STA511 Homework #2

Ezgi Karaesmen

10/14/15

1. No materials need to be submitted for this problem.

2. 100 observations from a beta distribution ($\alpha = 3, \beta = 5$, i.e. `shape1` =3 and `shape2` =5) were simulated using the Newton Raphson Method with `R`. The stopping rule was set to $|x_i$ - $x_{i-1}| < 0.05$ and the starting point for the algorithm was 0.5. On average `3.29` iterations were required until the observation was accepted. Frequency and distribution of the accepted observations (i. e. solutions obtained from Newton Raphson Method) are presented in Figure 1.

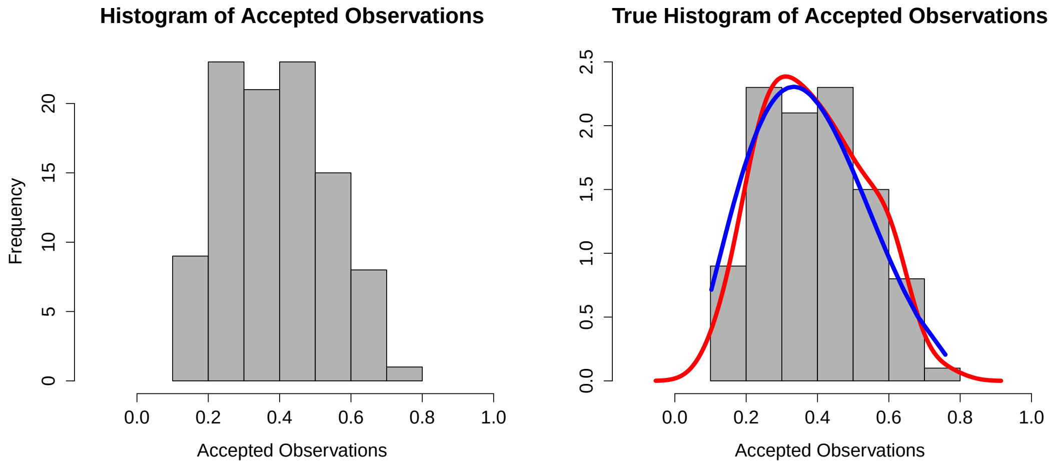

Figure 1: Histogram and true histogram of accepted observations (i.e solutions) are presented. Red curve on the true histogram represents the empirical density of the accepted observations, determined by the `density()` function in `R`. Blue curve represents the theoretical density (i.e probability) function of a beta distribution with $\alpha = 3, \beta = 5$, (i.e. `shape1` =3 and `shape2` =5). It can be see that the both curves overlap well, suggesting that the method was successful in generating random variables from a beta distribution.

`R` code for problem 2 concerning the Newton Raphson method and histogram of the accepted observations are presented below:

```
## HW2 _2. ##

nsims=100  # number of simulations
iter <- NULL
solutions <- NULL
random <- runif(1000)

for(n in 1:nsims){
  x0 <- 0.5 #starting point x-old
  x1 <- NULL # x-new
  i <- 1 # iterations
  N=200 # max number of iterations to compute the solution
  u <- random[n]
  while(i <= N){
    x1 <- x0 - (pbeta(x0, 3, 5) - u)/dbeta(x0, 3, 5)
    i = i+1
    if(abs(x1 - x0) < 0.05) {break}
    x0 = x1
  }
  solutions[n] <- x1  # Solutions from 100 simulations
  iter[n] <- i        # Number of required iterations to obtain solution
}
mean(iter) # Average number of iterations

## Figure 1 ##
library(MASS)
par(mfrow=c(1,2))
hist(solutions, col="gray70", xlim = c(-0.13, 1),
     xlab="Accepted Observations", main="Histogram of Accepted Observations")
truehist(solutions, col="gray70", ymax = 2.5, xlim = c(-0.13, 1),
         xlab="Accepted Observations", main="True Histogram of Accepted Observations")
## Predicted density function of the observations
lines(density(solutions)$x,density(solutions)$y, col="red", lwd=5)
beta.pdf <- data.frame(solutions, dbeta(solutions, 3, 5))
colnames(beta.pdf) <- c("x", "y")
beta.pdf<-beta.pdf[order(beta.pdf$x),]
# Theoretical PDF of the observations
lines(beta.pdf$x, beta.pdf$y, col="blue", lwd=5)
```

3. The Chi-Squared test and Kolmogorov-Smirnov test (KS test) are compared for accuracy in testing random number sequences according to Type I error (3.a) and power (3.b).

