ECON 580 **Shabab Ahmed**
Due Date: Friday, October 16 at 8AM

Homework 2

PROBLEM 1:

A gambler plays the following game: a fair coin is tossed repeatedly. On the k-th toss, if the coin shows tails, she receives 0, if it shows heads she receives $2/(3^k)$, Let X_k be the gambler's total winnings after k stages.

(a) Find the pmf and cdf for X_1.

(b) Find the pmf and cdf for X_2.

(c) Using simulation in `R`, or otherwise, find either approximately or exactly the pmf and cdf for X_5. Express your answers as plots.

Hint: you may find the `R` *function* `plot.ecdf` *useful.*

SOLUTION:
a) After the 1st toss, the only two possible outcomes are: heads or tails.

If the toss results in a heads, she receives $\frac{2}{3}$.

If the toss results in tails, then she receives 0.

Therefore, $X_1 = \{0, \frac{2}{3}\}$. Also, $P(X_1 = 0) =$ probability of tails $= \frac{1}{2}$.Similarly, $P(X_1 = \frac{2}{3}) =$ probability of heads $= \frac{1}{2}$.

Thus, the pmf for X_1 is:

$$f(x) = \begin{cases} \frac{1}{2} & \text{for } x = 0 \\ \frac{1}{2} & \text{for } x = \frac{2}{3} \\ 0 & \text{otherwise} \end{cases}$$

$$\implies f(x) = \begin{cases} \frac{1}{2} & \text{for } x \in \{0, \frac{2}{3}\} \\ 0 & \text{otherwise} \end{cases}$$

Now, we move onto the CDF. Recall, that for a discrete random variable the cdf is discontinuous from the left:

$$F(x) = \sum_{k \leq x} f(k).$$

Therefore, $F(x) = 0$ for $x < 0$. For $0 \leq x < \frac{2}{3}$, $F(x) = \frac{1}{2}$ since $f(0) = \frac{1}{2}$. Now, let $x = \frac{2}{3}$. Then, $F(x) = f(\frac{2}{3}) + f(0) = \frac{1}{2} + \frac{1}{2} = 1$. In fact, using the above formula for discrete random variable and our pmf we have that $F(x) = 1$ for $x > \frac{2}{3}$. Combining all of these gives us the following CDF for X_1:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{2} & \text{for } 0 \leq x < \frac{2}{3} \\ 1 & \text{for } x \geq \frac{2}{3} \end{cases}.$$

---

b) Now, we are consider the case of two tosses. We will denote a toss of heads as H and a toss of tails as T. The possible outcomes are following:

$$(H,H), (T,H), (H,T), (H,H)$$

Notice:

$$P((H,H)) = P((T,H)) = P((H,T)) = P((T,T)) = \frac{1}{4}.$$

Now, we will calculate X_2 for the given outcomes.

(H,H) :

$$X_2 = \frac{2}{3} + \frac{2}{9} = \frac{8}{9}$$

(H,T) :

$$X_2 = \frac{2}{3} + 0 = \frac{2}{3}$$

(T,H) :

$$X_2 = 0 + \frac{2}{9} = \frac{2}{9}$$

(T, T) :

$$X_2 = 0 + 0 = 0$$

Thus, $X_2 = \{0, \frac{2}{9}, \frac{2}{3}, \frac{8}{9}\}$.

Given that,

$$P((H, H)) = P((T, H)) = P((H, T)) = P((T, T)) = \frac{1}{4}.$$

we have that

$$P(X_2 = 0) = P(X_2 = \frac{2}{9}) = P(X_2 = \frac{2}{3}) = P(X_2 = \frac{8}{9}) = \frac{1}{4}.$$

Therefore, the pmf of X_2 is:

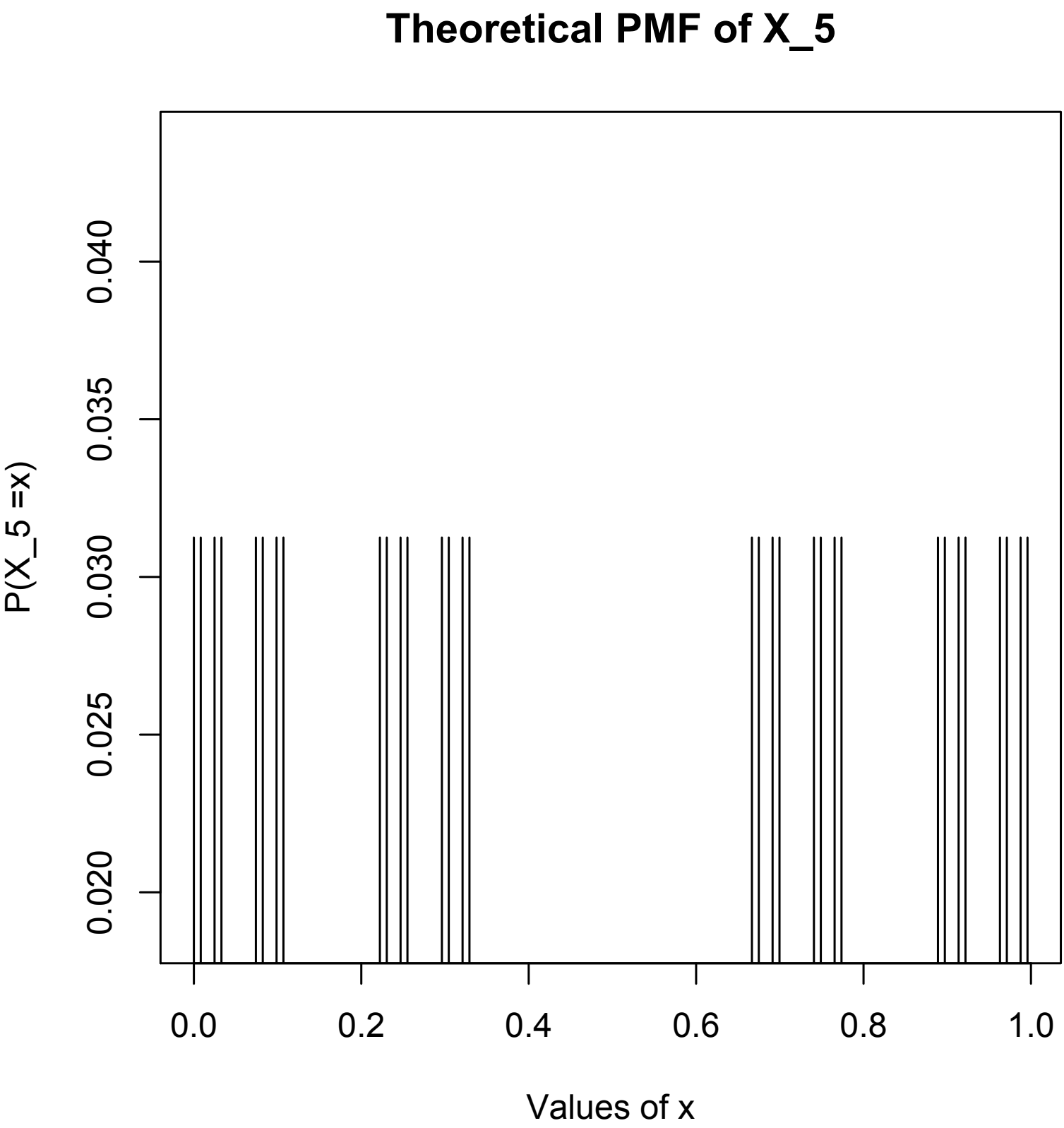
$$f(x) = \begin{cases} \frac{1}{4} & \text{for } x \in \{0, \frac{2}{9}, \frac{2}{3}, \frac{8}{9}\} \\ 0 & \text{otherwise} \end{cases}$$

