Problem Set 1

Zahra Khan

Due: February 11, 2026

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 23:59 on Wednesday February 11, 2026. No late assignments will be accepted.

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 ith 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:

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

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).

Before condcuting the test it is crucial to outline the Null and Alternative hypothesis.

Null Hypohtesis: The datasets are the same or follow the same, specific distribution. In this case, we are testing if our dataset follows a Normal Distribution.

Althernative Hypothesis: The datasets differ and follow a different distribtuion. In this ase, our dataset does not follow a Normal Distribution.

```r
set.seed(123)
data <- (rcauchy(1000, location = 0, scale = 1))
KSpval <- function(data) {
  data <- data[[1]] #explicitly assuming one column to prevent ecdf error
  k <- length(data) #setting the length of K
  ECDF <- ecdf(data) #creating empirical distribution of observed data
  empiricalCDF <- ECDF(data) #creating empirical distribution
  d_value <- max(abs(empiricalCDF - pnorm(data))) #test statistic
  i <- 1:k #taking a vectorized approach instead of a loop for calculating the terms
  terms <- exp(-((2*i - 1)^2) * pi^2)/(8*(d_value^2)) #calculating the terms for all 100 k's
  d_obs <- sqrt(2 * pi)/d_value * sum(terms) #calculating the observation
return(d_obs)
}

# [1] 0.0006529358
```

After conducting the test, we have recieved a p-value of 0.0006529358, which at a threshold of alpha = 0.05 is a statistically significanty result. This means we have sufficent evidence to reject the null hypothesis that the datasets are the same distribution. Meaning, this result indicates that our dataset does not follow a normal distribution.

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.

```
}

# [1] 0.0006529358
```