University of Tehran
ECE Department

Statistical Inference

Spring 2021

INTRODUCTION

In this project, we intend to study and analyze a series of real datasets with what you learned in this course. To begin analyzing a dataset, the first step is to get familiar with it. In the first step, this acquaintance can be made by observing the features of the dataset and distribution of the values and visualizing the data to make initial guesses about it. In the next step, by performing statistical tests, we make sure our guesses are correct and make our claims with certainty.

Datasets Description

Dataset name	Filename	Description
Healthcare	health.csv	This dataset includes some information regarding the health situations of around 5000 individuals as well as how much they yearly spend on their health bills.
University admissions	admission.csv	This dataset contains some information about some students applying for university admissions.
Students' performance	student.csv	This dataset includes the information about a sample of students studying in two different institutes as well as their grades in three different exams.

IMPORTANT NOTICES

- Use the R language in answering questions. Submit your codes in a separate file next to your report. Reports without R codes are pointless.
- In some datasets, you need to clean the data and convert the format and data type to more appropriate formats. So do this before answering the questions and explain the steps at the beginning of your report.
- If you need more categorical variables, you can add a new one to the dataset using some of your numerical variables. In this case, you need to describe the way you created the categorical variable from the numerical variable.
- In most of the questions, you should use the ggplot2 library to visualize and produce the desired charts.
- For each question, you need to fully explain your answer. An important part of the score will be attributed to your description. Drawing charts and performing calculations without sufficient explanations will result in losing the score. These descriptions show how much you understand the dataset. If you see interesting things in the diagrams, don't forget to mention them.
- When performing statistical tests, be sure to check the requirements for that test and write it down in your answer.

Question 0

By answering these questions, you will get valuable information about your dataset:

A. Briefly describe your dataset and why studying your dataset can be interesting?

B. How many variables (features) and cases does your dataset have?

C. Is there any missing value in your data? Provide a summary of a portion of missing values for each variable (feature) and describe how you handle these missing values for each variable (on what basis).

D. Using this elementary view of your dataset, which variables do you think maybe the most relevant (contain some important information)? Why?

Question 1

From your dataset choose a numerical variable and answer the following questions:

A. Plot a histogram with an appropriate bin size, then overlay that with the curve of a density. Based on your plot talks about the modularity of this variable.

B. Discuss the properties of the distribution of this variable. then compare it with the normal distribution, using an appropriate plot to show differences.

C. Calculate skewness and describe it.

D. Visualize the outliers then try to find what's the meaning of them.

E. Calculate mean, median, variance, standard deviation and describe each one.

F. Draw a density plot of this variable and add lines for the mean and median to it. What is the relationship between the mean, median, and density of this variable?

G. Categorize this variable into four intervals based on its mean and plot a pie chart that visualizes the frequency of these four categories. Your chart should be colorized and the labels should contain each category with its percentage.

H. Determine the upper and lower quartiles, whiskers, and the IQR by drawing a boxplot.

Question 2

From your dataset choose a categorical variable and answer the following questions:

A. find the frequency of each category and its percentage.

B. Plot a barplot for this variable and add percentage marks to it. Use different colors for each category.

C. Sort the categories by their frequencies, then using a horizontal barplot to show the result.

D. Plot a violin plot for this variable.

Question 3

From your dataset choose two numerical variables and answer the following questions:

