ASDS - Applied Statistics with R

Fall 2018, YSU

Happy first Stat Homework!
Hope you will enjoy solving Stat homework problems ☺

Homework Grading Policy

You will have both theoretical problems, and practical problems to be solved by using R. Please submit the paper version of your homework by the deadline, and submit your program files by the end of the same day using the Google Drive folder that instructor will share with you.

ASDS - Applied Statistics with R

Fall 2018, YSU

Homework No. 01

Due time/date: 21:20, 14 September, 2018

Note: Please use R only in the case the statement of the problem contains (R) at the beginning. Otherwise, show your calculations on the paper. Supplementary Problems will not be graded, but you are very advised to solve them and to discuss later with TA or Instructor.

Problem 1. Describe the population and a representative sample for each of these cases. Describe also a sampling method.

a. We want to know the percentage of smokers in Armenia.
b. We want to know the percentage of female smokers in Armenia.
c. We want to study the relationship between the years of study and salary in Armenia.
d. We want to study how dangerous are right-hand-drive (right-wheel) cars in the sense of accidents in Armenia.

Problem 2. Give a real-life example for each of the following type data:

a. Univariate (1D), numerical variable;
b. Univariate (1D), categorical variable;
c. Bivariate (2D), both variables are numerical;
d. Bivariate (2D), both variables are categorical;
e. Multivariate (>2D), both numerical and categorical variables.

Problem 3. We want to make a statistic about the mean age and salary of all workers in Armenia. We choose 200 profiles of Armenians at random in Linkedin, write a message to the corresponding person and ask about if the person is working, and about the age and salary (also promising to keep the data confidential). Then we calculate means for obtained answers. Will the result be acceptable? Is this sampling method acceptable for the task? Explain your reasoning.

Problem 4. We have a population of size N, and want to choose a sample of size n ($n < N$).

a. How many different samples we can have?
b. We will use random sampling (the probability of any object/person to be chosen from the population is the same). What is the probability that some fixed object/person will be chosen in our sample?

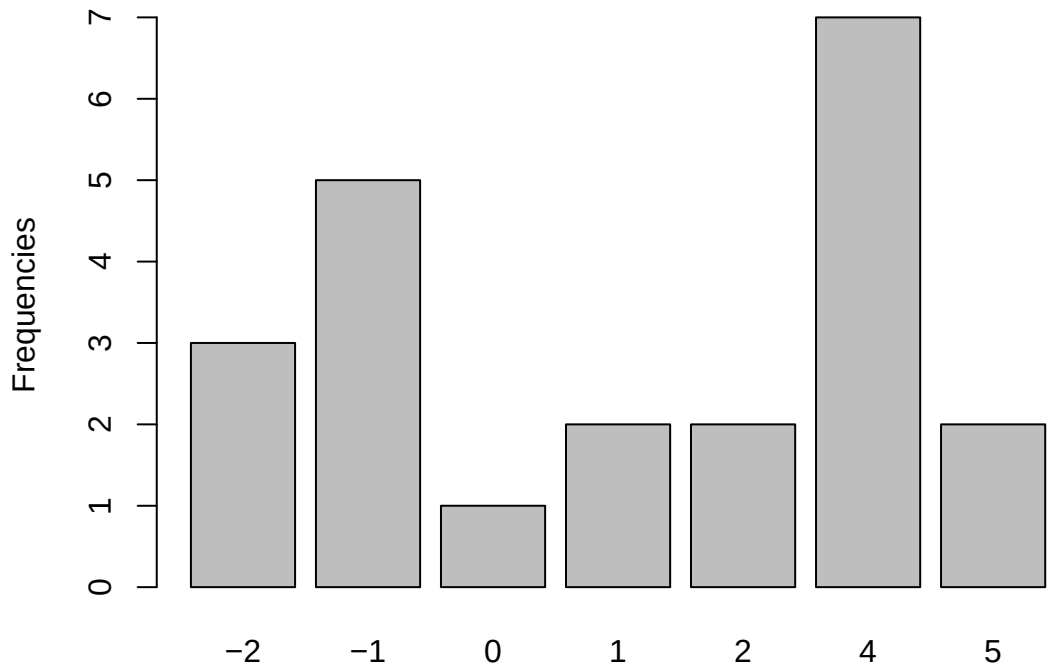

Figure 1: Barplot for a dataset

Problem 5. Fig. 1 gives the bar plot for some dataset:

a. How many observations we have?
b. What is the most frequent value in the dataset?
c. Construct the Frequency and Relative Frequency table for the dataset.
d. Can this data represent observations on the variable, which is, supposedly, normally distributed?

Problem 6. Download the Academic Faculty Salaries Survey from https://www.ams.org/profession/data/annual-survey/2017Survey-FacultySalaries-Report.pdf. On the fifth page (which is, actually, the 833rd page of the total survey), you will find two polygon plots for Statistics and Biostatistics Group Faculty Salaries. Describe, as much as possible, the information you can get from these plots (say, info about the Stat Full Profs, comparison between Stat Full Professors and Assistants Salaries, comparison between Stat and Biostat Full Prof Salaries etc.).

Problem 7. (Supplementary) One of my neighbors is suggesting me to buy his own crop of hazelnuts, from his own garden (bostan ⌣̈). He swears that the crop was an excellent one, and no huzelnut will have an empty core. I want to buy some 50kg of hazelnuts, but want to be sure that he is telling me the truth about the quality.

a. How to get the census about the ratio of empty/non-empty core hazelnuts?

b. Suggest me a non-destructive test to find the ratio of empty/non-empty core hazelnuts.

Problem 8. (Supplementary) We want to make a survey concerning the popularity of cheating at AUA: we want to know how many students used cheating at least once during their study. Well, although we will not record any names, there is a good chance that some students will not answer honestly to the question. Give suggestion how to design the survey to get acceptable results. Explain your reasoning.

Problem 9. (Supplementary) We have asked two persons to make a coin-tossing experiment (120 tosses) to obtain a random sequence of H, T-s of length 120. The recorded responses are:

$$HHTTHTTHTHHTTHTTTHHTHTHTTHHTHT$$
$$THTHTHHTHHTTHHTHHHTHTTHTTHHTHT$$
$$THHTHTTHHTHTHHTTHHHTTHTHTHTTTH$$
$$THTHTHTHTHHTTHTTHTHHTTHTTHHTHH$$

and

$$HHHTTTHHHTHTHHHHHHTHTTTHHTTHHT$$
$$HTHTTTHHTHTTHHHTHTTTTHTHHHTHHT$$
$$THHHTHHTTHHHHHHTHHTHTHHHTTTTTT$$
$$TTHHTHHHTHHHHTHHHHTHTHHTHHTHTH$$

One of the persons sent a fake sequence (was too lazy to perform the experiment). Who? Explain!