Statistical Inference

In this project, we intend to study and analyze a series of real datasets with what you learned in this course. The first step to begin analyzing a dataset is to get familiar with it. In the first step, this acquaintance can be made by observing the dataset features 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.

To answer each question, you have to fully explain the meaning of your analysis and interpret the generated plot and what you observe, even when it is not explicitly stated in the question. The more reasonable your analysis is, the more positive effect it has on the acquired grade of the corresponding question.

Note that there is no one way to solve each question correctly. Furthermore, whenever you need to do hypothesis testing, you must check all of the pre-requisite conditions (such as sample size, skewness, etc.). Finally, the validity of your results should be discussed.

Datasets Description

Choose one of the following datasets. For more information about your dataset, please refer to the mentioned references.

Dataset Name	Description
Breast Cancer	This dataset includes information about breast cancer features as well as the diagnosis of the type of cancer. (This dataset contains two features of race and marital status. For more information, go to [3])
Heart Disease	This dataset includes information about some heart disease features as well as the "target" field, which refers to the presence of heart disease in the patient. [4]
Non-Voters	This directory contains the data behind the story Why Many Americans Don't Vote. The poll was conducted among a sample of U.S. citizens that oversampled young, Black and Hispanic respondents, with 8,327 respondents, and was weighted according to general population benchmarks for U.S. citizens. A voter company matched what respondents said and what they actually did. The data included here is the final sample we used: 5,239 respondents who matched to the voter file and whose verified vote history we have, and 597 respondents who did not match to the voter file and described themselves as voting "rarely" or "never," all of whom have been eligible for at least 4 elections. [5]
Nutrition Studies	This directory contains data and code behind the story You Can't Trust What You Read About Nutrition. Many diet and nutrition studies include multiple variables with vast amounts of data, making it easy to p-hack your way to false results. We learned this firsthand when we invited readers to take a survey about their eating habits known as the food frequency questionnaire and answer a few other questions about themselves, so we ended up with 54 complete responses. [6]

Dataset Name	Description
Airbnb Open Data	Airbnb, Inc is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking. The company was founded in 2008. Airbnb is a shortened version of its original name, AirBedandBreakfast.com.Since 2008, guests and hosts have used Airbnb to travel in a more unique, personalized way. As part of the Airbnb Inside initiative, this dataset describes the listing activity of homestays in New York City. [7]
Car Insurance Claim Prediction	This is the training dataset that contains all the independent and target features. It contains information on policyholders having the attributes like policy tenure, age of the car, age of the car owner, the population density of the city, make and model of the car, power, engine type, etc., and the target variable indicating whether the policyholder files a claim in the next 6 months or not. [8]
Housing Price in Beijing	This dataset describes features like Longitude, Latitude, DOM (days on market), Price, SquareMeter, Living Room, Kitchen, Bathroom, Building Type, Construction time, etc., for houses in Beijing. These features can help predict the price of a house with specific characteristics. [9]
Maximum Credit	This dataset is about the Maximum Credit that is available for different people with different backgrounds and characteristics, like their current loan amount, their annual income, how many years they have used credit, whether they have been bankrupt or not, etc. [10]

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 how you created the categorical variable from the numerical variable.

- In most questions, you should use the ggplot2 library to visualize and produce the desired charts. You can find more about this elegant visualization package in references [1], [2].

- 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 may be the most relevant (contain some important information)? Why?

Question 1

Choose one numerical variable from your dataset and answer the following question:

A. Plot a histogram with an appropriate bin size, then overlay that with the curve of a density, is the distribution approximately normal?

B. Based on the previous plot in section “A” talk about the modularity and skewness of this variable and describe it.

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

D. What are the outliers in this variable? Determine the outliers and their quantity, then try to interpret their meaning.

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

F. 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.

G. 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?

Question 2

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

A. Create a frequency table for this variable.

B. Plot a horizontal bar plot, sort by frequency, and use different colors for each category.

C. Plot a bar plot for this variable and add percentage marks to it.

D. Plot a violin plot for this variable

Question 3

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

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

B. Select a categorical variable, and determine the samples either by the symbol or by the color (or by both) in a scatter plot that has been drawn in section “A”. Does the relation still hold for different categories?

C. Calculate the correlation coefficient for these two variables. Using the “cor.test” function, we can also test the significance of a correlation. Are the variables correlated? According to the test, what is shown by the p-value, and what is the intuition of the p-value?

D. A hexbin (Figure 1) plot with marginal distribution is like a two-dimensional histogram. The data is divided into bins, and the color strength of each bin represents the number of data points in that bin. Also, each dimension has its own distribution in front of its axis. Draw the hexbin plot with marginal distribution for chosen variables. What is your interpretation? Discuss the bin size and how it changes the result.