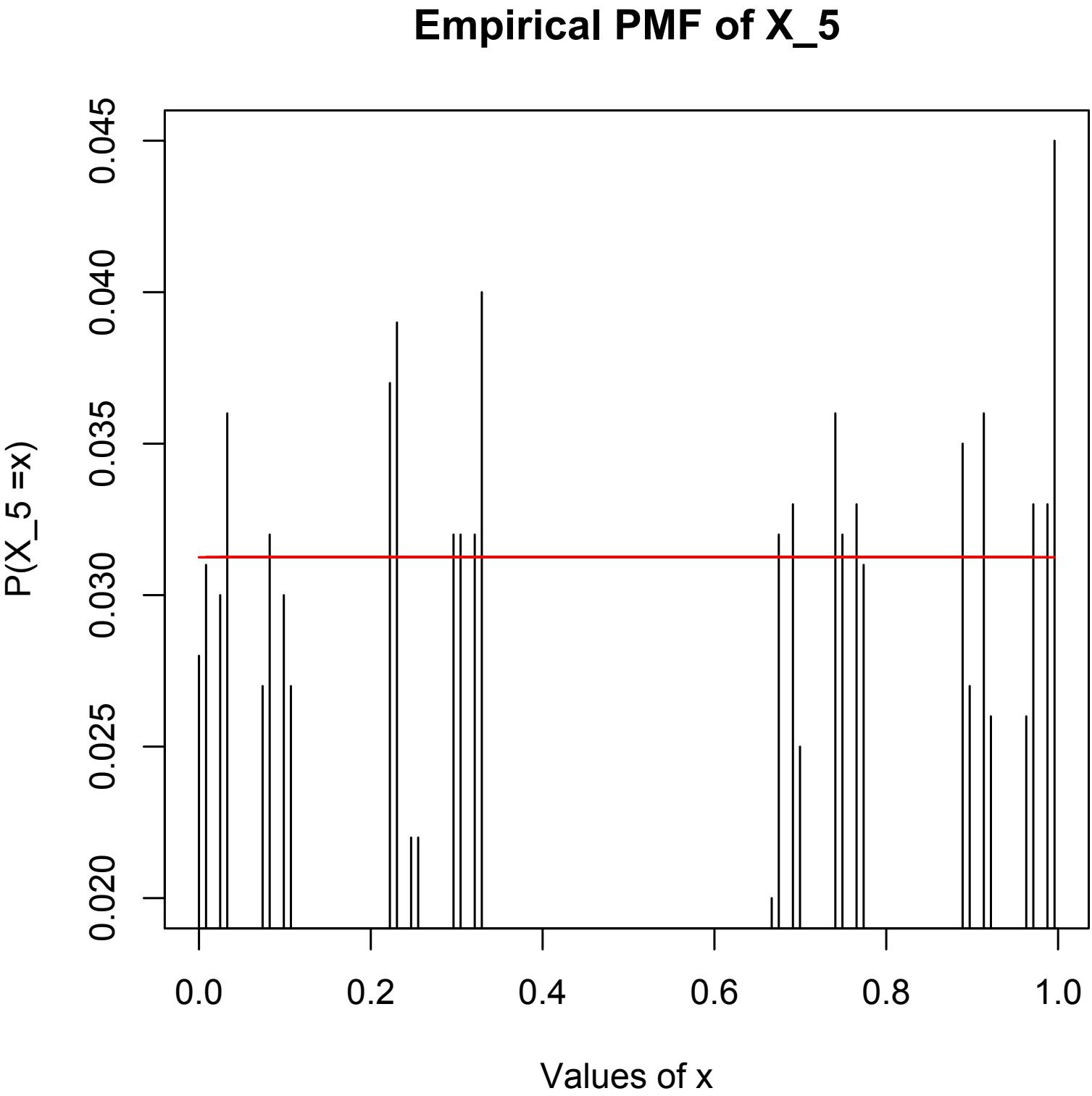
Now, we move onto finding the CDF.

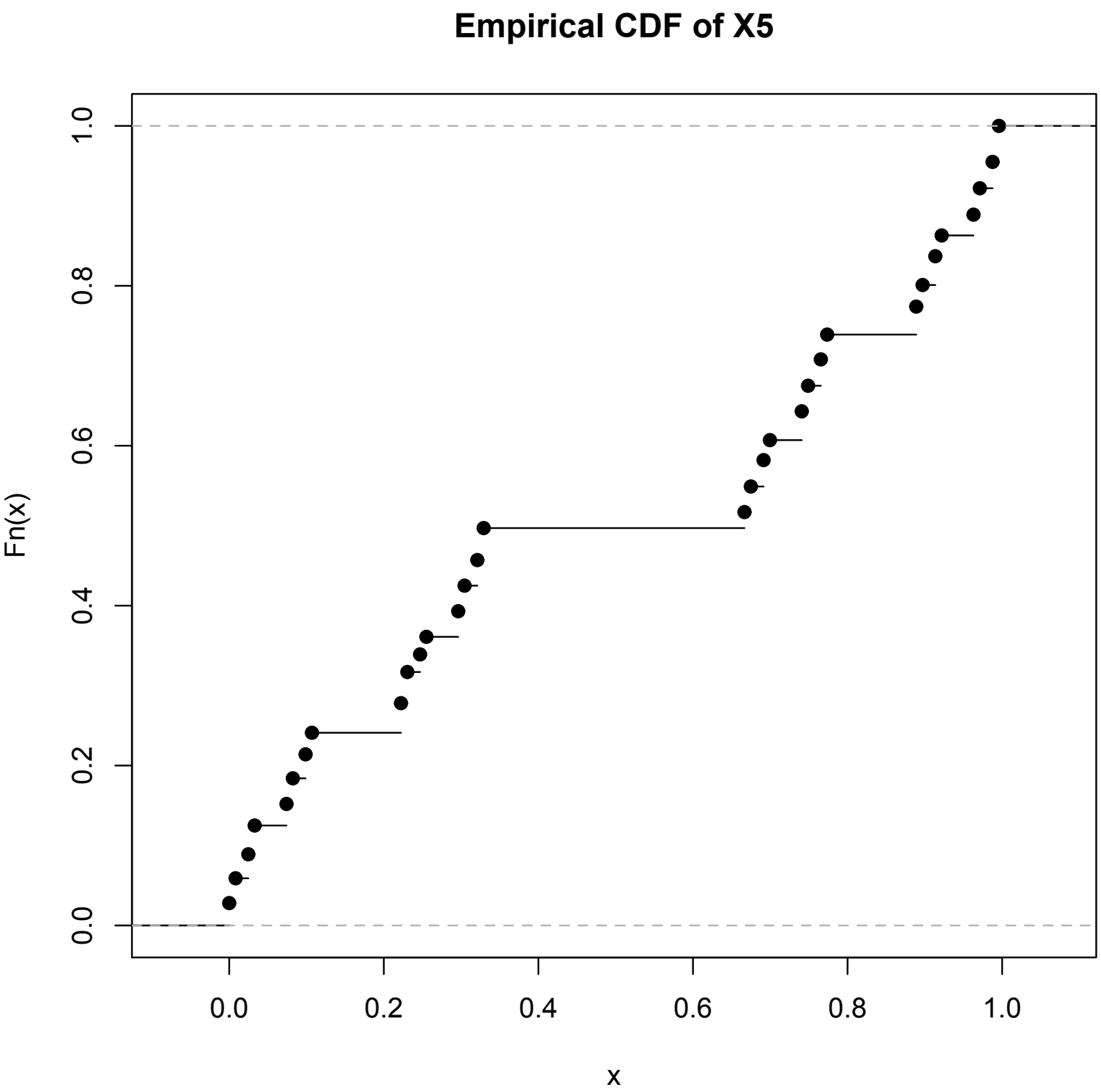
$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{4} & \text{for } 0 \leq x < \frac{2}{9} \\ \frac{1}{2} & \text{for } \frac{2}{9} \leq x < \frac{2}{3} \\ \frac{3}{4} & \text{for } \frac{2}{3} \leq x < \frac{8}{9} \\ 1 & \text{for } x \geq \frac{8}{9} \end{cases}.$$

---

c) We used simulations in `R` to approximately find the CDF for X_5. For X_5, we are tossing 5 coins and so there are $2^5 = 32$ possibilities and each has a probability of $\frac{1}{32}$. We will use this for the PMF and we will extract the possible values for X_5 by extracting the unique values from the simulation with each having probability $\frac{1}{32}$. This unique values represents the different possibilities of the 5 coin tosses and the corresponding total winnings. For example, $X_5 = 0$ will correspond (T, T, T, T, T) and $P(X_5 = 0) = \frac{1}{32}$. We include both the theoretical and the empirical pmf in this case. The red line in the empirical pmf indicates the probability according to theory.


