Introduction to Statistics and Data Analysis with R - Homework #1

Adi Sarid and Afek Adler

2019-11-03

This homework sheet is due on the 18th of November. You may submit your answers in pairs. Submission will be performed electronically via the moodle.

We urge you to start solving this sheet as soon as possible and, if you have any questions, come to visit us in reception hours next week.

The exercise is divided into two parts: Technical (programming in `R`) and theoretical.

Technical (programming in `R`)

Question 1:

Please read the following chapters in R4DS - https://r4ds.had.co.nz:

1. Introduction
2. Explore - Introduction
3. Explore - Data visualizations **up to 3.6**

Solve exercise 3.6.1, and **submit the code for question 6** ("*Recreate the R code necessary to generate the following graphs*").

Question 2:

In this question, you will get acquainted (or reminded of) the following distributions:

- Normal distribution $N(\mu, \sigma)$
- Student's t t_{df}
- Chi-square χ^2

Complete the blanks (`___`) in the following code, to generate $n = 100$ random values from each of these distributions with:

- Normal with $\mu = 3, \sigma = 1.5$
- Student's-t with df $= 10$
- Chi-square with df $= 12$

Tip: if you type a `?` followed by the command name in the console, you will see its documentation. I.e., type `?rnorm` to see the help on the random number generator for the normal distribution.

Complete the blanks:

```
set.seed(0)  # we set the seed of the random generator so that your results will be consistent

random_normal <- rnorm(n = ___, mean = ___, ___ = 1.5)
random_t <- rt(n = ___, df = ___)
random_chi <- rchisq(n = ___, df = ___)
```

Plot by completing the blanks:

Plot each of these samples using `ggplot2`. Think, what `geom` would you use to plot the distribution of the sample?

```
# if you don't have the tidyverse package first install by running
# install.packages("tidyverse")

library(tidyverse)

all_random_data <- tibble(random_normal, random_t, random_chi)

ggplot(all_random_data, aes(random_normal)) +
  geom____()

ggplot(all_random_data, aes(random_t)) +
  ---

ggplot(all_random_data, aes(random_chi)) +
  ---
```

Answer these:

1. Is the original distribution symmetric? does the plots look symmetric, why?
2. Generally speaking (not relating to the specific sample you obtained), what is the relationship between the mean and median of each of these distributions?
3. What would happen if we increase n from 100 to 1000?
   a. How would the distribution look like?
   b. Why?
   c. Modify your code and visualize the updates.

Theoretical

Question 3:

In the smallest branch of the smallest bank, the number of customers in the queue (waiting customers), is a random variable $Q \in \{0, 1, 2\}$. You cannot have more than 2 customers waiting in the queue, because the've been downsizing and the branch is really small.

The distribution of Q is dependent on a parameter θ.

$$Q = \begin{cases} 0 & \text{w.p. } 4\theta^2 \\ 1 & \text{w.p. } 4\theta - 8\theta^2 \\ 2 & \text{w.p. } 1 - 4\theta + 4\theta^2 \end{cases}$$

The bank's headquarters randomly sampled the queue during five independent times. The results were $\{0, 1, 0, 0, 0\}$ customers in the queue.

Answer the following questions:

1. Find an unbiased estimator $\hat{\Theta}$ for the parameter θ for a sample of size $n = 5$. What is $\hat{\Theta}$ based on the current sample? (you should get 0.45)

2. Find an unbiased estimator for the expected number of customers waiting in the queue based on a sample of size $n = 5$. What is the estimate of the expected number of customers, based on the current sample? (0.2)
3. Find an estimator for θ in the maximum likelihood estimation method. (0.45)

Question 4:

let X be a random Bernoulli variable. It's probability density function can be formulated as follows:

$$f(x;p) = \begin{cases} p^x(1-p)^{1-x} & x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Show that $X = 1$ with probability p and that $X = 0$ with probability $1 - p$
2. Suppose we get a random sample of size n from a Bernulli distribution. What is the likelihood function $L(p)$ of the sample? (what is the probability that $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$)
3. Apply the log transformation on this likelihood function, what do you get?
4. Find the p that maximizes $\log L(p)$

Question 5:

For the following probability density function:

$$f(x) = \begin{cases} \frac{2}{\theta^2}(\theta - x) & 0 < x < \theta \\ 0 & \text{else} \end{cases}$$

Find θ by the method of moments.

Question 6:

For the exponential distribution:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Find λ by MLE and by the Method of Moments.