Home Exercises 4

Your Name

18.10.2021

Write your name at the beginning of the file as "author:".

1. Return to Moodle by **9.00am, Mon 18.10.** (to section "BEFORE").
2. Watch the exercise session video available in Moodle by **10.00am, Mon 18.10.**
3. If you observe during the exercise session that your answers need some correction, return a corrected version to Moodle (to section "AFTER") by **9.00 am, Mon 25.10.**

Problem 1.

Read in the data from "prostate.txt" using command

```
pr = read.table("prostate.txt", as.is = TRUE, header = TRUE)
```

when the file is in your current working directory. (You can use `getwd()` to check your working directory and then move the data file there.)

These data are from *Stamey et al. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients, Journal of Urology 141(5), 1076-1083.* They studied the level of prostate specific antigen (PSA) and a number of clinical measures in 97 men who were about to receive a radical prostatectomy. The variables include the log(arithm) of PSA (lpsa), log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

(1) Check the structure of the data using `str(pr)`. Find out what is a Gleason score (e.g. use Google). Make a boxplot of `lpsa` levels for each value of Gleason score, i.e., on the x-axis you should have values 6,7,8 and 9 and for each of them you should have a boxplot that describes the distribution of the `lpsa` values for that Gleason score. (Hint: `boxplot(pr$lpsa ~ pr$gleason, ylab = "lpsa", xlab = "gleason score")`.)

(2) Test using `t.test()` whether mean `lpsa` values are similar in individuals with `gleason == 6` as in individuals with `gleason > 6`. What is the 95%CI for the difference in means and what is the P-value? What is your conclusion? (Hint: you can index data.table (`pr`) using conditions such as `pr$gleason == 6`. So pick column `lpsa` for all rows where `pr$gleason == 6` by `pr[pr$gleason == 6, "lpsa"]`.)

(3) Test using `t.test()` whether mean `lpsa` values are similar in individuals with `pgg45 <= 40` as in individuals with `pgg45 > 40`. What is the 95%CI for the difference in means and what is the P-value? What is your conclusion?

Problem 2.

Continue with the prostate cancer data set from Problem 1. We will look at two variables, one which is quite well modeled by a Normal distribution and the other which is not.

(1) Plot a histogram of `lpsa` values. Does it look like Normal? What is its mean and sd? If `lpsa` were Normally distributed with current mean and sd, what would be a point below which 60% of the population values would reside? (Hint: use `qnorm( )` with estimated mean and sd.) What is the corresponding empirical value from the data? (Hint: use `quantile( )`.) Make a QQ-plot of `lpsa` to compare it to a Normal distribution. Interpret the QQ-plot.

(2) Repeat the analyses of part (1) but changing `lpsa` to `lbph`.

Problem 3.

Read in 'systbp_ldlc.txt'. Plot LDL-cholesterol (ldlc) for females using a histogram. (Females have `sex == 2` in the data; males have `sex == 1`.) What is the mean and standard deviation of LDL-C in females? Does LDL-C look Normally distributed in females? (Make a QQ-plot and Shapiro-Wilk test.) Does `log(ldlc)` look Normally distributed in females? Answer the same question for males. Is there a difference in means between males and females (apply a t-test to the original LDL-C values and report the difference, its 95%CI and P-value)? Use also the non-parametric Wilcoxon rank sum test to compare the distribution of `ldlc` in males and in females.

Hint: Adapt the codes applied to blood pressure in Lecture 4.

Problem 4.

A sample of 20 drivers was chosen, and their reaction times in an obstacle course were measured (in seconds) before and after drinking two beers. The purpose of this study was to check whether drivers are impaired after drinking two beers.

Read in 'beers.csv' using `x = read.csv("beers.csv", header=T)`. Use `head(x)` to get an idea what's in there. Visualize the data by applying `stripchart()` to the differences of reaction times within individuals. Use these data to test whether drinking two beers changes the reaction time while driving. Give numerical estimate of the effect with confidence intervals and explain your conlcusion.

This example is taken from http://bolt.mph.ufl.edu/6050-6052/unit-4b/module-13/paired-t-test/ (EXAMPLE: Drinking and Driving).