# ASDS - Applied Statistics with R

## Fall 2018, YSU

# Homework No. 03

Due time/date: 21:20, 19 October, 2018

**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.** For any distribution, we call a number $m \in \mathbb{R}$ to be a median of that distribution, if

$$\mathbb{P}(X \leq m) \geq \frac{1}{2} \qquad \mathbb{P}(X \geq m) \geq \frac{1}{2},$$

for a r.v. $X$ with that distribution.

a. Prove that if our distribution is continuous with CDF $F(x)$ and PDF $f(x)$, then $m$ is a median if and only if

$$F(m) = \frac{1}{2} \qquad \text{or, equivalently,} \qquad \int_{-\infty}^{m} f(x)dx = \int_{m}^{+\infty} f(x)dx = \frac{1}{2}.$$

b. Give examples of continuous distributions (analytically or geometrically, through $F$ or $f$) with unique and non-unique medians.

c. Find a median of $Bernoulli(\frac{1}{3})$;

d. Find a median of $Binom(2, \frac{1}{2})$;

e. Find a median of $Unif([0,1])$.

**Problem 2.** Assume we have a dataset $x : x_1, x_2, ..., x_n$, and for some numbers $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$, the dataset $y = \alpha \cdot x + \beta$, i.e., $y_i = \alpha \cdot x_i + \beta$, $i = 1, ..., n$.

a. Assume $X$ is a r.v. and $Y = \alpha \cdot X + \beta$. Express $Var(Y)$ in terms of $Var(X), \alpha, \beta$;

b. For our datasets, express $var(y)$ in terms of $var(x), \alpha, \beta$;

c. Choose $\alpha, \beta$ in the part **a.** such that $Y$ is a *standardized* r.v., i.e., $\mathbb{E}(Y) = 0$ and $Var(Y) = 1$;

d. For the part **b.**, choose $\alpha, \beta$ such that $y$ is a *standardized* dataset, i.e., $mean(y) = 0$ and $var(y) = 1$.

**Problem 3.** Prove that if we will take for a dataset $x$: $x_1, ..., x_n$,

$$var(x) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \overline{x})^2, \qquad \overline{x} = mean(x),$$

then

$$var(x) = mean(x^2) - \big(mean(x)\big)^2,$$

where by $x^2$ we mean the dataset $x_1^2, x_2^2, ..., x_n^2$.

**Problem 4.** Construct two different datasets with the same mean and variance.

**Problem 5.** (**R**) At the end of the second part of my lecture notes on Moodle, page 70, I presented a problem from the well-known mathematical journal American Mathematical Monthly, Volume 124, 2017 - Issue 2. In fact, I do not know the answer, so I am asking you to make experiments to get the idea about the shape of the distribution. To that end, first you need to assume that the nested formula is finite (say, it goes up to $X_{1000}$), then generate some sample $X_1, X_2, ...., X_{1000}$, calculate the corresponding value of the expression. Then generate another sample $X_1, X_2, ...., X_{1000}$, again calculate the expression value. And do this many-many times. At the end you need to draw the distribution of the values of the expression for all experiments.

Give both the picture of the distribution and the code.

**Problem 6.** (**R**) Generate 200 normally distributed random numbers with the mean 0.7 and standard deviation 1.1. Calculate the number of outliers in that dataset. Repeat the experiment 100 times, and calculate the mean number of outliers.

**Problem 7.** Give an example of a dataset such that its 50% quantile is not equal to the Median of that dataset.

**Problem 8.** Calculate the 10%, 30% and 70% quantiles of the dataset

$$1, 1, 2, -3, -1, 0, 0, 3, 1$$

**Problem 9.** Construct a dataset of 12 elements, such that the 30% quantile of that dataset is 4, and 70% quantile is 5.

**Problem 10.**

a. Calculate the 0.5-th quantile of the distribution $Exp(2)$;

b. Calculate the $\alpha$-th quantile of the distribution $Unif[a, b]$;

c. (Supplementary) Calculate the 0.4-th quantile of the distribution $Poisson(1)$.

**Problem 11.** We have the following observations for two variables $x$ and $y$:

$$x : 0, 1, 2, 1, 0, -2, 1, \qquad y : -3, 2, 1, 2, 1, -2, 0$$

a. Give the Scatterplot for this data;

b. Calculate the sample covariance and correlation coefficient for $x, y$.