Theoretical PMF of X_5
P(X_5 =x)
0.020
0.025
0.030
0.035
0.040
0.0
0.2
0.4
0.6
0.8
1.0
Values of x


Empirical PMF of X_5
P(X_5 =x)
0.020
0.025
0.030
0.035
0.040
0.045
0.0
0.2
0.4
0.6
0.8
1.0
Values of x


Empirical CDF of X5
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
x
0.0
0.2
0.4
0.6
0.8
1.0

R Code:

```
set.seed(5)
nsims <- 1000  # number of simulations
seq <- 1:nsims
n <-5 # number of tosses
X5 <-c() # initializing list of total winnings
for (i in seq){
x5 = 0
fair <- sample(1:2, n, replace=TRUE)  # 5 tosses of the coin
seq1 <- 1:length(fair)
for (j in seq1){
if (fair[j] == 1) x5 <- x5 + (2/3^j)  # 1 = heads and sum the winnings
}
X5[i] <- x5
}


plot.ecdf(X5, main = "Empirical CDF of X5")


x_vals<- unique(X5)

# extracting unique values from X_5; this represents all the possible values X_5
can take

# there are 32 different event and fair coin means
# probability of each event is 1/32, that is, the probability that X_5
#takes on one of the values is 1/32. We have extracted the values from
#X_5 above

prob <- rep(1/length(x_vals), length(x_vals))

plot(x_vals, prob, xlab = "Values of x", ylab = "P(X_5 =x)", main = "Theoretical PMF
of X_5", type = 'h')

m <- 1:length(x_vals)
probs <- c()

for (i in m){
```

```
bool <- X5 == x_vals[i] # creates 1 if true 0 if false
p = mean(bool)  # mean will give the proportion
probs[i] = p
}

plot(x_vals, probs, xlab = "Values of x", ylab = "P(X_5 =x)", main = "Empirical
PMF of X_5", type = 'h')
points(x_vals, prob, col = "red", type = 'l')
```

PROBLEM 2:

(a) Simulate samples of size 500 from an Exponential random variable with parameter $\lambda = 2$ in `R` using `rexp`. Construct an empirical CDF from the data using the `plot.ecdf` command in `R`.

Recall that the empirical CDF instead of plotting $P(X \leq x)$ for a random variable X, it plots the proportion of observations in the data that are $\leq x$.

(b) Use the `rexp` or `qexp` functions to add the CDF for an exponentially distributed variable with parameter $\lambda = 2$ to your plot.

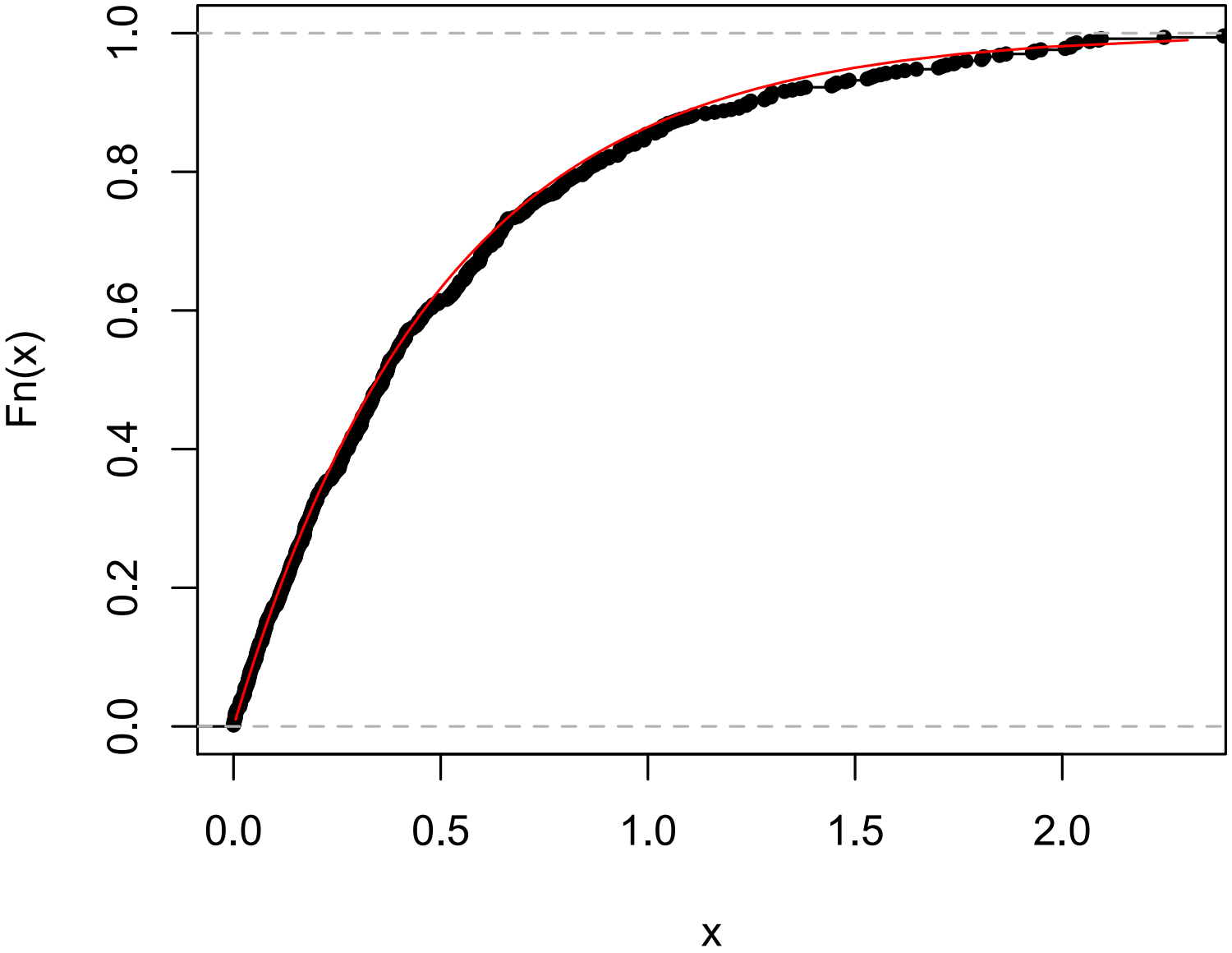
Your solution should contain the few lines of `R` code that you used, together with a single plot containing the empirical CDF and the Exponential CDF superimposed.

SOLUTION:

We have attached the single plot containing the empirical CDF and the Exponential CDF superimposed. The red line is the Exponential CDF while the black line is the empirical CDF.

`R Code:`

```
set.seed(2)
n <- 500  # sample size

# Exponential random variable with parameter lambda = 2
lambda <-2
data <- rexp(n, lambda)

# Theoretical CDF

probs <- seq(from=0.01,to=0.99,by=0.01)
xvals <- qexp(probs, lambda)

plot.ecdf(data,xlim=c(min(xvals),max(xvals)), main ="CDF for Exponential Distribution
with lambda = 2", pch=20)
points(xvals,probs,col="red",type="l")
```


CDF for Exponential Distribution with lambda = 2
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
x

PROBLEM 3:

Suppose that we have a dataset consisting of n numbers. Let the numbers, in order from smallest to largest, be $x_{(1)}, \ldots, x_{(n)}$ and assume that the numbers lie between 0 and 1 and there are no ties so that $0 < x_{(1)} < x_{(2)} < \cdots < x_{(n)} < 1$.

(a) Let $\mathbb{F}(t)$ be the empirical CDF for this dataset. Express $\mathbb{F}(t)$ in terms of $x_{(1)}, \ldots, x_{(n)}$.

