Statistics for Business Analytics I
R Laboratory Homework Assignment II

Introduction

This assignment is designed to solidify your understanding of some of the concepts in statistics that you learned in this class. As such, we will sometimes ask you a question that requires an explanation. Please write your explanation as regular English text in your markdown file, not as comment in your code. Please do not spend time writing long explanations, as 1-3 short sentences will generally more than suffice. To do well in this assignment, please be sure to review the material from the labs we have done together as you will find significant help from these. For the first two questions in particular, I also recommend you consult your course notes. There are just 4 questions to this assignment that cover, in order: confidence intervals/hypothesis testing, the central limit theorem, ANOVA, and multiple linear regression. Finally, you should remind yourself of the instructions on how to submit an assignment by looking at the instructions from the first assignment. Good luck!

Questions

Question 1 (15 points)

In this first question, we will give you a series of scenario for which we want you to select the right approach possible and implement it in R. Then you will have to report on something like a p-value or to interpret the results. The material will cover confidence intervals and hypothesis testing.

1) The ABC Corporation claims that its new gasoline additive for passenger cars will enhance the mileage per tankful of gasoline if a can of their product is added to a full tank of gasoline. Five vehicles were tested, and the number of miles per tankful was measured. Before using the additive, the cars had the following values: 370, 385, 375, 380, 378. After using the additive, the corresponding miles per tank that the vehicles registered were 387, 397, 380, 392, 389. Run a suitable hypothesis test that checks whether there is an increase in efficiency as the corporation claims and comment on the results. Use a 95% confidence interval.
2) From a random sample of 200 voters it was found that 110 were in favor of a particular piece of legislation. Is opinion equally divided on this legislative issue? Use $\alpha = 0.05$. What is the p-value for this test?
3) Fenway Park needs to purchase spotlights that exhibit long life as well as uniformity of operating life. Past experience dictates that the variance of bulb life is 230 $(\text{hours})^2$. A sample of n = 16 bulbs is obtained from a new

vendor wishing to get the lighting contract. It was found that the mean operating life for this sample was 1020 hours^2. Construct a symmetric 98% confidence interval around this mean.

4) Suppose that you did not know the variance in the previous problem but you measured the variance in your sample to be 275 (hours)^2. Recompute the confidence interval with this new information and setting.
5) In a random sample of 500 tulip bulbs taken from a normal population 476 of them bloomed. For α=0.05, would you reject the claim that at least 90% of the bulbs will bloom? What is the p-value for this test?
6) Generate 100 samples from a gamma distribution with shape parameter 1 and scale parameter 1. Now generate 100 samples from an exponential distribution with $\lambda = 1$. Finally, run a hypothesis test that these two samples came from the same distribution and explain the result.
7) Suppose that the scores of students on a specialized dexterity test were $92, 83, 95, 96, 85, 61, 76, 80, 92, 87$. Run a hypothesis test that the median of the underlying distribution is 82 with 95% confidence.

Question 2 (10 points)

In this question we want to reinforce your understanding of the central limit theorem (CLT), which is a very important theorem in statistics. Please refresh your notes about what the theorem says. In what follows, our goal is to witness how the distribution of sample means approaches the normal distribution when the size of the samples is large enough and to see that this holds for different distributions (in your notes you saw this just with the exponential distribution).

1) Before generating random samples, use the set.seed() command to allow your experiments to be reproducible.
2) Now create 10000 samples of size 1000 each, drawn from a Poisson distribution with parameter $\lambda = 2.7$. Please try to do this with a function from the apply family, rather than using a for-loop if possible.
3) Compute a vector with the means of these samples
4) Plot a density (HINT: look at my ggplot2 examples for this) of this vector of means. Is it what you expected?
5) Compute the mean of your sample means and state what the expected value of this is
6) Finally, run a hypothesis test to check the normality of your vector of means

Repeat and answer the questions in steps 1 - 6 above with a Binomial distribution parameterized by size = 20 and p = 0.3.

Question 3 (10 points)

In this question we want you to perform a one-way ANOVA in the same vein as we did in the lab. The data set for this example is in the accompanying CSV file

titled anova_data.csv. This contains data on several different countries of the world collected by the International Agency for Research on Cancer in 2002. For this problem, we are especially interested in how the average employment rate varies for countries in different parts of the world. There is a column that lists continent, which has Eastern and Western Europe (EE and WE respectively) separately so it is not traditional continents but we will use it as is.

Here are the steps we want you to take:

1) Load the csv file into a data frame in R. For your R markdown file please assume that the file is in your working directory (i.e. do not use full paths)
2) Plot the mean value of *employrate* for each different continent
3) Run one way ANOVA on the *employrate* variable using *continent* as the grouping factor and in your own words carefully explain what the results of the ANOVA tell us
4) Now run a Tukey test and compute the number of continent pairs whose means are significantly different from each other.
5) Finally, test the homeogeneity of the variances and comment on how this might affect the analysis
6) Have a think about what sort of conclusions you can draw from these data. For example, is it right to conclude that living in Eastern Europe causes people to have a lower chance of finding a job when they graduate?

Question 4 (15 points)

In this question we want you to review the basics of multiple linear regression. The data set we will use is the Boston housing data set (https://archive.ics.uci.edu/ml/datasets/Housing). The accompanying file, *housing.data* contains the data which you can also find by following the previous link.

1) Load the data in *housing.data* into a data frame
2) Use the first column from the following table to give suitable names to the columns in your data frame (the table has the variables in order)

CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of African-Americans by town
LSTAT	% Lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

3) Train a linear regression model with the final column (MEDV) as the dependent variable
4) How well would you say that your model fits the data? Argue your case by using appropriate numbers from the model's summary.
5) The model's summary also gives you the results of various t-tests run on the model coefficients. What inferences can you draw from what you see?
6) Carry out stepwise regression on this model and report which variables remain in the model
7) What is the coefficient of variable CRIM in your model (after stepwise regression)?
8) Draw a Q-Q plot for your model's residuals and comment on the results.