**Problem 12.** I have calculated correlation coefficients for 6 different pairs of datasets. Fig. 1 gives the scatterplots.

Now, the calculated correlation coefficients are:

$$-0.8849832, \quad 0.9454197, \quad -0.05221834, \quad 0.7774705, \quad 0.002113799, \quad 0.9778802.$$

Can you identify which coefficient is for which case? Explain your reasoning.

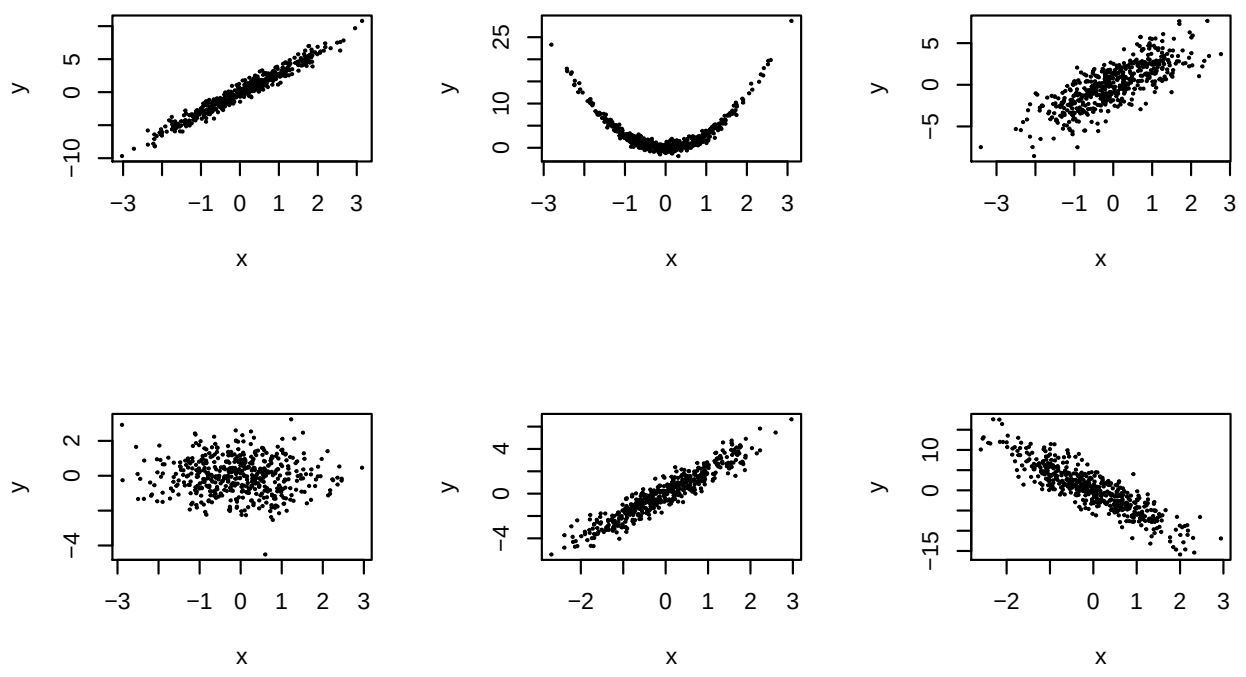


**Figure 1:** Scatterplots for 6 Datasets.

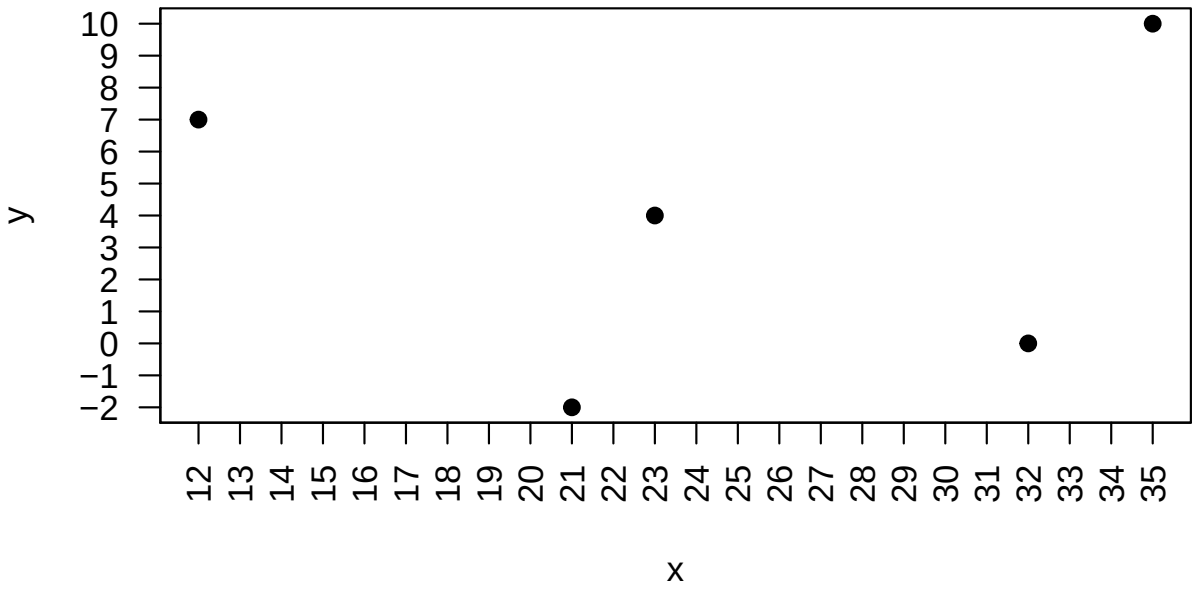


**Figure 2:** Scatterplot

**Problem 13.** We are given a scatterplot for some dataset, see Fig. 2 (all datapoints have integer coordinates).

Recover the dataset (i.e., find $x$ and $y$) and calculate the covariance between $x$ and $y$.

**Problem 14.** Prove that for any 2D dataset $(x, y)$,

a. $var(x+y) = var(x) + 2 \cdot cov(x,y) + var(y)$;
b. if $x$ and $y$ are uncorrelated, then $var(x+y) = var(x) + var(y)$

**Problem 15.** Construct a bivariate dataset with uncorrelated variables. The number of datapoints need to be larger than 2. Give the scatterplot of your dataset.

**Problem 16.** (**R**) The *mtcars* dataset in **R** is from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

a. Give the Scatterplot of the variable *disp* vs. *wt*;
b. Calculate the correlation coefficient between the *wt* and *disp*; explain the result;
c. Calculate the covariance between the *wt* and *disp*
