Problem Set 1

Applied Stats II

Due: February 12, 2023

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 Sunday February 19, 2023. 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:

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

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).

The code for the function is as follows:

```
kolmogorov_smirnov <- function (dat)  {
  # get number of elements in data
  N <- length(dat)
  # convert data to ECDF
  ECDF <- ecdf(dat)
  empiricalCDF <- ECDF(dat)
  # generate test statistic
  D <- max(abs(empiricalCDF - pnorm(dat)))
  Dmin <- min(empiricalCDF - pnorm(dat))
  Dmax <- max(empiricalCDF - pnorm(dat))

  #calculate critical value
  K <- D * sqrt(N)

  # get probability of calculated test statistic
  k = seq(1,N)

  #calculate q value for P(K)
  qval <- sum(exp(-1*((2*k - 1)^2)*(pi^2)/(8 * K^2)))
  qval <- qval * sqrt(2 * pi) * (1/K)

  # return results
  res <- tibble('Dval'= D, 'Dmax' = Dmax, 'Dmin' = Dmin,
    'Kval'=K, 'pval'=1-qval)
  return(res)
}
```

The test data was generated using the `rcauchy` function, and is summarised in Table 1.

```
  # create empirical distribution of observed data
  set.seed(123)
  N <- 1000
  data <- rcauchy(n=N, location = 0, scale = 1)


```

Table 1: Data Summary Table

Statistic	N	Mean	St. Dev.	Min	Max
Observed_CDF	1,000	0.500	0.289	0.001	1.000
Normal	1,000	0.505	0.358	0.000	1.000

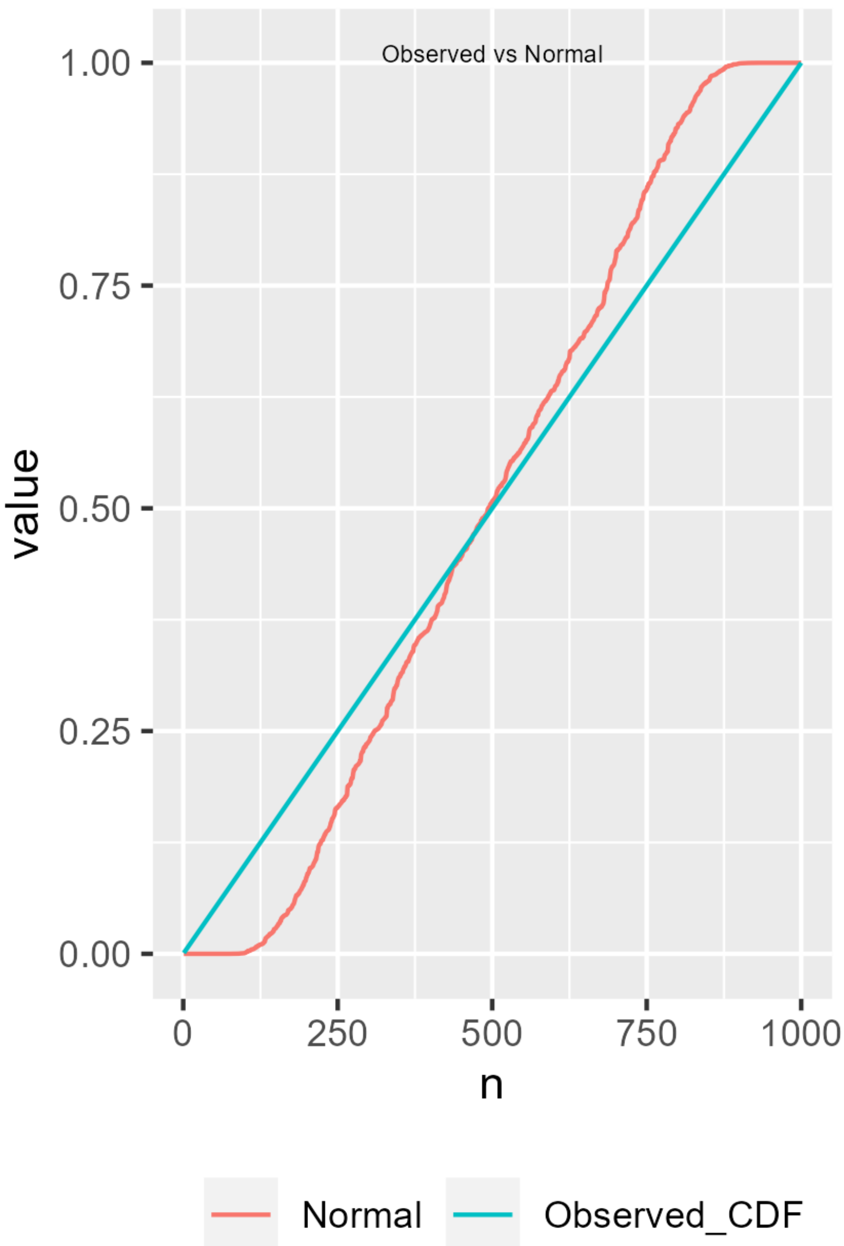

Figure 1: Observed and model (Normal) data

A hypothesis test was carried out using the `kolmogorov_smirnov`, function:
Function call:

```
ks_results <-kolmogorov_smirnov(data)

# A tibble: 1 * 5
  Dval   Dmax    Dmin   Kval       pval
<dbl> <dbl>   <dbl> <dbl>      <dbl>
1 0.135  0.124  -0.135   4.26  2.22e-16

```

Hypothesis Test

1. H_0 the observed data is from the normal distribution $N(\mu = \bar{x}, \sigma = sd)$
2. H_a the observed data is not normally distributed
3. the α value is 0.05
4. the test statistic K is 4.26048
5. the `pvalue` is 2.22×10^{-16} (ie the probability of observing these values if the data was normally distributed is approximately 0)
6. as `pvalue` is less than α we reject the null hypothesis

There is insufficient evidence to support the hypothesis that our observed data is normally distributed.

The results are summarised in Table 2.

The results were checked using the `cont_ks_test` function from the `KSgeneral` package, and the same results were obtained.[1]

```

 KSgeneral::cont_ks_test(data, "pnorm")

     One-sample Kolmogorov-Smirnov test

     data:  data
     D = 0.13573, p-value < 2.2e-16
     alternative hypothesis: two-sided

```

[1]these agreed with the results with the default `ks.test` function.

Table 2: Kolmogorov-Smirnov Test results - one sample

	1
Dval	0.13473
Dmax	0.12356
Dmin	-0.13473
Kval	4.26048
pval	0
k_alpha	0.04301

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.

```
#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{i/n - F_(i); F_(i-1) - (i - 1)/n}
```