Hint: count the number of observations that are $\leq t$.

(b) If a random variable X has the Rectangular distribution on $[0, 1]$, find the CDF $F(t)$ for X. *Hint: integrate the pdf for X.*

Suppose somebody gives you the following set of $n = 10$ observations:

$$0.03, 0.11, 0.42, 0.44, 0.47, 0.66, 0.75, 0.88, 0.89, 0.90$$

This person claims that this is a set of independent samples from a Rectangular $[0, 1]$ distribution (that they have ordered). If you were skeptical about this claim, it would be natural to compare the empirical CDF $\mathbb{F}(t)$ and the (theoretical) CDF $F(t)$.

(c) Plot the empirical CDF $\mathbb{F}(t)$ and the (theoretical) CDF $F(t)$ on a single plot. *Hint: your plot can be contained inside the unit square.*

(d) One way to measure the difference between these CDFs is to look for the largest difference between them:

$$K = \sup_t |\mathbb{F}(t) - F(t)|.$$

Note: sup is used here because $\mathbb{F}(t)$ is discontinuous.

Find the value of K for the dataset given above. Also report the value of t where the supremum is attained.

(Aside) K as defined here is the Kolmogorov-Smirnov statistic for testing whether a set of observations are a random sample from a Rectangular $[0, 1]$ distribution.

SOLUTION:

a)

Let $\mathbb{F}(t)$ be the empirical CDF for the dataset. Since, $0 < x_{(1)} < x_{(2)} < \cdots < x_{(n)} < 1$ we know that there are no observations that are less than $x_{(1)}$. The CDF is given as the following:

$$\mathbb{F}(t) = \begin{cases} 0 & \text{for } x < x_{(1)} \\ \frac{1}{n} & \text{for } x_{(1)} \leq t < x_{(2)} \\ \frac{2}{n} & \text{for } x_{(2)} \leq t < x_{(3)} \\ \frac{3}{n} & \text{for } x_{(3)} \leq t < x_{(4)} \\ \vdots & \\ \frac{n-1}{n} & \text{for } x_{(n-1)} \leq t < x_{(n)} \\ 1 & \text{for } t \geq x_{(n)} \end{cases}$$

This can be concisely written as:

$$\mathbb{F}(t) = \begin{cases} 0 & \text{for } x < x_{(1)} \\ \frac{k}{n} & \text{for } x_{(k)} \leq t < x_{(k+1)};\ k = 1, 2, 3, \cdots, n-1\ . \\ 1 & \text{for } t \geq x_{(n)} \end{cases}$$

---

b) The pdf for the Rectangular distribution on $[0, 1]$ is given by:

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

We will integrate it to find the CDF $F(t)$ for X.

$$F(t) = \int_{-\infty}^{t} f(x)\, dx$$

If $t < 0$:

$$F(t) = \int_{-\infty}^{t} f(x)\, dx = \int_{-\infty}^{t} 0\, dx = 0$$

If $0 \leq t \leq 1$:

$$F(t) = \int_{-\infty}^{t} f(x)\, dx = \int_{-\infty}^{0} f(x)\, dx + \int_{0}^{t} f(x)\, dx = 0 = \int_{0}^{t} 1\, dx = x\Big|_{0}^{t} = t$$

If $t > 1$:

$$F(t) = \int_{-\infty}^{t} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{t} f(x)\,dx$$

$$\implies F(t) = \int_{0}^{1} f(x)\,dx = \int_{0}^{1} 1\,dx = x\big|_{0}^{1} = 1.$$

Therefore,

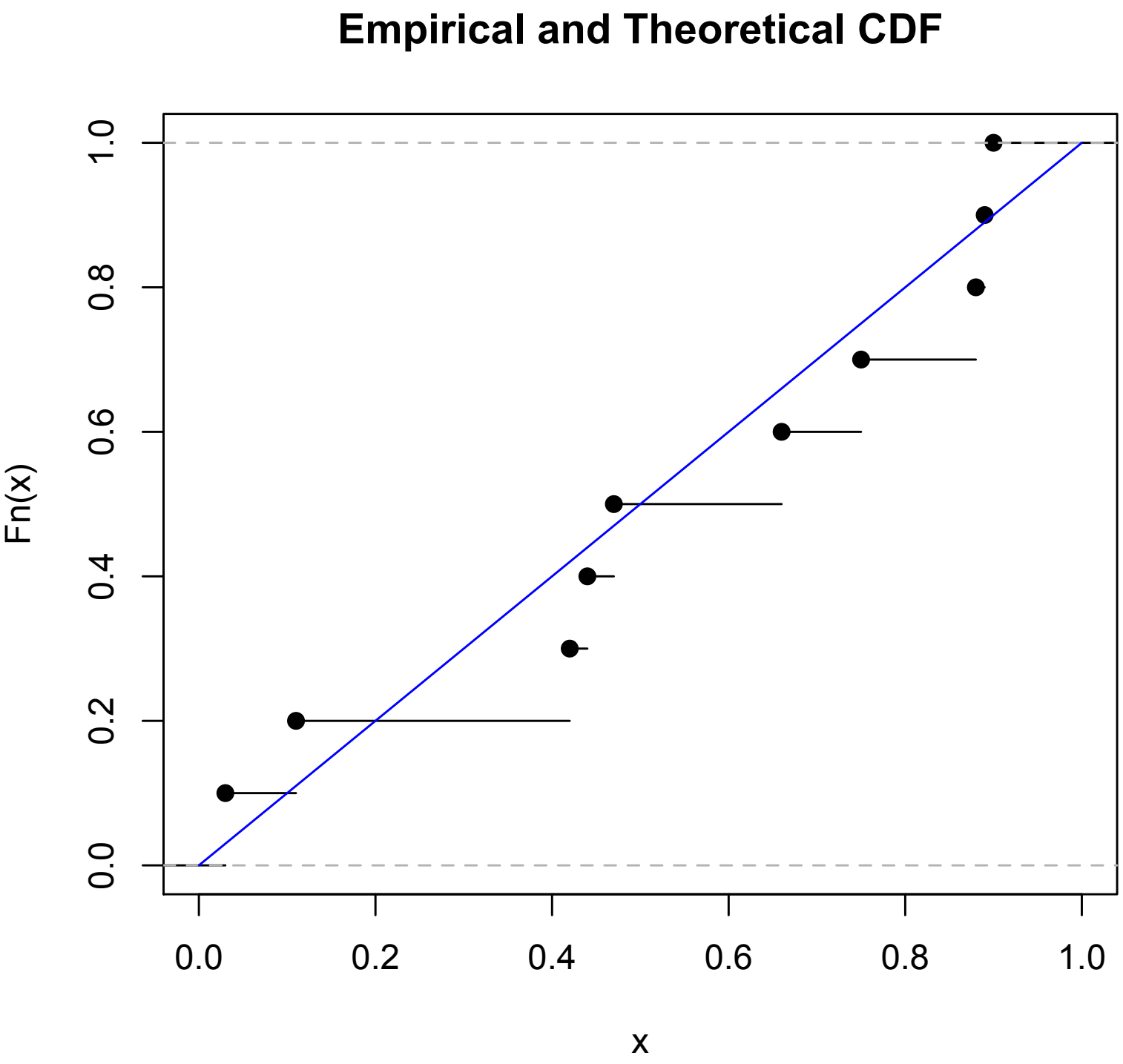
$$F(t) = \begin{cases} 0 & t < 0 \\ t & 0 \le t \le 1 \\ 1 & t > 1 \end{cases}.$$

---

c) The following is the plot containing the empirical CDF $\mathbb{F}(t)$ and the theoretical CDF $F(t)$. The blue line represents the theoretical CDF $F(t)$ which we had calculated above. It is just the $F(t) = t$ we had calculated above on $[0, 1]$. We could have used the `qunif` function here but I just used the computed CDF from part b). The black dots and lines represents the empirical CDF $\mathbb{F}(t)$.

