CS5701(Quantitative Data Analysis)

Lab 2: Part 2

The aim of this part of the lab exercise is to give you practical experience in working with confidence intervals in R Studio.

1 Confidence Intervals

1. Open R Studio and open a new project
2. Save the auto-mpg.csv file in the same folder as this new project
3. Open a new R Markdown file to use for this analysis
4. Read in the auto-mpg.csv data set into a new R notebook
5. Explore the data (using summary statistics and plots) and check which columns from the continuous variables look normally distributed
6. Compute a 95% confidence interval for the mean for one of the variable that appears to be normally distributed
7. The model years range from the 70s and the 80s. What proportion of cars are from the 80s?
8. Compute a 90 % confidence interval for the proportion of cars from the 80s?

Extension Compute the confidence intervals for the mean using the "bootstrap method".

1.1 About the data

- mpg: continuous
- cylinders: multi-valued discrete
- displacement: continuous
- horsepower: continuous
- weight: continuous
- acceleration: continuous
- model year: multi-valued discrete
- origin: multi-valued discrete
- car name: string (unique for each instance)

Dataset source: UCI Machine Learning Repository

2 Distributions - OPTIONAL Extension

1. Assume that among diabetics the fasting blood level of glucose is approximately normally distributed with a mean of 105 mg per 100 ml and a standard deviation of 9mg per 100ml.

   (a) Plot the density function for this distribution.

   (b) What proportion of diabetics have levels between 90 and 125 mg per 100ml? (This quantity is represented by the area under the normal curve between x=90 and x=125)

   (c) You should see from the plot that most of the area under the normal curve is contained between the two vertical lines. Therefore we should expect that the proportion of diabetics with levels between 90 and 125 mg per 100 ml to be high. Use the `pnorm` function to calculate this proportion. (you may need to use it twice)

   (d) What level cuts off the lower 10% of diabetics? (hint: use `qnorm`)

3 R code

You may find this code useful to complete this worksheet:

```
#to read in a csv file
x<-read.csv("filename.csv")

# to explore the contents of the data
summary(x)
# confidence interval for proportion
prop.test(x, n, p = NULL, conf.level = 0.95)

#confidence interval for mean
t.test(x,conf.level=0.9)
# for the binomial distribution
# the probability mass function for a given number of successes $v$, number of
    trials $n$ and probability of success $p$
dbinom(v,n,p)
#the cumulative distribution function for the same as above
pbinom(v,n,p)

#For the Normal distribution (the PDF) in order to find the y-values
    corresponding to a range of x values $(a,b)$
#with a mean of $m$ and a standard deviation of $sd$
dnorm(x, mean=m, sd=sd)

#the CDF for a given value of x
pnorm(x, mean, sd)

#the x value that corresponds to a given CDF
qnorm(p, mean, sd)

#for Z value the standard normal distribution x values:
qnorm(1- significance level)
```