# Problem Set 1

Applied Stats II

Due: February 14, 2022

## Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub in `.pdf` form.
- This problem set is due before class on Monday February 14, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

## Question 1

The Kolmogorov-Smirnov test uses cumulative distribution statistics test the similarity of the empirical distribution of some observed data and a specified PDF, and serves as a goodness of fit test. The test statistic is created by:

$$D = \max_{i=1:n} \left\{ \frac{i}{n} - F_{(i)}, F_{(i)} - \frac{i-1}{n} \right\}$$

where $F$ is the theoretical cumulative distribution of the distribution being tested and $F_{(i)}$ is the $i$th ordered value. Intuitively, the statistic takes the largest absolute difference between the two distribution functions across all $x$ values. Large values indicate dissimilarity and the rejection of the hypothesis that the empirical distribution matches the queried theoretical distribution. The p-value is calculated from the Kolmogorov- Smirnoff CDF:

$$p(D \leq x) \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2 \pi^2 / (8x^2)}$$

which generally requires approximation methods (see Marsaglia, Tsang, and Wang 2003). This so-called non-parametric test (this label comes from the fact that the distribution of the test statistic does not depend on the distribution of the data being tested) performs poorly in small samples, but works well in a simulation environment. Write an `R` function that implements this test where the reference distribution is normal. Using `R` generate 1,000 Cauchy random variables (`rcauchy(1000, location = 0, scale = 1)`) and perform the test (remember, use the same seed, something like `set.seed(123)`, whenever you're generating your own data).

As a hint, you can create the empirical distribution and theoretical CDF using this code:

```r
# create empirical distribution of observed data
ECDF <- ecdf(data)
empiricalCDF <- ECDF(data)
# generate test statistic
D <- max(abs(empiricalCDF - pnorm(data)))


```

MY ANSWERS

QUESTION 1

```r
  ##########################################
############## Question 1 ##############

# libraries
library(dgof) # for performing K-S test
library(tidyverse)
library(stargazer)

# MY TEST FOR ECDF
set.seed(2020)
x <- rnorm(5)
ecdf(x)
plot(ecdf(x))
#################

# Empirical Cumulative Distribution Function EDCF
# This computes the ECDF of a numeric input vector

set.seed(123) # set seed for reproducibility
# set.seed(123) from problem set 1

# rcauchy() is used to compute random cauchy density among a range of inputs
# not very well explained
data <- rcauchy(1000, location = 0, scale = 1) # as given in problem set 1
data <- sort(data)
# applying the ecdf function to calculate the ecdf values of the R data
vector_edcf <- ecdf(data)
vector_edcf

plot(ecdf(data)) # create the ecdf plot
```

```r
# the plot is S shaped

# the Kolmogorov-Smirnov test is used in situations where a comparison has to be made
# between an observed sample distribution and theoretical distribution
ks_data <- max(abs(vector_edcf(data) - pnorm(data)))
print(ks_data)
# gives 0.1347281 and this is the test statistic

# or using R ks.test() function
ks.test(data, "pnorm")
# THIS GIVES
## data:  data
## D = 0.13573, p-value = 2.22e-16
## alternative hypothesis: two-sided

# we can see that the test statistic is 0.13573 and the corresponding
# p-value is 2.2e-16. Since the p-value is less than .05,
# we reject the null hypothesis.
# We have sufficient evidence to say that the sample data does not come
# from a normal distribution.

# I found it difficult to turn the K-S CDF function into R code
# considering that x values are to do with the largest absolute differences
# between the two distribution functions.
# CDF = The cumulative distribution function (CDF) of a random variable evaluated
# at x, is the probability that x will take a value less than or equal to x.


# x = seq(0, 1, by = 0.01)
# plot(x, vector_edcf(x))

# 2 pi


```

# Question 2

Estimate an OLS regression in `R` that uses the Newton-Raphson algorithm (specifically `BFGS`, which is a quasi-Newton method), and show that you get the equivalent results to using `lm`. Use the code below to create your data.

```r
set.seed(123)
# create empirical distribution of observed data
ECDF <- ecdf(data)
```

QUESTION 2

```r
############## Question 2 ##############

set.seed(123) # as given in the problem set 1 - question 2

# ECDF <- ecdf(data)
data <- data.frame(x = runif(200, 1, 10)) # as given in the problem set 1 -
    question 2

data$y <- 0 + 2.75*data$x + rnorm(200, 0, 1.5)

head(data) # prints out the 6 data values using the head() function
# `      x         y
# 1 2.397377  5.099089
# 2 8.612659 22.124880
# 3 2.929424  8.028945
# 4 7.028859 19.131100
# 5 6.559808 14.215458
# 6 1.449998  5.548355

## Estimate the OLS Regression

ols_reg <- lm(data$x ~ data$y, data) # linear model lm()

# or
ols_reg1 <- lm(y ~ x, data = data)
#summary(data)
summary(ols_reg)
summary(ols_reg1)


```