`R Code:`

```
val <- c(0.03,0.11,0.42, 0.44,0.47,0.66,0.75, 0.88, 0.89, 0.90) # values

# theoretical cdf F(t) =t on the interval [0,1]
x <- seq(0,1,0.00001)
plot.ecdf(val, xlim= c(0,1),main = "Empirical and Theoretical CDF")
points(x, x, type = 'l', col = 'blue')
```


Empirical and Theoretical CDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
x

d)

$$K = \sup_t |\mathbb{F}(t) - F(t)| .$$

The value of K in our dataset is 0.12 and the t that achieves it is 0.42. We used the following code for this:

```
# Just for this dataset

F <- ecdf(val)
diff <- abs(F(val)- val)
K <- max(diff)

# finding the corresponding t :

# finding the index for which the max difference occurs

t_index <- which(diff== K)

t <- val[t_index]
print(K)
print(t)
```

We could also consider K for all $t \in [0,1]$. In that case, we find that $K = 0.21999$ and $t = 0.41999$. We did this by discretizing the interval $[0,1]$ using a fine scale and then calculating the difference for those discretized points on the interval. Here is the code for this part:

```
# For all t in [0,1]

F <- ecdf(val)
diff <- abs(F(x)- x)
K <- max(diff)

# finding the corresponding t :

# finding the index for which the max difference occurs

t_index <- which(diff== K)

t <- x[t_index]
```

```
print(K)
print(t)
```

PROBLEM 4:

Goldberger question 2.12.

For each of the following, use the cdf approach to obtain the pdf of Y:

(a) X distributed exponential, $Y = 2X$.

(b) X distributed rectangular on $(0, 1)$, $Y = -\log(X)$.

(c) X distributed standard normal, $Y = X^2$.

SOLUTION:

a) Suppose X distributed exponential and $Y = 2X$. We will first derive the CDF of Y:

$$F_Y(y) = P(Y \leq y) = P(2X \leq y) = P(X \leq \frac{y}{2}) = F_X(\frac{y}{2})$$

Then,

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(\frac{y}{2}) = \frac{1}{2}f_X(\frac{y}{2})$$

The pdf of exponential distribution is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda > 0$ is a constant.

Thus, the pdf, $f_Y(y)$, of Y is the following:

$$f_Y(y) = \frac{1}{2}\begin{cases} \lambda e^{-\lambda \frac{y}{2}} & \text{for } \frac{y}{2} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\implies f_Y(y) = \begin{cases} \frac{1}{2}\lambda e^{-\lambda \frac{y}{2}} & \text{for } y > 0 \\ 0 & \text{otherwise} \end{cases}.$$

b) X distributed rectangular on $(0,1)$, $Y = -\log(X)$. We will start by calculating the CDF of Y.

$$F_Y(y) = P(Y \leq y) = P(-\log(X) \leq y) = P(\log(X) \geq -y) = P(X \geq e^{-y}) = 1 - P(X < e^{-y})$$

$$= 1 - P(X \leq e^{-y}) \text{ since } X \text{ is a continuous random variable and } P(X = e^{-y}) = 0$$

$$= 1 - F_X(e^{-y})$$

Then, we differentiate to find the pdf of Y:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}(1 - F_X(e^{-y})) = -e^{-y} \times (-f_X(e^{-y}))$$

$$= e^{-y}f_X(e^{-y}).$$

The pdf for rectangular distribution on $(0,1)$ is given by:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the pdf for Y is:

$$f_Y(y) = e^{-y}f_X(e^{-y}) = e^{-y}\begin{cases} 1 & \text{for } 0 < e^{-y} < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} e^{-y} & \text{for } \log(0) < -y < \log(1) \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} e^{-y} & \text{for } -\log(1) < y < -\log(0) \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} e^{-y} & \text{for } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

---

c)X distributed standard normal, $Y = X^2$.

We will start by finding the CDF for Y:

If $y < 0$, then:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = 0$$

If $y \geq 0$ then:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = P(X \leq \sqrt{y}) - P(X < -\sqrt{y})$$

$$= F_X(\sqrt{y}) - P(X \leq -\sqrt{y}) \text{ (since } X \text{ is a continuous random variable } P(X = -\sqrt{y}) = 0)$$

$$\implies F_Y(y) = F_X(\sqrt{y}) - F_X(\sqrt{-y})$$

Therefore, the pdf for Y is (for $y \geq 0$) :

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{2\sqrt{y}} [f_X(\sqrt{y}) + f_X(-\sqrt{y})]$$

The pdf for standard normal distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Plugging this into $f_Y(y)$ (for $y \geq 0$):

$$f_Y(y) = \frac{1}{2\sqrt{y}}[\frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}} + \frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}] = \frac{e^{-\frac{y}{2}}}{\sqrt{2\pi y}}$$

So, the pdf for Y is the following:

$$f_Y(y) = \begin{cases} 0 & \text{for } y < 0 \\ \frac{e^{-\frac{y}{2}}}{\sqrt{2\pi y}} & \text{for } y \geq 0 \end{cases}$$

PROBLEM 5:

In this exercise the goal is to simulate data from a given distribution in order to verify the formulae given in Goldberger Table 3.1

(a) Show via simulation that if X has a Poisson distribution with $\lambda = 2$ then $E(X) = V(X) = 2$. *Hint: Use the* `rpois` *function in* `R`.

By simulating a few different sized datasets show that as the number of simulations increase, the mean and variance for your simulated data both become closer to 2.

(b) Show via simulation that if X has a Binomial distribution with $n = 400, p = 0.3$ then $E(X) = np = 120$ and $V(X) = np(1-p) = 84$. *Hint: Use the* `rbinom` *function in* `R`.

Again by simulating different sized datasets show that as the number of simulations increase, the mean and variance for your simulated data approach their respective population values.

SOLUTION:

a) We aim to show via simulation that if X has a Poission distribution with $\lambda = 2$, then $E(X) = V(X) = 2$. We use sample size from 10 to 10000 with increments of 20 and we did find that with increasing sample size the mean and variance of the simulated data goes closer to 2. This can be seen in the figures below: as we increase the number of simulations, we get the mean and the variance gets closer to 2. The points seem to converge to the blue line represented by the theoretical answers as we increase the number of samples. The `R` `code is provided below:`

`R Code:`

```
set.seed(42)
simulations <-  seq(10,10000, 20) # sample sizes
lambda <- 2 # parameter
E <- c()
V <-c()
n <- 1:length(simulations)
for (i in n){
val <- rpois(simulations[i],lambda)
e <- mean(val)
v <- var(val)
E[i] <- e
V[i] <- v
}
```

```
theoretical <- rep(2, length(simulations)) # theoretical = 2


# Plotting the Expectation and Variance


plot(simulations, E, pch = 20, main ="Poisson Distribution with lambda =2",
xlab="Sample Size", ylab = "E(X)")
points(simulations, theoretical, col="blue", type = "l")


plot(simulations, V, pch = 20, main ="Poisson Distribution with lambda =2",
xlab="Sample Size", ylab = "V(X)")
points(simulations, theoretical, col="blue", type = "l")
```