A. First, try to guess the relationship between these two variables only with their descriptions.

B. Draw a scatter plot for two variables and describe the relationship between them based on the plot.

C. Calculate the correlation coefficient for these two variables.

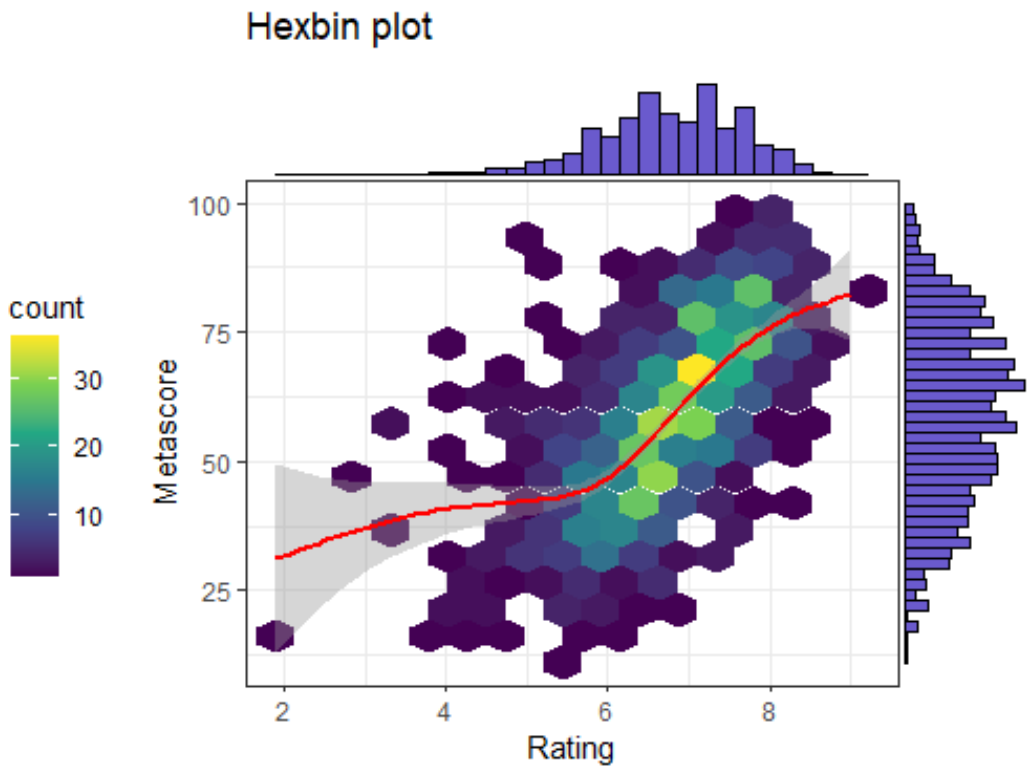
D. Describe the calculated correlation coefficient and examine your answer in part a.

E. test the significance of a correlation calculated in part c. what is shown by the p-value and what is the intuition of the p-value?

F. Select a categorical variable, and determine the samples by both symbol and color in a scatter plot that has been drawn in section “b”.

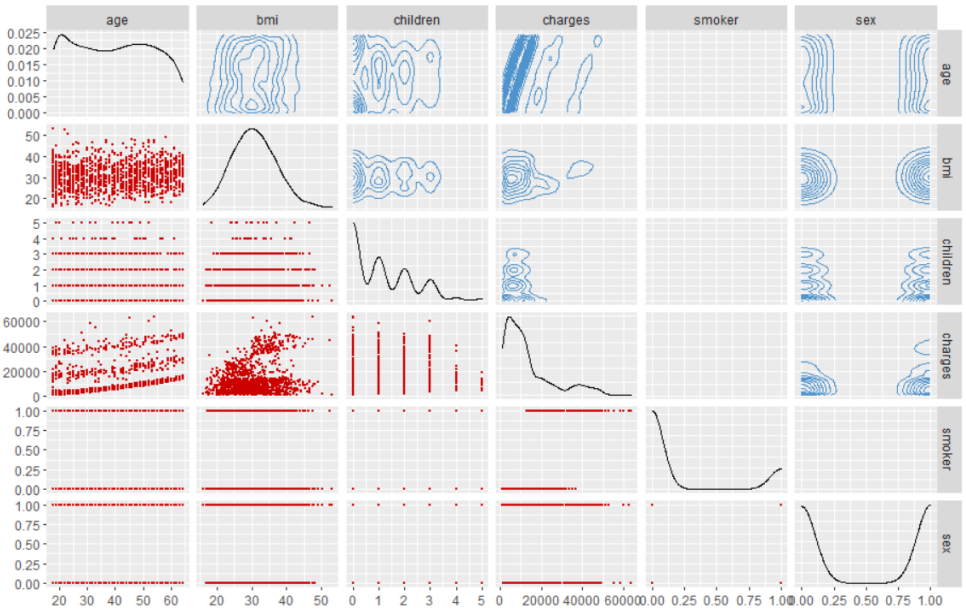
G. A hexbin plot is like a two-dimensional histogram. The data is divided into bins, and the number of data points in each bin is represented by color or shading. Draw a hexbin plot with marginal distribution and a fitting curve for your chosen variables. How do you interpret the resulting graph? Discuss the bin size and how it changes the result.