d. Calculate the variance of the variable *disp*

**Problem 17.** (**R**) Generate 200 random numbers from the distribution $\mathcal{N}(3,5)$, store it in the vector $x$, and also generate 300 random numbers from the distribution $Unif[-2,8]$, and store it in $y$.

a. Calculate, using **R**, the $\alpha$-th quantiles $q^x_\alpha, q^y_\alpha$, for $x$ and $y$ respectively, for $\alpha = 0.05, 0.1, 0.15, 0.2, 0.25, ..., 0.95$
b. Plot the points $(q^x_\alpha, q^y_\alpha)$. Are these points on the same line?
c. Calculate also, using **R**, the $\alpha$-th quantiles $q^{\mathcal{N}}_\alpha$, for the Standard Normal Distribution, i.e., for $\mathcal{N}(0,1)$, for the same values of $\alpha$ as above;
d. Plot on the same window 2 graphs: one needs to be the graph of points $(q^x_\alpha, q^{\mathcal{N}}_\alpha)$, and the other one - for $(q^y_\alpha, q^{\mathcal{N}}_\alpha)$. Explain the results.

Hint: To put 2 graphs in one window, you can use the command $par(mfrow = c(1,2))$ before doing the plots.

**Problem 18.** (Supplementary)

a. Assume $X$ is a r.v. and $Y = \alpha \cdot X + \beta$ for some numbers $\alpha, \beta \in \mathbb{R}$. Express the quantiles of $Y$ through the quantiles of $X$, $\alpha, \beta$;
b. Assume $x$ is a dataset and $y = \alpha \cdot x + \beta$ for some numbers $\alpha, \beta \in \mathbb{R}$. Express the quantiles of $y$ through the quantiles of $x$, $\alpha, \beta$.