---

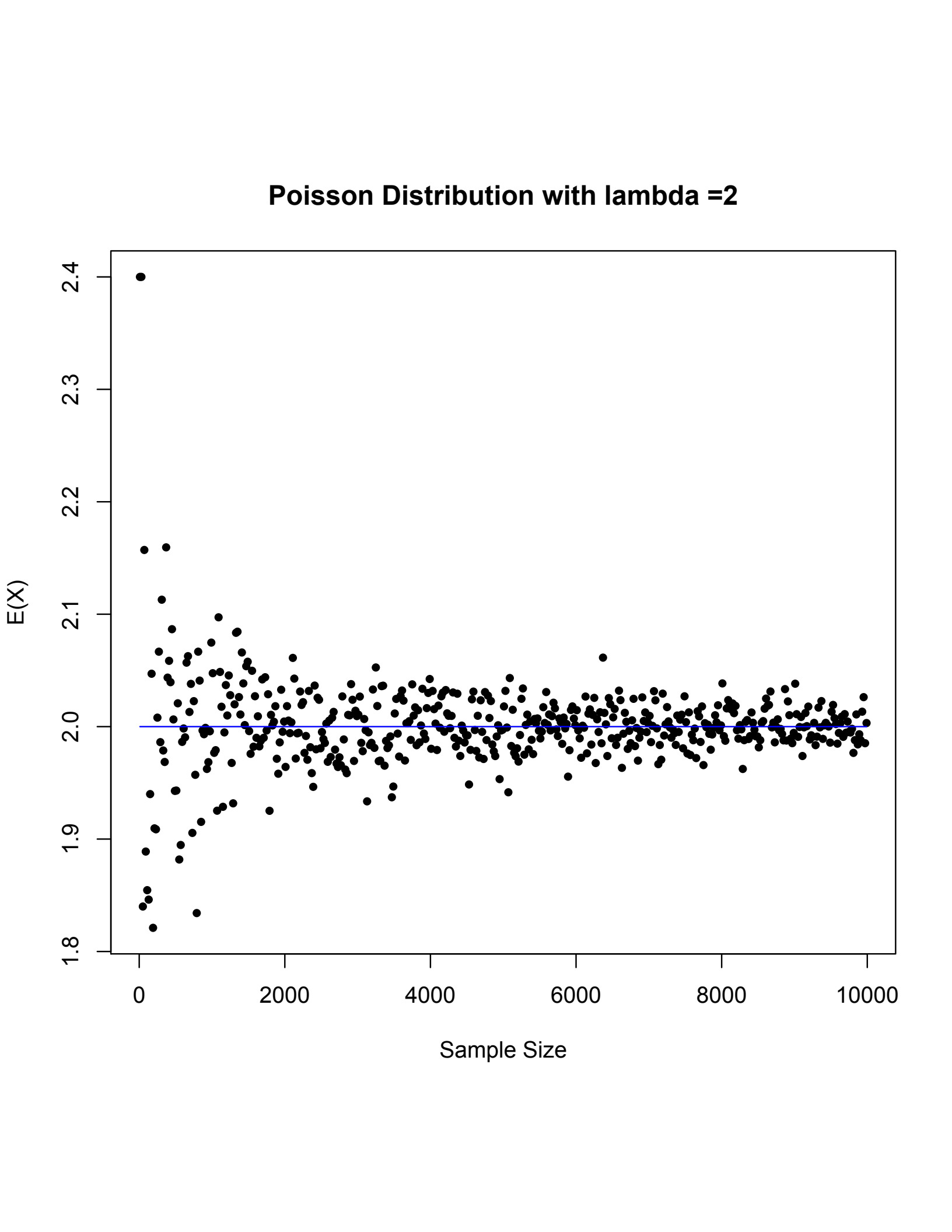
b)

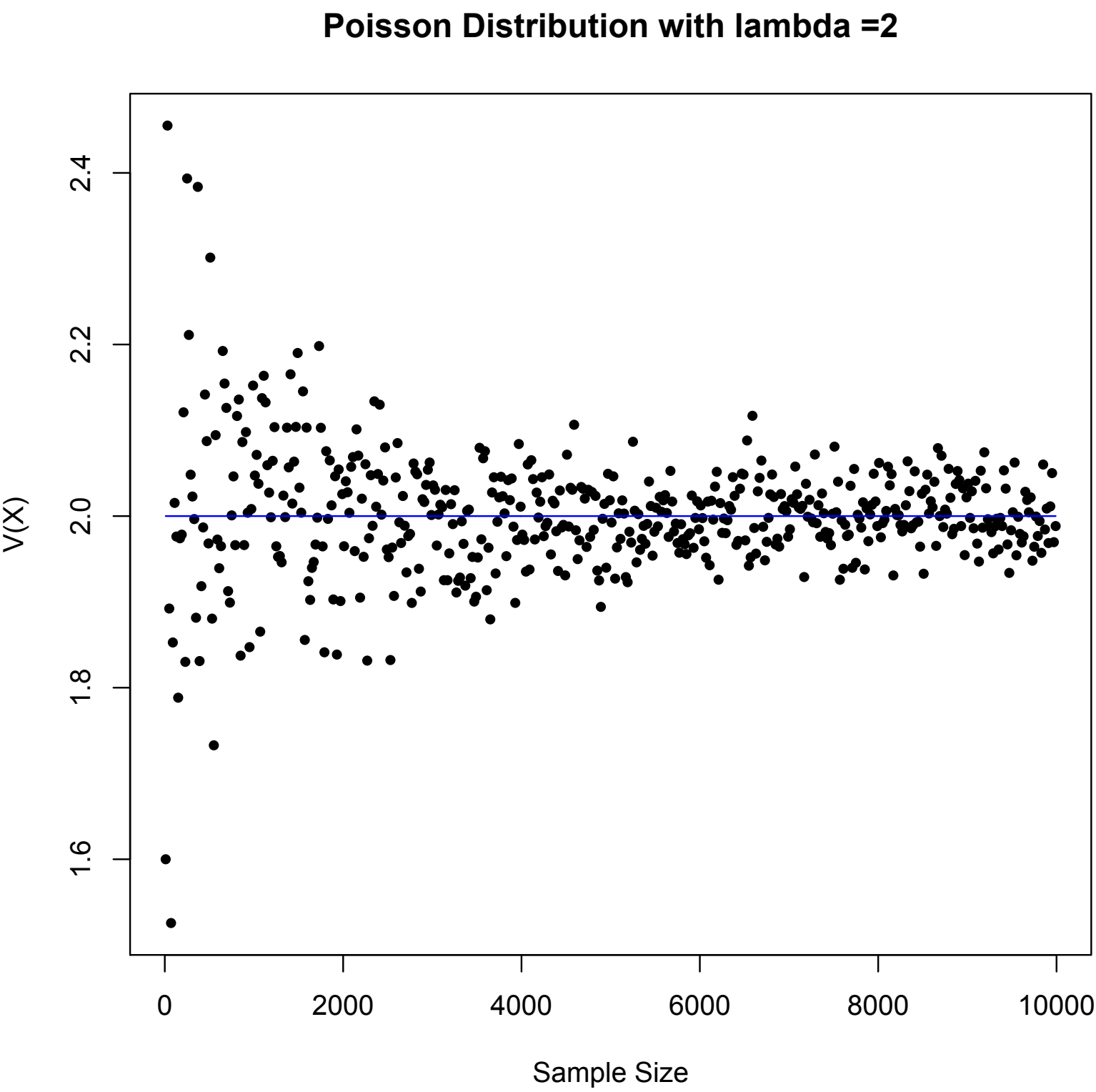
We show that if X has a Binomial distribtuion with $n = 400$ and $p = 0.3$ then $E(X) = np = 120$ and $V(X) = np(1-p) = 84$. We simulated different sized datasets, namely, 10 to 10000 in increments of 20 and found that the simulated data approach their respective population values. This is shown by the data points of mean and variance for the samples getting closer to the true value represented by the blue lines as sample size increases.