E. 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.

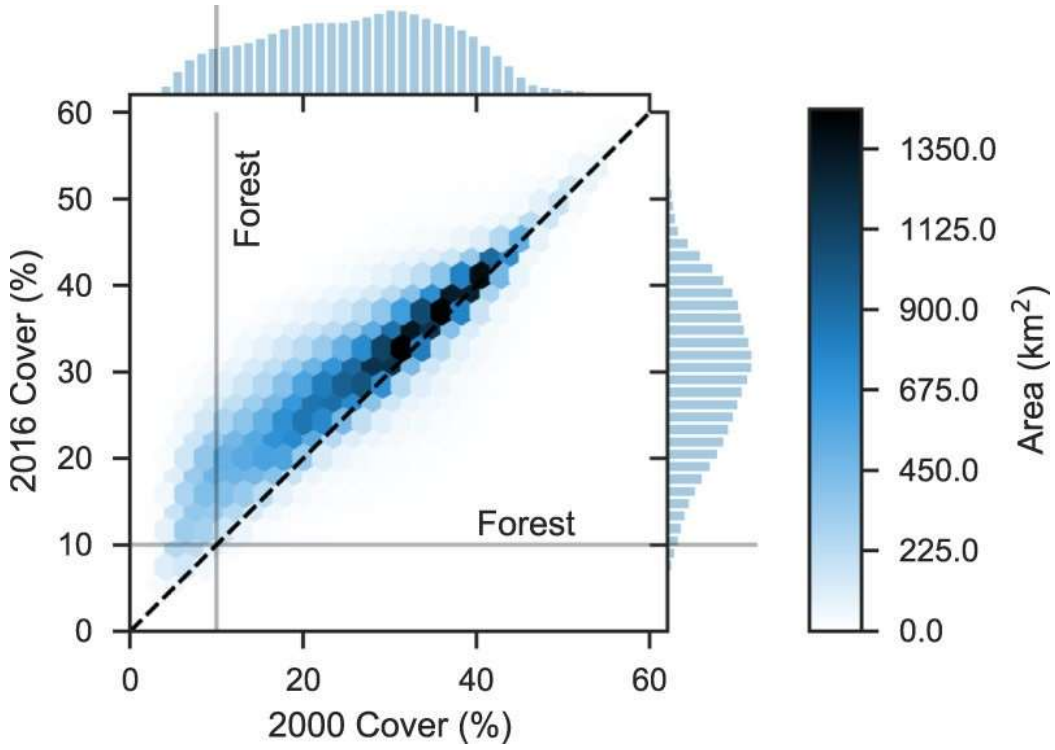


Figure 1. Hexbin (2D histogram) and marginal histograms of canopy cover in 2000 and 2016. Forest is defined as areas with at least 10% cover

Question 4

Please answer the following questions:

A. 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 the positive correlation and blue for the negative correlation. Highlight significant correlations.