    (a) Random uniform numbers between 0 and 1 were generated using the `runif()` command in `R` with sample size of `n = 10, 25, 50, 100, 200, 500, 1000, 2000, 5000, 10000`. At each sample size, simulation was repeated `500` times and the number of *p-values* less than 0.05 for both the Chi-squared test (with 10 bins) and the Kolmogorov-Smirnov test were counted upon `500` iterations for different sample sizes. A table summarizing count results with sample size as the rows and tests as the columns was produced (Table 1).

Table 1: Frequencies of *p-values* less than 0.05 that are obtained from Kolmogorov-Smirnov and Chi-squared tests for different sample sizes. Number of iterations was set to `500`.

	Kolmogorov-Smirnov Test	Chi-Squared Test
n=10	0.04	0.02
n=25	0.05	0.03
n=50	0.02	0.04
n=100	0.04	0.04
n=200	0.04	0.04
n=500	0.03	0.05
n=1000	0.04	0.05
n=2000	0.07	0.06
n=5000	0.06	0.08
n=10000	0.06	0.03

`R` code for problem 3 (a) is presented below:

```
## HW2 _ 3 (a) ##

sims=500 #number of simulations
size <- c(10, 25, 50, 100, 200, 500, 1000, 2000, 5000, 10000) # Sample sizes
results <- data.frame(matrix(c(1:2*length(size)), nrow = length(size), ncol = 2)) #Empty matrix
colnames(results) <- c("Kolmogorov-Smirnov Test", "Chi-Squared Test")
rownames(results) <- paste("n", size, sep="=")

for(n in 1:length(size)){
  p_KS <- NULL
  p_CS <- NULL
  for(i in 1:sims){
    x <- runif(size[n])
    p_KS[i] <- ks.test(x, "punif", 0, 1)$p.value
    xcount <- hist(x, breaks=10, plot=F)$counts
    p_CS[i] <- chisq.test(xcount)$p.value
  }
  KS <- sum((p_KS <= 0.05)*1)
  CS <- sum((p_CS <= 0.05)*1)
  results[n, 1:2] <- cbind(KS, CS)
}
results <- results/500 # frequencies
library(xtable)
xtable(results) # print results in LaTeX table format
```

(b) Non-random numbers were generated using `rbeta()` with $\alpha = 4, \beta = 6$ (i.e. `shape1` =3 and `shape2` =5) with sample size of `n = 10, 25, 50, 100, 200, 500, 1000, 2000, 5000, 10000`. At each sample size, simulation was repeated `500` times and the number of $p - values$ less than `0.05` for both the Chi-squared and KS test were counted. A table summarizing the count results with sample size as the rows and tests as the columns was produced (Table 2).

Table 2: Frequencies of *p-values* less than 0.05 that are obtained from Kolmogorov-Smirnov and Chi-squared tests for different sample sizes. Number of iterations was set to `500`.