H. Draw the 2D density plot for chosen variables. How do you interpret the resulting graph? Describe the advantages and disadvantages of the 2D density and hexbin graph.

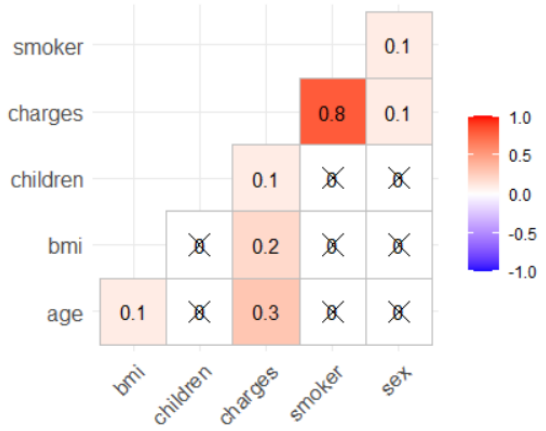


Question 4

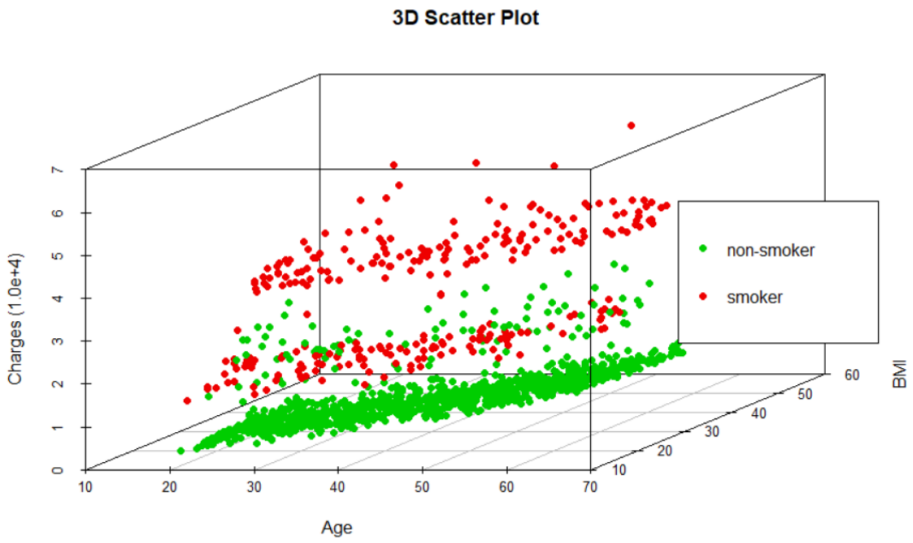
A. Display all the bivariate relations between the variables using a correlogram where each element is a scatter-plot between two variables. Can you find any meaningful pattern between them?



B. Create a heatmap correlogram from your variables. Annotate each cell with their corresponding Pearson's correlation coefficients and p-value as well. Use red for positive correlation and blue for the negative correlation. Highlight significant correlations.