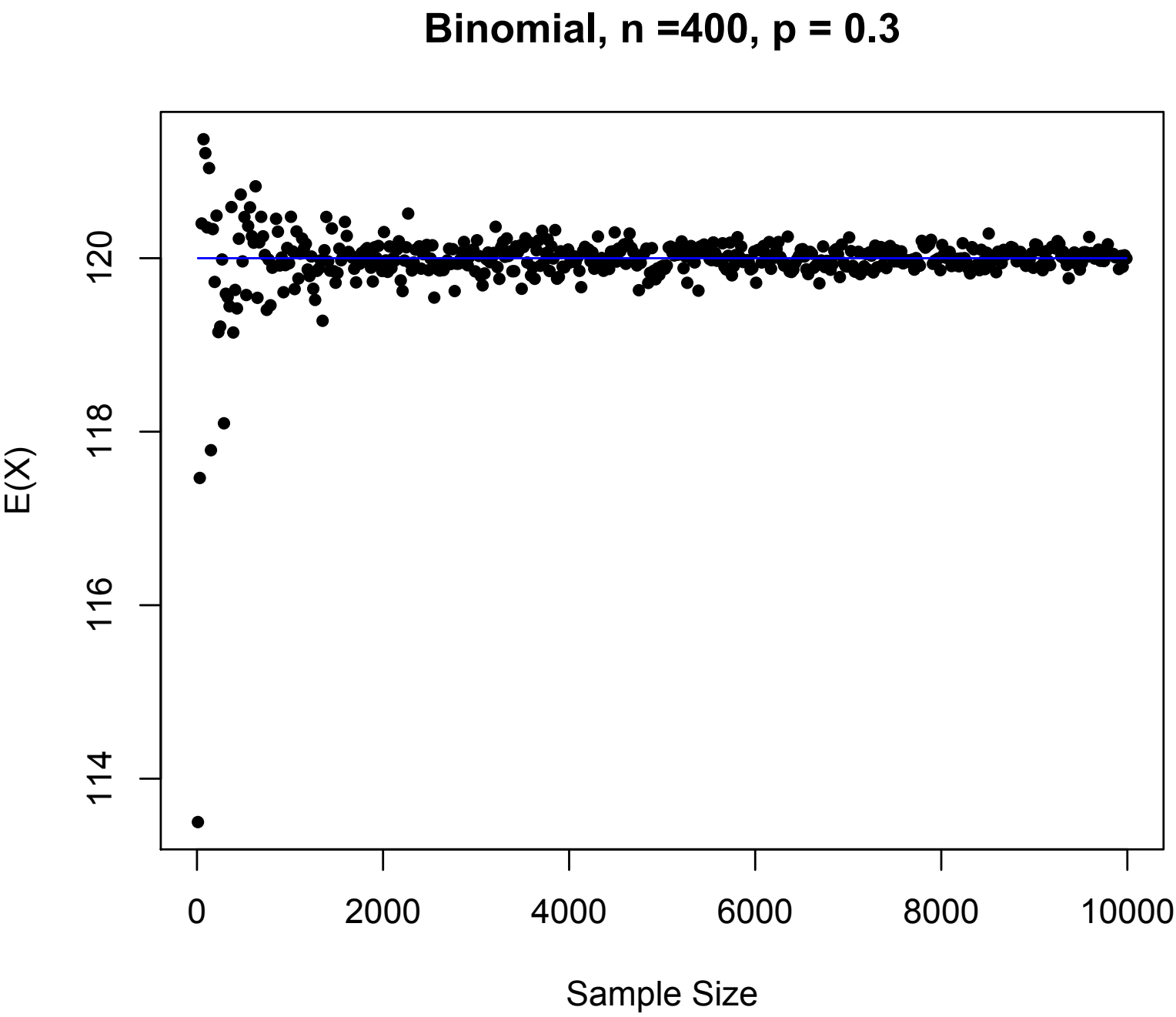
R Code:

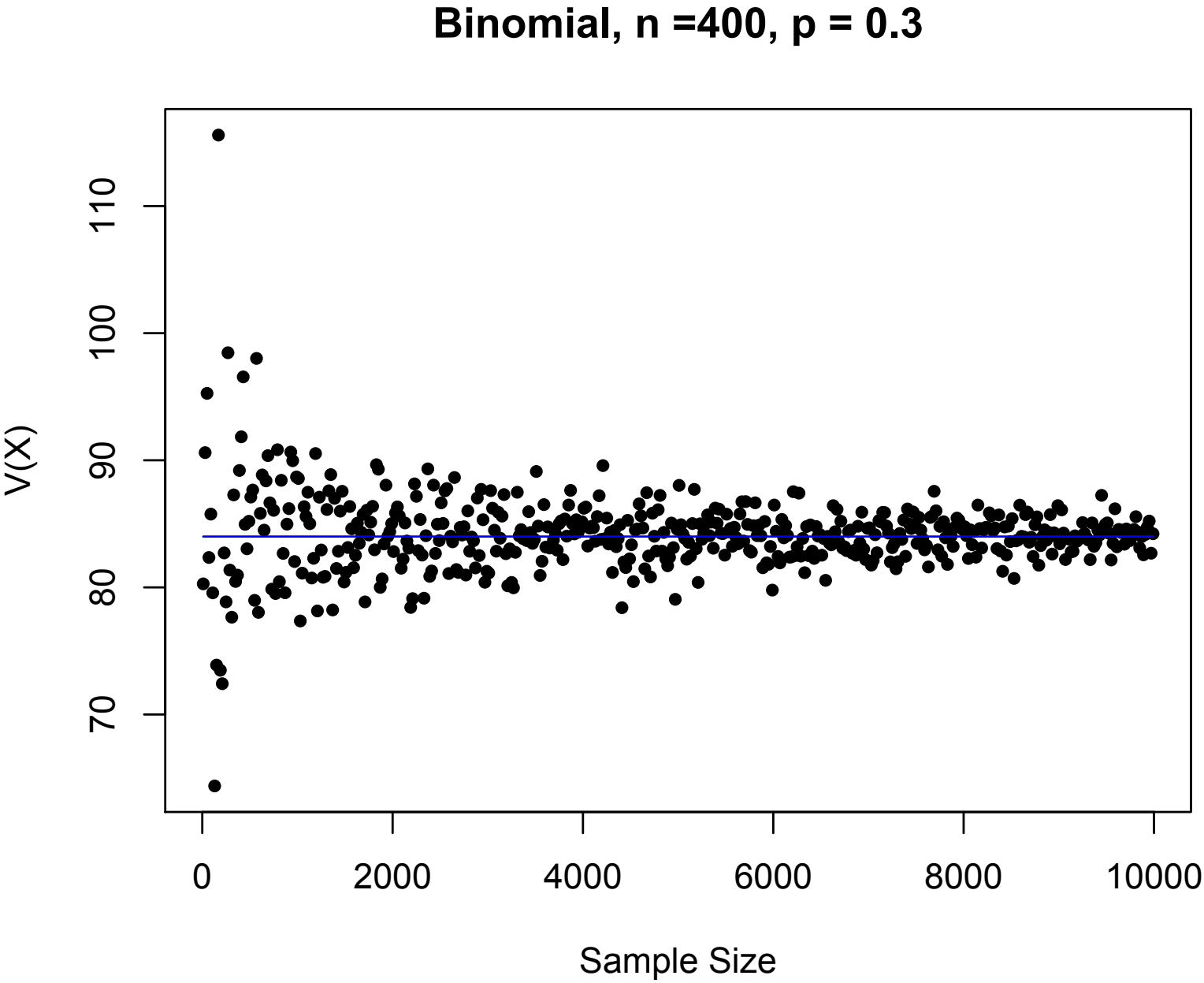
```
set.seed(42)
simulations <-  seq(10,10000, 20) # sample sizes

n <-400

p <- 0.3

E <- c()
V <-c()

m <- 1:length(simulations)
for (i in m){
val <- rbinom(simulations[i], n, p)
e <- mean(val)
v <- var(val)
E[i] <- e
```

```
V[i] <- v
}

# Calculating theoretical expectation and variance

theo_E <-rep(n*p, length(simulations)) # theoretical = np
theo_V <- rep(n*p*(1-p), length(simulations)) #theoretical = np(1-p)



# Plotting the Expectation and Variance

plot(simulations, E, pch = 20, main ="Binomial, n =400, p = 0.3", xlab="Sample
Size", ylab = "E(X)")
points(simulations, theo_E, col="blue", type = "l")


plot(simulations, V, pch = 20, main ="Binomial, n =400, p = 0.3", xlab="Sample
Size", ylab = "V(X)")
points(simulations, theo_V, col="blue", type = "l")
```


Poisson Distribution with lambda =2
E(X)
1.8
1.9
2.0
2.1
2.2
2.3
2.4
0
2000
4000
6000
8000
10000
Sample Size


Poisson Distribution with lambda =2
V(X)
1.6
1.8
2.0
2.2
2.4
0
2000
4000
6000
8000
10000
Sample Size


Binomial, n =400, p = 0.3
E(X)
114
116
118
120
0
2000
4000
6000
8000
10000
Sample Size


Binomial, n =400, p = 0.3
V(X)
70
80
90
100
110
0
2000
4000
6000
8000
10000
Sample Size

PROBLEM 6:

Suppose that you have a (possibly) biased coin that gives Heads with probability p, where p is unknown ($0 < p < 1$).