	Kolmogorov-Smirnov Test	Chi-Squared Test
n=10	1.00	0.02
n=25	1.00	0.12
n=50	1.00	0.56
n=100	1.00	0.99
n=200	1.00	1.00
n=500	1.00	1.00
n=1000	1.00	1.00
n=2000	1.00	1.00
n=5000	1.00	1.00
n=10000	1.00	1.00

R code for problem 3 (b) is presented below:

```
## HW2 _ 3 (b) ##

sims=500 # Number of simulations
size <- c(10, 25, 50, 100, 200, 500, 1000, 2000, 5000, 10000)
results <- data.frame(matrix(c(1:2*length(size)), nrow = length(size), ncol = 2))
colnames(results) <- c("Kolmogorov-Smirnov Test", "Chi-Squared Test")
rownames(results) <- paste("n", size, sep="=")

for(n in 1:length(size)){
  p_KS <- NULL
  p_CS <- NULL
  for(i in 1:sims){
    x <- round(rbeta(size[n], 4, 6), 3)
    p_KS[i] <- ks.test(x, "pbeta", 0, 1)$p.value
    xcount <- hist(x, breaks=10, plot=F)$counts
    p_CS[i] <- chisq.test(xcount)$p.value
  }
  KS <- sum((p_KS <= 0.05)*1)
  CS <- sum((p_CS <= 0.05)*1)
  results[n, 1:2] <- cbind(KS, CS)
}

results <- results/sims
library(xtable)
xtable(results)
```

4. A simulation to compare the Type I error for a Chi-squared test for a random number generator on 0 to 1 was performed. Two variables were studied for the simulations: sample size and number of bins. A table summarizing the count results with sample size as the columns and bin numbers as the rows was produced (Table 3). Additionally, number of $p - values$ lower than 0.05 for each size with different number of bins was visualized as barplots (Figure 2). It can be seen that the number of bins affect the number of p-values for smaller sample sizes at 10 and 25. However $p - value$ frequency is stabilized around 0.05 for higher sample sizes (Table 3, Figure 2). For sample sizes lower than 50, high number of bins increases the error rate of the Chi-squared test. Additionally, smaller number of bins for higher sample sizes show a higher variability in the $p - value$ frequencies compared to higher number of bins.

Table 3: Frequencies of *p-values* less than 0.05 that are obtained from Chi-squared test for different bin and sample sizes. Number of iterations was set to 500.

Bins	Sample Size								
	10	25	50	100	200	500	1000	2000	5000
5	0.03	0.06	0.05	0.06	0.04	0.04	0.05	0.06	0.05
10	0.01	0.05	0.04	0.05	0.07	0.05	0.04	0.05	0.05
15	0.03	0.03	0.04	0.06	0.04	0.07	0.03	0.05	0.07
20	0.02	0.04	0.05	0.07	0.05	0.07	0.04	0.04	0.03
25	0.03	0.04	0.03	0.05	0.04	0.06	0.06	0.05	0.06
30	0.02	0.03	0.04	0.05	0.05	0.06	0.04	0.04	0.06
35	0.03	0.03	0.04	0.04	0.03	0.03	0.04	0.06	0.05
40	0.05	0.04	0.04	0.04	0.06	0.05	0.06	0.05	0.05
45	0.03	0.02	0.03	0.03	0.05	0.05	0.06	0.04	0.05
50	0.05	0.04	0.05	0.04	0.05	0.04	0.06	0.03	0.05
100	0.04	0.05	0.03	0.06	0.04	0.06	0.04	0.05	0.05
200	0.01	0.06	0.04	0.04	0.04	0.06	0.04	0.04	0.05
300	0.05	0.09	0.06	0.03	0.05	0.04	0.05	0.05	0.05
400	0.09	0.06	0.04	0.05	0.04	0.04	0.05	0.04	0.04
500	0.08	0.07	0.05	0.04	0.05	0.04	0.03	0.05	0.06

