Statistical Inference

Project Phase 1

University of Tehran

Spring 2019

Datasets

One of the following datasets is assigned to you. The ID of your assigned dataset is your student ID modulus 4.

For example:
810197342 mod 4 = 2 -> “House Sales” dataset

ID	Dataset	#Obs	#Var	Description
0	Google Play Store	10.8k	13	App information scraped from the Google Play Store.
1	IBM HR Analytics	1470	35	A dataset designed to analyze employee performance and attrition.
2	House Sales	21.6k	20	This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.
3	Adult Income	45.2k	15	Statistics of adult income and education from various countries.

Introduction

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 acquired grade of the corresponding question. Note that there is no one way to correctly solve each question.

Furthermore, whenever you need to do hypothesis testing you must check all of the pre-requisite conditions (such as sample size, skewness etc.). Finally validity of your results should be discussed.

Tools

We recommend `ggplot2` for visualizations and using either the `data.tables` or `dplyr` packages for easier data manipulation.

Question 0

Briefly describe your dataset and why studying your dataset can be interesting. Describe each variable existing in the dataset and specify its type. By doing this simple task, you gain an initial understanding of your dataset. Knowing your dataset is the first step in data science and it usually works as a helpful step toward more advanced analyses.

Using this elementary view of your dataset, which variables do you think may be the most relevant (contain some important information)? Why? Note that in this section, we only want you to express your intuition about the relationship between the variables without doing any quantitative analysis.

Does your dataset have missing values? Provide a summary on portion of missing values for each variable (feature) and describe how you handle these missing values for each variable (on what basis).

Question 1 - Numerical Univariate Analysis

Select one numerical variable from your dataset.

1. Plot a histogram for this numerical variable with an appropriate bin size.
2. Plot the density plot for this numerical variable
3. Describe modality and skewness (calculate skewness).
4. Calculate mean , variance, standard deviation.
5. Draw the boxplot, determine the upper and lower quartiles, whiskers, and the IQR.
6. Determine the outliers for this variable.

Question 2 - Bivariate Correlation

Select two numerical variables from your dataset.

1. Draw a scatterplot for these two variables.
2. Describe the two variables relation in words. Can you explain this relation?
3. Select a categorical variable and either by symbol or by colour (or both); distinguish the samples in the scatterplot. Does the relation still hold for different categories?
4. Calculate the correlation coefficient between these two variables. Using the `cor.test` function we can also test the significance of a correlation. According to this test, what is the probability that the two variables, in the population from which this sample was taken, are in fact not correlated?
5. What is the best estimate for the population correlation coefficient? With only 5% probability of being wrong, what are the lowest and highest values this coefficient could in fact have? Does this test prove causation?
6. Draw a 2D density plot and a 2D histogram for the two variables. How do you interpret the resulting graphs? What are the advantages and disadvantages of each plot?

Question 3 - Multivariate Correlation

Consider a group (more than 3) of numeric variables from your dataset.

1. Display all the bivariate relations between the variables using a correlogram[1] where each element is a scatter-plot between two variables.
2. Describe the relations between the variables. Are there any interesting patterns?
3. Create a heatmap correlogram from your variables. Annotate each cell with their corresponding Pearson's correlation coefficients. Use red for positive correlation and blue for negative correlation.
4. These simple correlations show how each two variables are related, but this leaves open the question as to whether there are any underlying relations between the entire set. Give an example of how bivariate correlation may fail to take into account the relations between a group of variables.

Question 4 - Categorical Univariate Analysis

Choose a categorical variable.

1. Plot the barplot for this variable.
2. Create a frequency table for this variable.
3. Plot a violin plot for this variable.

[1] https://www.r-graph-gallery.com/correlogram/

Question 5 - Categorical Bivariate Analysis

For each of the following chart types, select two categorical variables from your dataset whose relationship can be best described by that chart and then draw the chart:

1. Grouped bar chart
2. Contingency table
3. Segmented bar plot
4. Spine plot

Question 6 - Distribution Analysis among Groups

In certain datasets we would like to compare the distribution of a numerical variable within different categories of another variable. Visualizing a distribution can be achieved using various plots such as box-plots, jitter points and violin plots. Select a sample size of 200 and a categorical and numerical variable from your dataset.

1. Use the mentioned plots to show how the distribution varies within the different categories of your categorical variable
2. Compare the strengths and weaknesses of these plots.
3. Interpret the data from these plots. What have you learned about the variables?

Question 7 - Univariate Hypothesis Testing

Choose a numerical variable from your dataset.

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

2. Interpret this confidence interval.

3. Design a hypothesis test for the mean of this variable with a power of 50%. Calculate the number of samples required, take the samples and calculate the p-value. Confirm or reject your assumption.

4. Calculate the Type I and Type II error.

Question 8 - Bivariate Hypothesis testing

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 which are of corresponding quantity (e.g sales total for two different years). We would like to use this data to compare the average quantity between the two variables.

1. Should we use a one-sided or a two-sided test? Explain your reasoning.
2. Should we use a t-test or a z-test? Explain your reasoning.
3. Calculate a 95% confidence interval for the difference of means
4. Design a hypothesis test to see if these data provide convincing evidence of a difference between mean values. Does the result agree with 95% the confidence interval?

Question 9 - Analysis of Variance among Groups

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,

1. Use the ANOVA test, compare the mean value of the numerical variable in the groups.
2. 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%.