Using R in Financial Statistics (spring 2019)
Final Report

Due time: June 28 (Friday), 13:00

Part 1

The dataset "survey" is stored in the R package MASS. This data frame contains the responses of 237 Statistics students at the University of Adelaide to a number of questions, and the below variables are available: Sex, Wr.Hnd, NW.Hnd, W.Hnd, Fold, Pulse, Clap, Exer, Smoke, Height, M.I and Age. You can use help(survey) to check the details of those variables. Construct a subset of data containing all observations without any missing value and answer the following questions based on this new data set. (25 points)

a) Report descriptive statistics of the data set obtained in (a).

b) Use boxplot to show the distributions of the height of male and female students.

c) Which numerical variables might have an influence on the student's pulse?

d) Is the probability of a student clapping his/her left hand on top less than 0.2?

e) Is the span of the writing hand in general larger than the span of the non-writing hand?

Hint: Note that in a) there are categorical variables. For d) and e) you need to use hypothesis testing.

Part 2

In probability theory, the central limit theorem (CLT) establishes that, in some situations, the mean of the independent random variables tends toward a normal distribution even if the original variables themselves are not

normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions. The following is a version of CLT. Let $X_1, X_2, \dots, X_n$ be a random sample of size n, that is, a sequence of independent and identically distributed (i.i.d.) random variables drawn from a distribution with expected value given by μ and finite variance given by σ^2. Let $\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$. According to CLT, $\bar{X}_n \sim N(\mu, \sigma^2/n)$ as n becomes large enough.

Now suppose each time we draw 10 independent samples from an exponential distribution with mean 5 and calculate the sample mean. We replicate the above procedure for 200 times. Answer the following questions. (20 points)

a) According to CLT, what is the approximated distribution of the sample means?

b) Draw the density plots of the sample means and its approximated distribution on one graph.

c) Show the qq plot of the distribution of the sample means.

d) What conclusion can you draw from this simulation?

Part 3

Suppose we have a stock whose current price is 10.3 dollars per share. We use X to denote the change of the stock price tomorrow (it equals current price - tomorrow's price, that is, the loss). Suppose that with probability 0.75 the random variable X follows a t-distribution with degree of freedom 4, with probability 0.15 it follows an exponential distribution with mean 2, and otherwise it follows a uniform distribution on the interval [-3, 3]. Answer the following questions. (15 marks)

a) Draw 5000 random samples of X. Here, if we have a sample whose value is greater than 10.3, we just set it to be 10.3. Show the density plot of the samples.

b) The Value-at-Risk (VaR) is defined to be the critical value L such that $P(X \geq L) = \alpha$, where α is the confidence level. The meaning of VaR is that we have confidence that the loss will not exceed the critical value L with probability $1 - \alpha$. Given $\alpha = 0.05$, find the VaR of the samples obtained in a).

c) The conditional Value-at-Risk (CVaR) is defined to be the expected value of the loss given that it is greater than the critical value L, i.e., $E(X \mid X > L)$. Find the CVaR of the samples obtained in a).

Part 4

The file "customer_analysis.csv" contains 1038 observations of the transactions made in a retail store. Given this data set, we want to study the customer purchase behavior. Specifically, we are interested in finding the variables that may affect the dependent variable (the amount of purchase) rather than making a prediction. The meanings of the variables in the data set are explained below.

- "Customer_ID": the id of the customer.
- "Gender": the gender of the customer with M meaning male and F meaning female.
- "Age": the age interval of the customer belongs to.
- "City_Category": the category of the city where the customer comes from.
- "Stay_In_Current_City_Years": the length of years that the customer stays in the current city.
- "Marital_Status": the marital status of the customer with 0 meaning single and 1 meaning married.
- "Purchase": the amount of dollars that the customer spent.

Conduct a regression analysis on this data set. Report your analysis and your conclusions from the following aspects: data exploration, modeling (baseline model and alternatives), results and interpretation of the fitted model, and model assumptions. You may use graphs and tables if necessary. (40 points)

侧重点是经济学上的意义，而不是数据挖掘，或者是 R 方够不够大

题目：R 语言期末报告之类

Submission

Save the source code as **final.R**

Write a report and attach *the official report title page*. Save this as **finalReport.pdf**

Due time: **June 28 (Friday), 13:00**

Note: Do NOT just copy your running results. Use your own words to explain your reasonings and conclusions with supporting information (graphs, tables, etc.). The official report title page can be downloaded from https://jwb.szu.edu.cn/info/1074/1077.htm

[Important]

1) Send the above two files together to email: hjpszu@163.com, with email title RFS_FinalReport_studentID_name
E.g.: RFS_FinalReport_2018123456_张三

You will receive an auto-reply only if your email is with a correct title.

2) In addition to 1), print out your *report* (not source code) with the *official title page*, and bring it to Wen Ke Lou 2613 between **13:00 and 17:00 of June 28 (Friday)**.