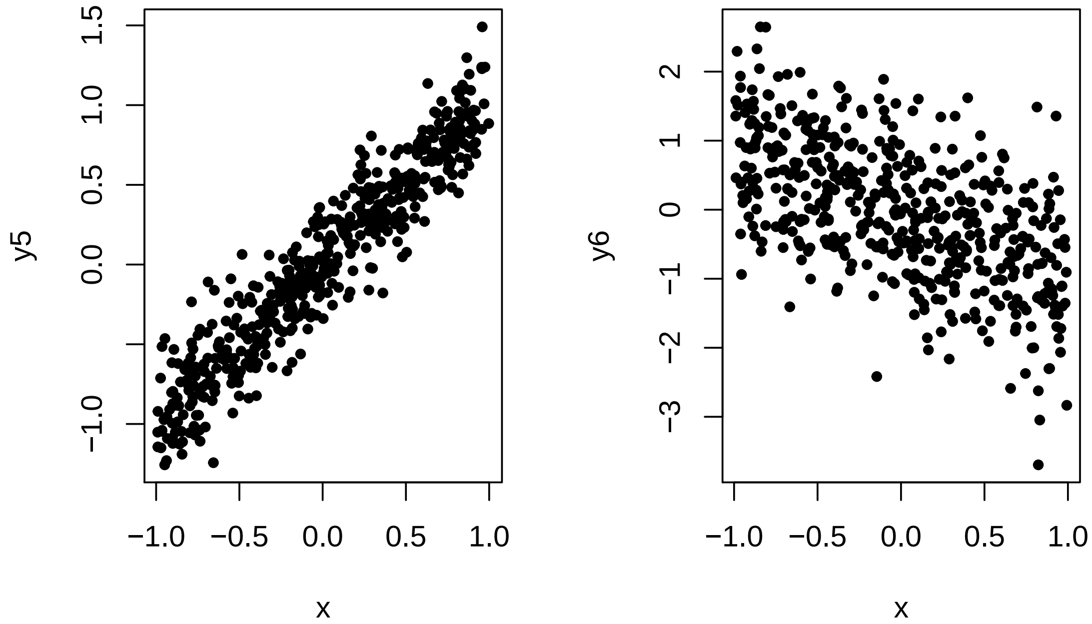
$$cor(x, y) = cov(z_x, z_y).$$

Note: Above, for Datasets, you need to take the same denominator when calculating sd, var and cov.

b.

- Below you can find Scatterplots for some Bivariate Datasets:





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

```
## [1] -0.94769390  0.24301610 -0.14739127 -0.59140222  0.02212758  0.94665452
```

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.

Here we want to define another measure of Correlation, the Spearman's ρ, and example of a Rank Correlation. As we have talked during our classes, our ordinary (Pearson's) Correlation Coefficient $cor(x,y)$ is measuring the *linear* relationship. Also, it is sensitive to outliers. Spearman's ρ is an alternative measure to capture the *monotonic* (not necessarily linear) relationship between the Datasets.

The definition is pretty simple: assume we have two Datasets of the same size,

$$x: \ x_1, x_2, ..., x_n \qquad and \qquad y: \ y_1, y_2, ..., y_n.$$

First we define the ranks of x and y: $rank(x)$ is the Dataset of positions (rank) of x_1, x_2, ..., x_n in the sorted array $sort(x)$. For example, if

$$x: \ 2, 1, 0, 4$$

then

$$sort(x): 0, 1, 2, 4,$$

hence, the rank of 2 in x is 3, since it will be the 3rd element in the sorted array. The rank of the next element of x, 1, is 2, since it is the second element in the $sort(x)$. Similarly, we will obtain

$$rank(x): \ 3, 2, 1, 4.$$

Another example: if

$$x: \ 1, 2, 3, 4, 0, 9,$$

then

$$rank(x): \ 2, 3, 4, 5, 1, 6.$$

Say, the element 3 in x will be the 4th element in the $sort(x)$, 0 in x will be the first element in $sort(x)$, etc.

Now, the Spearman's Correlation Coefficient ρ for Datasets x and y is defined by

$$\rho = \rho(x,y) = cor(rank(x), rank(y)).$$

So calculation of ρ is easy: first we calculate the ranks Datasets for x and y, $rank(x)$ and $rank(y)$, then calculate ordinary (Pearson's) COrrelation Coefficient between the ranks Datasets.

1. Calculate the Spearman's ρ for

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

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.

7. (Supplementary) Read about the Kendall's τ measure for the Correlation between 2 variables. Use the **R** `cor` function parameter `method` to calculate the τ for some Datasets.

Problem 2, Probability Refresher, RVs

a.

Let $X \sim Pois(2)$ and $Y \sim Exp(3)$. Calculate

1. $\mathbb{P}(X >= 2)$;
2. $\mathbb{E}(X^2)$;

Hint: You can use the Variance!

3. $\mathbb{P}(Y < 3)$
4. Assuming X and Y are independent, calculate $\mathbb{E}(XY)$.

b. (R)

Assume I made an Ad on FB and I want to model the number of clicks during a day on my Ad. I have calculated that the average number of clicks in a day is 34.3.

1. Suggest a model for the number of clicks
2. Calculate the probability that I will have more than 40 clicks tomorrow.
3. Generate a possible scenario for the number of clicks for each day of the next week.

Problem 3. Convergence of r.v.s

a.

Assume X_n, $n \geq 3$, is a r.v. with the following PMF:

Values of X_n	$-\frac{1}{n}$	$3 + \frac{n+1}{n^2+1}$
$\mathbb{P}(X_n = x)$	$\frac{1}{3} - \frac{1}{n}$	$\frac{2}{3} + \frac{1}{n}$

Check, using only the definitions, if X_n converges to some limit in three senses: in Probability, in Quadratic Mean and in Distributions.

b.

Assume $X_n \sim Exp(\frac{1}{n})$ and $Y_n \sim Exp(n)$ (assume also that all r.v.s are defined on the same Probability Space). Check, using only the definitions, if

1. $X_n \xrightarrow{\mathbb{P}} 0$ and $Y_n \xrightarrow{\mathbb{P}} 0$
2. $X_n \xrightarrow{qm} 0$ and $Y_n \xrightarrow{qm} 0$
3. $X_n \xrightarrow{D} 0$ and $Y_n \xrightarrow{D} 0$

Note: You can "prove" or "disprove" the convergence in Distributions using graphs in **R**: say, you can plot the CDFs for different values of n to see the dynamics.

c. (R)

Assume $X_n \sim \mathcal{N}(0, \frac{1}{n})$. Guess the limit in Distributions of X_n, and "prove" that X_n indeed tends to your guess, in the Distributions sense, geometrically. To that end, you need to plot the CDFs for different increasing values of n, on the same graph, and also the CDF of the limit. Use different colors/line types for different n-s and the limit. Add also the legend (explanation which line is for which CDF).

Note: In the `plot` function, you can change the line type by using the `lty` parameter. Try, for example `lty=1`, `lty=2`, `lty=3`,... . To add a legend to a graph, use the `legend` function, see, e.g., this link.