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 27th, 2021
Solution: May 4th, 2021

Applied Data Analysis

R-Laboratory 3

Central Limit Theorem – Simple Linear Models

Useful packages and functions:

- `table()`
- `barplot()`
- `sprintf()`
- `rbinom()`
- `title()`
- `axis()`
- `dplyr`
- `dplyr::mutate()`
- `pairs()`
- `abline()`
- `lm()`
- `MASS`
- `MASS::ginv()`
- `predict()`
- `poly()`
- `I()`
- `qf()`

Task 9

(a) Draw a random sample of size $m = 30$ from a $\mathcal{B}(n, p)$-distribution, the Binomial distribution with parameter $n = 12$ and $p = 0.7$, applying the R-function `rbinom`.

(b) Construct the bar plot and add the probability mass function (pmf) of the generating distribution.

(c) Calculate the mean ($\bar{x}$) and the variance (s^2) of your sample and write them in the title of the figure. Furthermore, calculate

$$T_{m,n,p} := \sqrt{m}\left(\frac{\bar{x} - np}{\sqrt{np(1-p)}}\right)$$

directing the output to the console.

(d) Write a function with arguments m, n and p which draws a new random sample of size m from a $\mathcal{B}(n, p)$-distribution and returns the value of $T_{m,n,p}$.

(e) Apply the function from (d) 10,000 times for $m \in \{5, 30, 500\}$, with $n = 12$ and $p = 0.7$. For each m, create a histogram with 16 breaks for the returned values. What do you observe?

Task 10

(a) Download the CSV-file *Solar.csv* from the RWTHmoodle space of the course Applied Data Analysis. Import the data as a `data.frame` object into the R workspace and transform the attribute `batch` to type `factor`.

(b) Create a scatterplot matrix of the attributes `Pmax`, `Imax`, `Umax`, `Isc` and `Uoc`. Differentiate the points by `batch` using colors.

(c) Create Box-plots for `Uoc` for each batch in one figure.

(d) For the data of *Solar.csv*, create an (`Pmax`, `Isc`) scatterplot. Differentiate the points by `batch` using colors and add a linear regression line. Compute the parameter vector

  (i) via Example I.4.6 and Theorem I.4.9 of the lecture,

  (ii) via the function `lm`.

  Hint: `lm` needs an argument `formula`. An object of class `formula` takes the form "*response*∼*terms*", where *terms* describes the predictors for *response*. The intercept of a linear model is given as default. If there is no intercept in the model you need to add "-1" to *terms*. The `formula` `Isc`∼`Pmax` describes the simple linear regression model above.

(e) Add corresponding colored regression lines based on the observations from batch 1 and batch 4.

(f) Predict the missing values of `Isc` based on the regression in (d).

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

Task 11

(a) Download the CSV-file *rent.csv* from RWTHmoodle. Import the data as a `data.frame` object into the R workspace.

(b) Create a scatterplot of the attributes `rent.sqm` (y-axis) and `space` (x-axis). Add a linear regression line (you may use the function `lm`) to the scatterplot. Does the linear regression describes the data well? Is there a transformation of one of the two variables which possibly allows the creation of a better fitting linear model?
Hint: Create a scatterplot of `rent.sqm` (y-axis) and `1/space` (x-axis).

(c) Create a regression model with the approach

$$\texttt{rent.sqm} = a + \frac{b}{\texttt{space}}$$

for real valued parameters $a, b \in \mathbb{R}$ (it is a linear model in the parameters). Add the regression curve to the first scatterplot in (b). Does this model provide a better description of the relation between `rent.sqm` and `space` than the simple linear regression of (b) based on your visual impression?
Hint: You can add the term $\frac{b}{\texttt{space}}$ to the formula of linear model by adding `I(1/space)` to the argument `formula` of `lm`.

Task 12

(a) Download the white-space-separated file *cars2.dat* from RWTHmoodle. Import the data as a `data.frame` object into the R workspace.

(b) Create a scatterplot of the attributes `dist` (y-axis) and `speed` (x-axis) of the `cars2` data set.

(c) Add a quadratic regression curve to the scatterplot by using a linear model with the approach

$$\texttt{dist} = a + b \cdot \texttt{speed} + c \cdot \texttt{speed}^2 \qquad (+)$$

for real valued parameters $a, b, c \in \mathbb{R}$ (it is linear in the parameters).
Hint: You can add the term $c \cdot \texttt{speed}^2$ to the formula of linear model by adding `I(speed^2)` to the argument `formula` of `lm`. Alternatively, you can use the function `poly` to create a polynomial predictor for a linear model. In the latter case, it is recommended to compute the points for the regression curve using the function `predict`.

(d) Test the hypotheses

$$\mathrm{H}_0 \colon c = 0 \qquad \text{versus} \qquad \mathrm{H}_1 \colon c \neq 0$$

on the significance level $\alpha = 0.05$ for the parameter c of the linear model with the approach (+) via the F-test of Testing procedure I.4.40 of the lecture. Consider the conditions of the F-test to be satisfied. Does the test reject the null hypothesis?