C. Choose 3 numerical and 1 categorical variable from your dataset. Draw a 3D scatterplot for the numerical variables and use the categorical variable as points' color. Describe the relation between them.



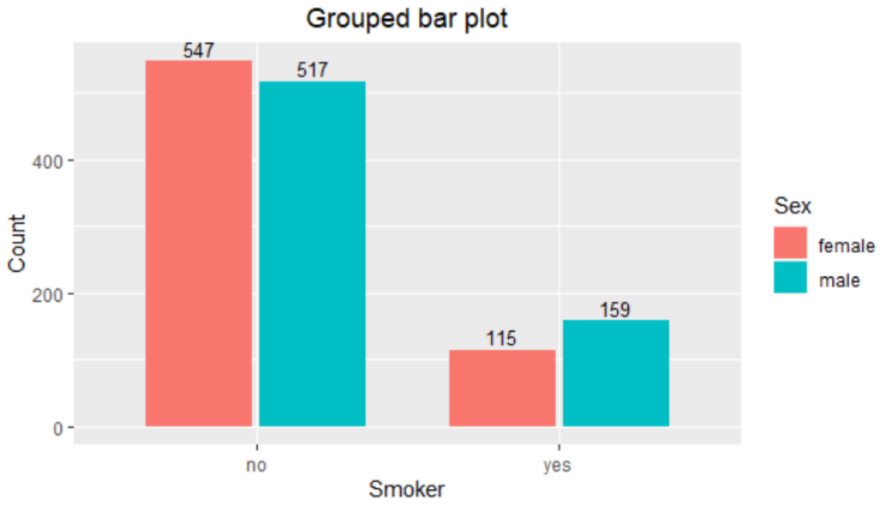
Question 5

For each below chart types, consider two categorical variables from your dataset that could be better described than others and then draw the chart:

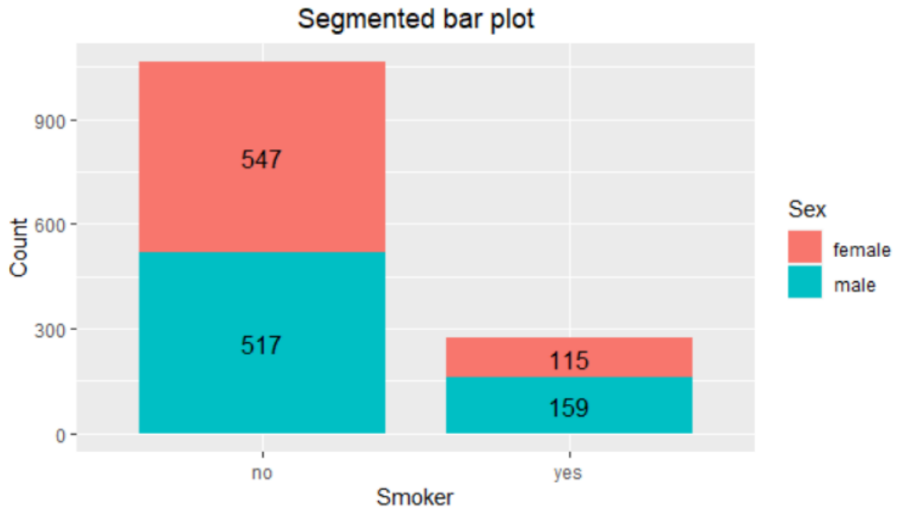
A. Frequency/Contingency table

	non-smoker	smoker	total
female	547	115	662
male	517	159	676
total	1064	274	1338

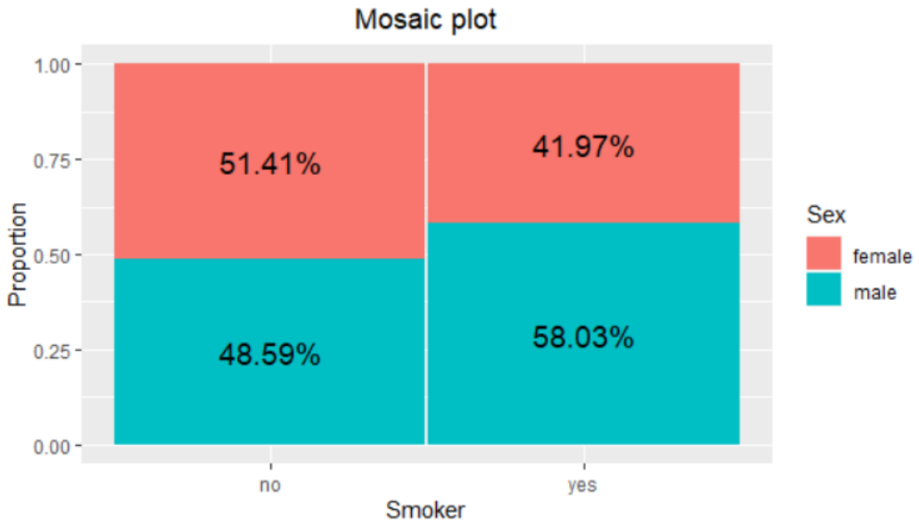
B. Grouped bar chart



C. Segmented bar plot



D. Mosaic plot



Question 6

Choose a numerical variable in your dataset:

A. Calculate a 95% confidence interval for the mean value of this variable.

B. Interpret this confidence interval. In this context, what does a 95% confidence level mean?

C. Plot the histogram of the variable and mark the mean of all the samples as a vertical line on top of the histogram plot. You must also mark the confidence intervals on the plot as two vertical lines.

D. For the mean value of this numerical variable, design a hypothesis test and by finding the p-value, confirm or reject your assumption. What does this p-value signify?

E. Based on the confidence interval you calculated in part a, does the data support the hypothesis that you have designed? Explain it.

F. Calculate type II error. What does this value mean?

G. Calculate the power and Explain the relationship between the power and the effect size.

Question 7

In this question, you will conduct a paired and a non-paired hypothesis test for two numerical variables.

A. Choose a random sample of 25 data points from the dataset and in this sample, select two numerical variables. In another word, you choose 25 pairs of samples in a way that every sample comprises two values associated with the two variables you've picked before, and, for each sample, the two values correspond to a single row in the dataset.

    a. Should we use a t-test or z-test? Explain it.
    b. By conducting a hypothesis test, explain whether there is a significant difference between the mean values of these two variables.

B. Now, draw 100 independent samples from the dataset for each of these two variables. Then, conduct a hypothesis test to inspect whether there is a significant difference between the mean values of these two variables. Are the results of the test consistent with the 95% confidence interval?

Question 8

Choose a numerical variable that has outliers and we cannot apply CLT based methods we have learned so far.

A. Calculate a 95% confidence interval for the median of this variable using the percentile method.

B. Pick a random sample of size 20. Then, using the bootstrapping method, calculate a 95% confidence interval for the mean of this variable using the standard error method.

C. Is there any noticeable difference between these two calculated confidence intervals? Explain your reasoning.

Question 9

Answer this question based on the dataset assigned to you. This requires that you verify your answer by performing statistical tests and satisfying the required conditions. Perform random sampling if necessary. Drawing a chart to clarify your answer has an extra score.

Healthcare Dataset

Inspect, using ANOVA analysis, whether there is any significant difference between the mean healthcare bills of the individuals associated with different work types. Explain your results.

University Admissions Dataset

Is there any difference between the admission chances of the applicants with different university ratings?
Use ANOVA analysis to inspect that and explain your results.

Students' Performance Dataset

Is there any difference between the average of the total scores (G1+G2+G3) of the students with different failure counts (failures)?
Use ANOVA analysis to inspect that and explain your results.