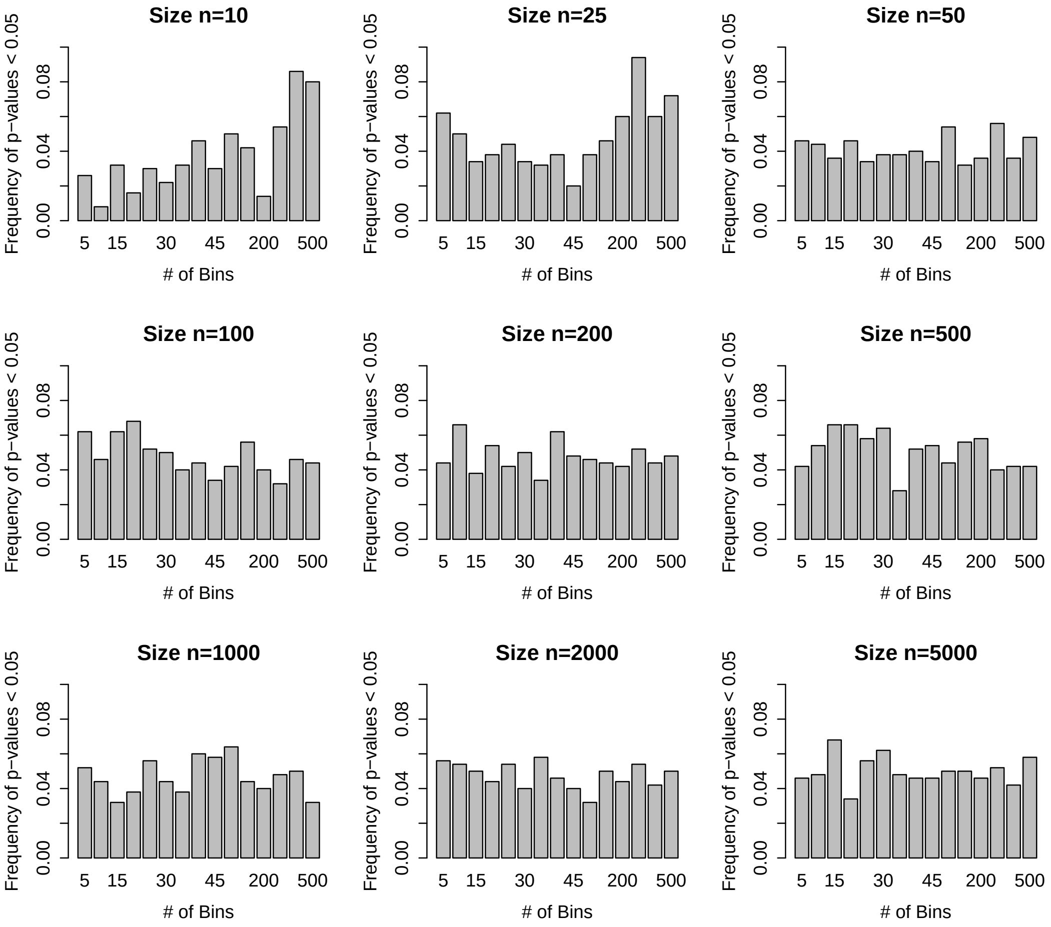

Figure 2: Barlpots of number of $p - values < 0.05$ for different sizes and different number of bins.

R code for problem 4 is presented below:

```
## HW2 _ 4 ##
set.seed(333)
sims=500
size <- c(10, 25, 50, 100, 200, 500, 1000, 2000, 5000)
bins <- c(seq(5, 50, by=5), seq(100, 500, by=100))
results <- data.frame(matrix(c(length(bins)*length(size)), # Empty data frame
            nrow = length(bins), ncol = length(size)))
colnames(results) <- as.character(size)
rownames(results) <- as.character(bins)

for(n in 1:length(size)){

  for(i in 1:length(bins)){

    p_CS <- NULL

    for(j in 1:sims){
      x <- round(runif(size[n]), 3)
      xcount <- hist(x, breaks=bins[i], plot=F)$counts
      p_CS[j] <- chisq.test(xcount)$p.value
    }
    CS <- sum((p_CS <= 0.05)*1)
    results[i,n] <- CS
  }
}
results <- results/sims #Frequencies
par(mfrow=c(3,3))
for(c in 1:length(size)){  #plotting multiple barplots
barplot(results[,c], names.arg = as.character(bins), main = paste("Size n", size[c], sep="="),
        xlab="# of Bins", ylab="Frequency of p-values < 0.05", ylim = c(0,0.1))
}

results <- results/sims
library(xtable)
xtable(results)
```