B. 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?

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

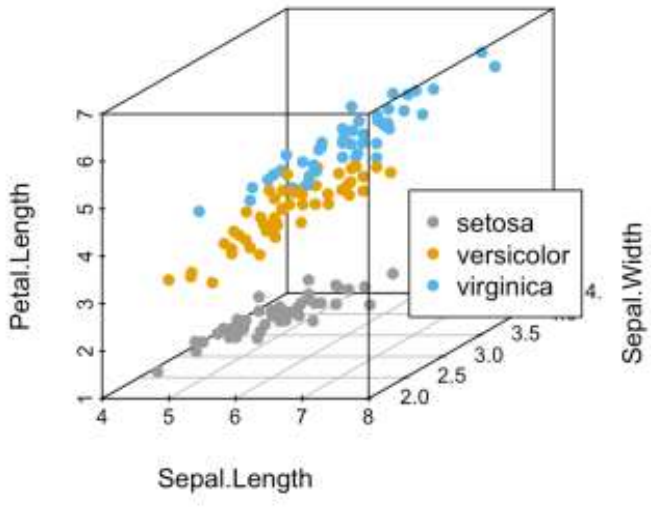


Figure 2. 3D-Scatterplot with a categorical variable coloring

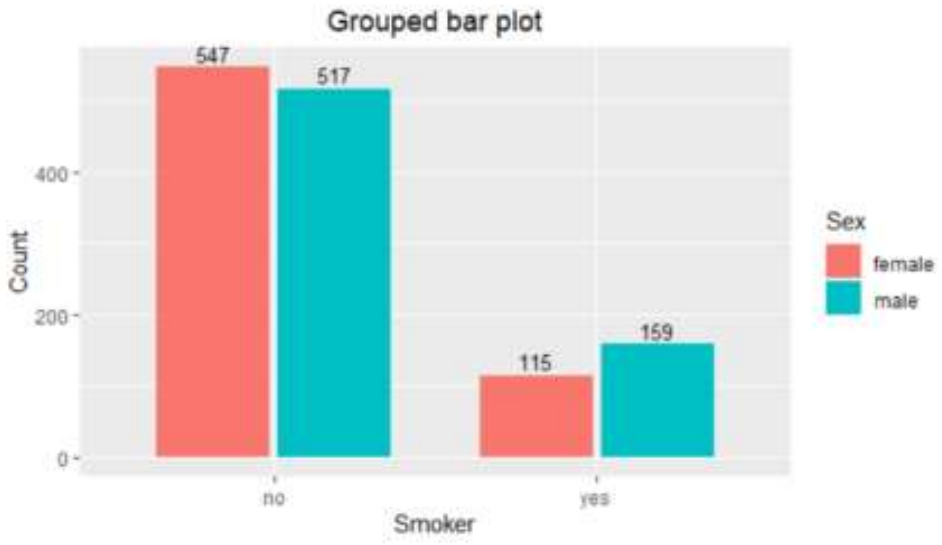
Question 5

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

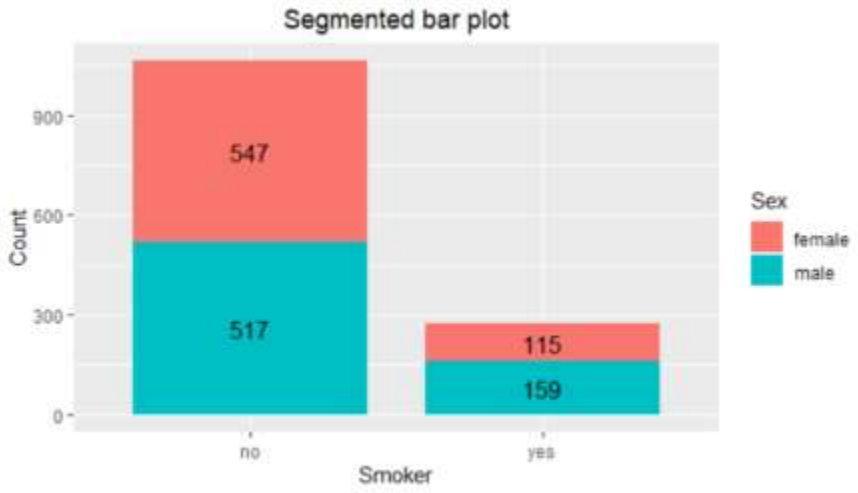
A. Contingency table

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

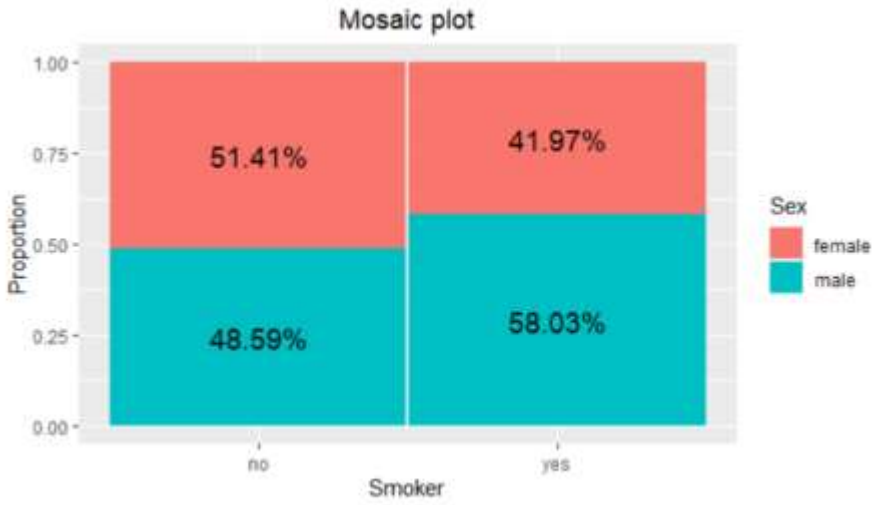
B. Grouped bar chart



C. Segmented bar plot



D. Mosaic plot



Question 6

In this question, you will conduct a hypothesis test for two numerical variables.

Choose a random sample of 25 data points from the dataset and choose two numerical variables that are not of a corresponding quantity. This data can be used to compare the average quantity between the two variables.

A. What is the best method for testing our hypothesis? Is it a t-test or a z-test? Explain it.

B. Design a hypothesis test to see if these data provide convincing evidence of a difference between mean values. Does the result agree with the 95% confidence interval?

Question 7

Choose a numerical variable from your dataset:

A. Calculate a 98% confidence interval for the mean of this variable.

B. Interpret this confidence interval. In this context, what does a 98% 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 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 and show the interval on the histogram.

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. Also, plot the bootstrap distribution on the dot-plot.

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

Question 9

Choose a numerical and a categorical variable with more than two levels. Divide observations of this dataset into different groups such that each group represents a level of the chosen categorical variable,

A. Use the ANOVA test and compare the mean value of the numerical variable in the groups.

B. Choose two of the groups, perform a hypothesis test for the mean difference of the selected numerical variable in these groups and calculate the p-value. Make a decision and explain the result using a significance level of 5%.

References

[1] Intro to Data Visualization with R & ggplot2 (Link)

[2] Data visualization with R and ggplot2: the R Graph Gallery (Link)

[3] Breast Cancer DataSet (Link)

[4] Heart Disease DataSet (Link)

[5] Non-Voters DataSet (Link)

[6] Nutrition Studies DataSet (Link)

[7] Airbnb Open Data DataSet (Link)

[8] Car Insurance Claim Prediction DataSet (Link)

[9] Housing price in Beijing DataSet (Link)

[10] Maximum Credit DataSet (Link)