(a) Describe a way to *simulate* flipping a fair coin using only the (possibly) biased coin. *Hint: By considering a sequence of* **pairs** *of flips of the possibly biased coin, find an event that has probability exactly* $1/2$ *(for all values of p).*

(b) Find the expected number of times that you are required to flip the possibly biased coin before it is determined that the event in (a) either has or has not happened.

(c) Implement the procedure from (a) in `R` with a coin for which $p = 0.2$. Use the command `rbinom(1,1,p)` as a way of 'flipping' this biased coin once. Run your procedure 1000 times. Report your code.

Hint: you are being asked to run your **procedure** 1000 *times: each 'run' of your procedure may require several (simulations of) flips of the biased coin.*

(d) Report the average number of the biased coin flips per run of the procedure over these 1000 runs.

SOLUTION:
a) Let us consider a sequence of pairs of flips of the biased coin. For a given pair of flips, we have the following possible outcomes:

$$(H,H),(T,T),(H,T),(T,H)$$

Let us look at the probabilities of each of these:

$$P(H,H) = p^2;\ P(T,T) = (1-p)^2;\ P(H,T) = p(1-p);\ P(T,H) = p(1-p)$$

We can notice that $P(T,H) = P(H,T)$. These two events have the same probability so if we can consider just these outcomes, that is, create a sample space that consists of only these two events then we should have simulated a fair coin. Let $S = \{A, B\}$ where $A = (H,T)$ and $B = (T,H)$ and let $P(A) = r$ and $P(B) = s$. We know that $P(A) = P(B)$ from above and also $P(A) + P(B) = 1$ because these two events exhaust our sample space. Therefore, $2r = 1 \implies r = \frac{1}{2} = s$.

Thus, $P(H,T) = \frac{1}{2} = P(T,H)$ if we exclude the other two possibilities from our sample space of tossing the biased coin twice. We can choose to ignore (H,H) and (T,T) and not

count those outcomes when we flip the coin twice. We can consider the outcome of (H, T) to represent Heads and the outcome of (T, H) to represent Tails.

---

b) We can consider this to be a geometric distribution as we are repeatedly tossing two coins and we want to find the number of times we need to flip two coins until the first time the two flips will be different. The probability of the two coins being different is the following:

$$P((H, T) \text{ or } (T, H)) = p(1-p) + (1-p)p = 2p(1-p)$$

From our calculation in class of $E[X]$ for geometric random variable X, we have that:

$$\text{Expectation} = \frac{1}{2p(1-p)}$$

Since, we are flipping two coins at a time, the expected number of times that we need to flip the possibly biased coin is $2 * \frac{1}{2p(1-p)} = \frac{1}{p(1-p)}$.

---

c) We provide the code for the procedure here. Basically, we keep flipping until we get that the first flip is not the same outcome as the second flip. We count each flip individually rather than counting two flips as one event.

R Code:

```
set.seed(42)
p <- 0.2 # probability of heads for biased coin
procedure <- 1:1000

tot_num_flips = 0

for (i in procedure){
num_flips <- 0
repeat{
x_val <- c()
x_val[1] <-  rbinom(1, 1, p) # 1st flip
x_val[2]<- rbinom(1,1,p)  # 2nd flip
num_flips <- num_flips + 2   # doing 2 flips
if (x_val[1] != x_val[2]) {  # if the flips are different
break
}
}
```

```
tot_num_flips <- tot_num_flips + num_flips
}
```

```
avg_num_flips <- tot_num_flips/length(procedure)
print(avg_num_flips)
```

---

d) The code from part c) gives us the answer. The average number of the biased coin flips per run of the procedure over these 1000 runs is 6.24. This makes sense because in part b) we said that the expected number of flips would be $\frac{1}{p(1-p)} = \frac{1}{0.2*0.8} = 6.25$.

PROBLEM 7:

Goldberger Qu. 3.2

For each of the following distributions for the random variable X, calculate $E(X)$ and $V(X)$:

(a) Discrete uniform, parameter $N = 9$.

(b) Binomial, parameters $n = 2$, $p = 0.4$.

(c) Binomial, parameters $n = 4$, $p = 0.6$.

(d) Poisson, parameter $\lambda = \frac{3}{2}$.

(e) Rectangular on the interval $[0, 2]$.

(f) Exponential, parameter $\lambda = 2$.

(g) Power on $[0, 1]$, parameter $\theta = 2$.

SOLUTION:
a) Discrete uniform, parameter $N = 9$.

Using Table 3.1, we have that for Discrete uniform, parameter N the expectation and the variance are given by the following formulae:

$$E(X) = \frac{(N+1)}{2}, \quad V(X) = \frac{(N^2-1)}{12}$$

In our case, $N = 9$. Hence,

$$E(X) = \frac{9+1}{2} = 5, \quad V(X) = \frac{81-1}{12} = \frac{80}{12} = \frac{20}{3}.$$

Therefore, $E(X) = 5$ and $V(X) = \frac{20}{3}$.

---

b) Binomial, parameters $n = 2$, $p = 0.4$.

From Table 3.1, we have that for Binomial with parameters n and p we have that:

$$E(X) = np, \quad V(X) = np(1-p)$$

So for $n = 2$ and $p = 0.4$ we have:

$$E(X) = 0.4 \times 2 = 0.8, \quad V(X) = 0.4 \times 2 \times (1 - 0.4) = 0.8 \times 0.6 = 0.48.$$

So, $E(X) = 0.8$ and $V(X) = 0.48$.

---

c) Binomial, parameters $n = 4$, $p = 0.6$.

Again, using the same formulae mentioned above we have:

$$E(X) = np = 4 \times 0.6 = 2.4, \quad V(X) = np(1 - p) = 2.4 \times 0.4 = 0.96.$$

Thus, $E(X) = 2.4$ and $V(X) = 0.96$.

---

d) Poisson, parameter $\lambda = \frac{3}{2}$

From Table 3.1, we have that for Poisson distribution with parameter λ:

$$E(X) = \lambda, \quad V(X) = \lambda.$$

In our case $\lambda = \frac{3}{2}$ and so we have that $E(X) = \frac{3}{2}$ and $V(X) = \frac{3}{2}$.

---

e) Rectangular on the interval [0,2]

Again, using Table 3.1, we have that for rectangular distribution on the interval $[a, b]$:

$$E(X) = \frac{(a + b)}{2}, \quad V(X) = \frac{(b - a)^2}{12}.$$

In the problem, we are given that $a = 0$ and $b = 2$. Plugging this into the equations above give us the following:

$$E(X) = \frac{0 + 2}{2} = 1, \quad V(X) = \frac{(2 - 0)^2}{12} = \frac{1}{3}.$$

Hence, $E(X) = 1$ and $V(X) = \frac{1}{3}$.

---

f) Exponential, parameter $\lambda = 2$.

Table 3.1 gives us that for an exponential distribution with given parameter λ:

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}.$$

We have that $\lambda = 2$ and hence:

$$E(X) = \frac{1}{2}, \quad V(X) = \frac{1}{2^2} = \frac{1}{4}.$$

So, $E(X) = \frac{1}{2}$ and $V(X) = \frac{1}{4}$.

---

g) Power on $[0, 1]$, parameter $\theta = 2$.

From Table 3.1, we have that for Power distribution on $[0, 1]$ and parameter θ:

$$E(X) = \frac{\theta}{(1+\theta)}, \quad V(X) = \frac{\theta}{[(1+\theta)^2(2+\theta)]}$$

Given $\theta = 2$ we get the following:

$$E(X) = \frac{2}{(1+2)} = \frac{2}{3}, \quad V(X) = \frac{2}{(1+2)^2(2+2)} = \frac{2}{9 \times 4} = \frac{1}{18}.$$

Therefore, $E(X) = \frac{2}{3}$ and $V(X) = \frac{1}{18}$.