Prof. Dr. E. Cramer, Prof. Dr. M. Kateri,
L. Kaufmann, M.Sc., T. van Bentum, M.Sc. ,
A. Müllenmeister

RWTH Aachen, SS 2021
Release date: April 20th, 2021
Solution: April 27th, 2021

Applied Data Analysis

R-Laboratory 2

Normal Distribution – Tidy Data – Data Preparation

Useful packages and functions:

- `density()`
- `hist()`
- `rnorm()`
- `sapply()`
- `dplyr`
- `dplyr::filter()`
- `dplyr::arrange()`
- `dplyr::fill()`
- `ggplot2`
- `ggplot2::ggplot()`
- `cut()`
- `regexpr()`
- `duplicated()`
- `which()`
- `boxplot()`
- `save()`
- `read.csv()`
- `write.csv()`
- `mvtnorm`
- `mvtnorm::rmvnorm()`
- `pairs()`
- `t()`
- `solve()`
- `MASS`
- `MASS::ginv()`
- `svd()`
- `diag()`

Task 5

(a) Draw random samples of size $n = 30, 100, 300$ from a $\mathcal{N}(\mu, \sigma^2)$ distribution with $\mu = 5$ and $\sigma^2 = 4$ and create a histogram for each sample size n.

(b) Add a density estimation using the function `density` and the probability density function of a $\mathcal{N}(5, 4)$ distribution to the histograms using different colors. What do you observe?

Task 6

(a) Draw random samples of size $n = 100$ from a $\mathcal{N}_4(\mathbf{0}, I_4)$ distribution.

Hint: You may use the functions `rnorm` and `matrix` or the function `rmvnorm` from the package `mvtnorm`.

(b) Initialize a vector $\boldsymbol{\mu} = (1, 0, 2, -1)'$ and matrices

$$\Sigma_1 = \begin{pmatrix} 4 & 2 & 2 & 3 \\ 2 & 3 & 2 & 1 \\ 2 & 2 & 5 & 2 \\ 3 & 1 & 2 & 3 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 4.5 & 4.75 & 2 & 2.25 \\ 4.75 & 5.25 & 2.75 & 3.25 \\ 2 & 2.75 & 2.75 & 3.5 \\ 2.25 & 3.25 & 3.5 & 4.5 \end{pmatrix}.$$

(c) Transform the random vectors from (a) to a sample from a $\mathcal{N}_4(\boldsymbol{\mu}, \Sigma_1)$ distribution and a sample from a $\mathcal{N}_4(\boldsymbol{\mu}, \Sigma_2)$ distribution. Do not generate new random numbers! Use a singular-value decomposition instead.

Remark: R computes the singular-value decomposition numerically. Replace eigenvalues smaller than `sqrt(.Machine$double.eps)`$= 1.490116 \cdot 10^{-08}$ with 0.

(d) Create three scatterplot matrices - one for each sample. What do you observe?

(e) Compute the Moore-Penrose general inverse of Σ_1 and Σ_2. If the inverse of Σ_1 and Σ_2 exists, does it coincide with the Moore-Penrose general inverse?

Task 7

(a) Download the CSV-files *Survey1a.csv* and *Survey1b.csv* from the RWTHmoodle space of the course Applied Data Analysis and import the data as a `data.frame` object into the R workspace.

(b) Transform the measured dimensions and the mean score of *Survey1a.csv* and *Survey1b.csv* to type `numeric` appropriately.

(c) Create a new `data.frame` called `data.survey` that contains the observations of *Survey1a.csv* and *Survey1b.csv*. Remember to fill missing values and to remove duplicated observations.

Hint: You may use the functions arrange, filter and fill from the package dplyr.

(d) For the data of `data.survey`, create an (`Age`, `DimSchool`) scatterplot (with the values of `Age` on the horizontal axis). Differentiate the points by sex with colors.

Hint: You may use the package ggplot2

(e) Create two Box-plots for `DimFriends` in one figure, one for male and one for female participants.

(f) Save the `data.frame` into an `.RData` file.

Task 8

(a) Download the file *credits.wsv* from RWTHmoodle and import the data as a `data.frame` object into the R workspace.

(b) Switch the coding for the binary variable `gastarb` in the `data.frame` object form 2 to 1 for Gastarbeiter and from 1 to 2 for a native worker.

(c) To score future credit applicants, a bank employee suggests the following discretization of the metric variables `time`, `amount` and `age` in the data set:

time	**score**	**amount**	**score**	**age**	**score**
$(0,6]$	10	$(0,500]$	10	$(0,25]$	1
$(6,12]$	9	$(500,1000]$	9	$(25,39]$	2
$(12,18]$	8	$(1000,1500]$	8	$(39,59]$	3
$(18,24]$	7	$(1500,2500]$	7	$(59,64]$	5
$(24,30]$	6	$(2500,5000]$	6	$(64,\infty)$	4
$(30,36]$	5	$(5000,7500]$	5		
$(36,42]$	4	$(7500,10000]$	4		
$(42,48]$	3	$(10000,15000]$	3		
$(48,54]$	2	$(15000,20000]$	2		
$(58,\infty)$	1	$(20000,\infty]$	1		

Create the three variables dtime, damount and dage by this discretization and include them to the data.frame object. The bank employee suggests as simple score to predict the repayment behavior (i.e. the value of repayment) of credit applicants the sum of the values of the following variables:
account, dtime, behavior, usage, damount, savings, employment, rate, famgen, guaran, residence, finance, dage, furthcred, home, prevcred, job, pers, phone, gastarb.
Considering the scores as quantitative variables, create a further variable simple.score by this approach and include it to the data.frame object.

(d) Compare the values of simple.score for the data points of both values of repayment. What is your first impression of this score?

(e) Save the data.frame into a CSV-file.

Note: We will revisit this data set later on and discuss the appropriateness of this predictor along with possible alternatives.