MANA130083.01 Nonparametric *Spring 2023*

Solution to Homework 2

Date: March 9, 2023 *Scribe: Jianliang He*

Warning: *This note is only used as a reference solution for the homework, and the solution to each question is not unique. The solution may contain factual and/or typographic errors and comments and criticism are kindly welcomed.*

Problem 1 *Recall that the Kolomogorov-Smirnov One-Sample Statistic* $D_n = \sup_x |F_X(x) - \hat{F}_n(x)|$ *is distribution free (*$\hat{F}_n$ *is the empirical CDF). Suppose* $n = 100$ *and the observed* $D_n = 0.04$*. Would you reject the null hypothesis* $H_0 : X_1, \ldots, X_n \sim F_X$ *at level* $\alpha = 0.05$*. Write a simulation to justify your answer.*

Solution:

```
# D: Kolomogorov-Smirnov One-Sample Statistic
# n: the number of samples
# nrep: the number of samples for simulation
kspvalue<-function(D,n,nrep){
  Dn <- c()
  for (i in 1:nrep) {
    x.obs <- rnorm(n,0,1)
    Fn.hat <- ecdf(x.obs)
    Dn[i] <- max(abs(Fn.hat(x.obs) - pnorm(x.obs)))
  }
  return(sum(Dn > D)/nrep)
}
# Calculate the p-value of the statistics under the problem setting
```

```r
D = 0.04; n = 100; nrep = 5000
kspvalue(D,n,nrep)
```

The result shows that the p-value is $0.983 > 0.05$ and we fail to reject the null hypothesis. $\square$

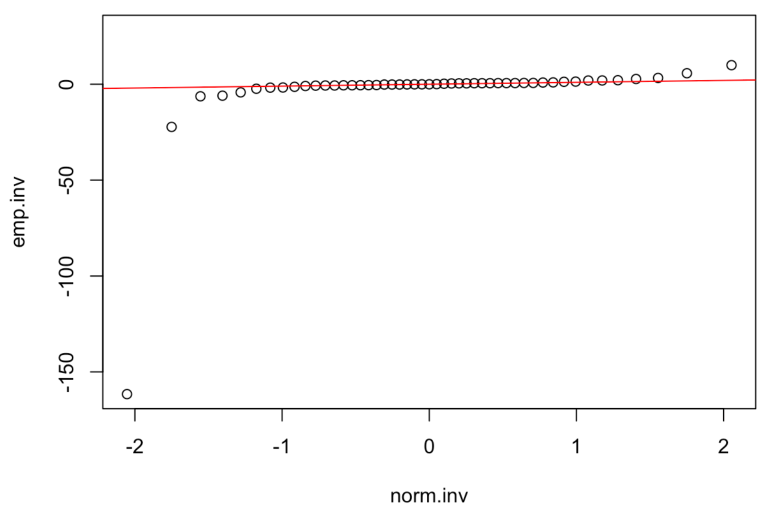
Problem 2 *Use the following r code to generate 50 random numbers from Cauchy distribution.*

```r
set.seed (2)
d <- rcauchy (50)
```

Use a QQ plot to see how good (or bad) your data fits a normal distribution.

Solution:

```r
emp.inv <- sort(d)
sel <- seq(1, 50, by=1)/50
norm.inv <- qnorm(sel)
plot(norm.inv, emp.inv)
abline(0,1,col="red")
```



where we only compare the data to the standard normal distribution. If you want to show the goodness of fit for the normal family, kindly use 'abline(a=mean(d), b=sd(d), col="red")'. $\square$

Problem 3 *Use a test discussed in class to test whether the following sequence is random at level* $\alpha = 0.1$,

$$A, B, B, A, A, A, A, A, A, B, A.$$

Use simulation to justify your answer.

Solution: (Solution is not unique and the following code is based on the number of run)

```
# n: the number of samples
# nrep: the number of samples for simulation
# sample: the original sample
# run: the number of run in the samples
randomtest.run<-function(n,nrep,sam,run){
  runs <- c()
  for (i in 1:nrep) {
    count <-1
    temp <- sample(sam, 11, replace = F)
    for (j in 1:(n-1)){
      count <- count + abs(temp[j]-temp[j+1])
    }
    runs[i] <- count
  }
  return(2*min(sum(runs < run), sum(runs > run))/nrep) # two-sided p-value
}
# Calculate the p-value of the statistics under the problem setting
n=11; nrep = 5000; run=5
sam <- c(1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1)
randomtest.run(n,nrep,samp,run)
```

The corresponding p-value is $0.489 > 0.1$ and we fail to reject the null. $\square$

Remark Please note that randomness implies independency but *does not suggest* sampling with equal probability (i.e. Bernoulli(p) with $p = 0.5$). Thus, in the simulation of Problem 3, you shall sample locally rather than creat new sequence with equal probability.