Homework 4

Alla Tambovtseva

Deadline: 18 February, 23:59

Submission

Part 1: answers should be submitted via Google forms: https://goo.gl/forms/jo2DLiwWJxADQ8js1.

Part 2: the pdf-file with answers, code and graphs should be uploaded via Dropbox: https://www.dropbox.com/request/3srFNyJQ1BEU3VwRbQgO.

The pdf-file should not be necessarily created in RStudio, it can be a Word/Open Office file converted to pdf. You can create a text file, write anwers, copy your R code & outputs and add graphs (export pictures via Export in Plots tab in the right bottom corner in RStudio) and save as pdf.

An alternative way, more convenient, but more demanding: create an Rmd-file with texts and code chuncks, then knit it to Word via Knit button in RStudio and export to pdf.

Part 1

Problem 1

(Taken from OpenIntro Statistics 3)

For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(a) In a survey, one hundred college students are asked how many hours per week they spend on the Internet.

(b) In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"

(c) In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.

(d) In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.

(e) In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

Problem 2

(Taken from OpenIntro Statistics 3)

A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram below shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.

2.1 What is the point estimate for the average number of credits taken per semester by students at this college? What about the median?

2.2 What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

2.3 Is a load of 16 credits unusually high for this college? What about 18 credits? Explain your reasoning. *Hint:* Observations farther than two standard deviations from the mean are usually considered to be unusual.

2.4 The college counselor takes another random sample of 100 students and this time finds a sample mean of 14.02 units. Should she be surprised that this sample statistic is slightly different than the one from the original sample? Explain your reasoning.

Part 2

Important: all tasks in Problem 1 (except task 1 where you should load data) should be done with `dplyr` functions and a pipe `%>%`.

Problem 1

1.1 Load a `Cowles` data set using this link to a csv-file.

These data come from a study of the personality determinants of volunteering for psychological research.

Variables:

- `neuroticism`: scale from Eysenck personality inventory (the higher is the value, the more neurotic a person is);
- `extraversion`: scale from Eysenck personality inventory (the higher is the value, the more extravert a person is);
- `sex`: a factor with levels: `female` and `male`;
- `volunteer`: volunteeing, a factor with levels: `no`; `yes`.

See the full description here. See the description of Eysenck personality inventory here (you can take an online psychological test as well to see how it works).

1.2 Add a column `female` to the loaded data frame. It should be binary, and value 1 correspond to females, and 0 - to males.

Hint: consider the following example:

```
v <- c('a', 'b', 'b', 'a')
ifelse(v == 'a', 1, 0)
```

```
## [1] 1 0 0 1
```

1.3 Calculate how many volunteers and non-volunteers are there in this data frame.

1.4 Calculate the average and the median values of extraversion index for males and females.

1.5 Which number is higher: number of female volunteers or number of male volunteers? Provide both you code and answers.

1.6 Using `dplyr` create a summary for the variable `neuroticism` for males and females separately. This summaries should include: mean value, median value, standard deviation.

Problem 2

2.1 Use `dplyr` to get the mean value of people's level of extraversion for volunteers and non-volunteers separately (use a pipe `%>%` and corresponding functions). Who are more extravert: volunteers or non-volunteers? Provide your code and comments.

2.2 Use a two sample Student's t-test (`t.test` in R) to decide whether a difference in the average level of extraversion of volunteers and non-volunteers is statistically significant.

(a) State the null hypothesis you are going to test. State the alternative hypothesis.

(b) Perform t-test and provide your code.

(c) Can you conclude that the difference in the average level of extraversion of volunteers and non-volunteers is significant at the 5% significance level? Explain your answer.