STT 810

Homework 3

Aditya Jain

2022-10-17

Contents

Question 1

For the following 4 discrete distributions, calculate the expected value and standard deviation in following two/ three ways. Compare to make sure you get approximately the same result for each.

1. Create a simulation of 1,000,000 outcomes, and calculate the average and standard deviation of the outcomes;
2. Use the formulas for expected values and variance for standard distributions (for example, for binomial, E(X) = np); and
3. For binomial distribution only use sequences of the possible outcomes to generate the expected value and standard deviation.

a. binomial, 12 trials, p = 0.2

```
# Method 1
xa <- rbinom(n = 1000000, size = 12, prob = 0.2)
paste(round(mean(xa),2), round(sd(xa),2), "are mean and standard deviation")
```

```
## [1] "2.4 1.39 are mean and standard deviation"
```

```
# Method 2
n= 12
p= 0.2
expected_value <- n*p
standard_deviation <- sqrt(expected_value*0.8)
paste(round(expected_value,2), round(standard_deviation,2),"are mean and standard deviation")
```

```
## [1] "2.4 1.39 are mean and standard deviation"
```

```
# Method 3
xa_seq <- lapply(1:1000000, function(x)
  sum(sample(x = c(0,1),size = 12, replace = TRUE, prob = c(0.8,0.2))))
xa_seq2 <- c()
for (i in 1:10000){
  xa_seq2[i] <- xa_seq[[i]]}
paste(round(mean(xa_seq2),2), round(sd(xa_seq2),2),"are mean and standard deviation")
```

```
## [1] "2.4 1.39 are mean and standard deviation"
```

b. Exponential, lamda = 0.03

```
# Method 1
xb <- rexp(n = 1000000, rate = 0.03)
paste(round(mean(xb),3), round(sd(xb),3), "are mean and standard deviation")
```

```
## [1] "33.316 33.338 are mean and standard deviation"
```

```r
# Method 2
expected_value <- 1/0.03
standard_deviation <- sqrt(1/(0.03^2))
paste(round(expected_value,3), round(standard_deviation,3),"are mean and standard deviation")
```

```
## [1] "33.333 33.333 are mean and standard deviation"
```

c. Poisson, rate = 0.4/minute, t = 20 minutes

```r
# Method 1
xc <- rpois(n = 1000000,lambda = 0.4*20)
paste(round(mean(xc),3), round(sd(xc),3), "are mean and standard deviation")
```

```
## [1] "7.998 2.828 are mean and standard deviation"
```

```r
# Method 2
expected_value <- 0.4*20
standard_deviation <- sqrt(0.4*20)
paste(round(expected_value,3), round(standard_deviation,3),"are mean and standard deviation")
```

```
## [1] "8 2.828 are mean and standard deviation"
```

d. Uniform, interval = [0, 6]

```r
# Method 1
xd <- runif(n = 1000000,min = 0, max = 6)
paste(round(mean(xd),3), round(sd(xd),3), "are mean and standard deviation")
```

```
## [1] "3.001 1.732 are mean and standard deviation"
```

```r
# Method 2
expected_value <- (0+6)/2
standard_deviation <- sqrt((1/12)*(6**2))
paste(round(expected_value,3), round(standard_deviation,3),"are mean and standard deviation")
```

```
## [1] "3 1.732 are mean and standard deviation"
```

Question 2

You are a manager for product development. You have a $500,000 budget to spend on research for a new product. There are two products you can research:

1. Product 1 has a 30% chance of producing a successful prototype. If a prototype is created, it has an 80% chance of making it to market. If the product makes it to market, the amount of sales it will produce is modeled as an exponential distribution with mean value $9,000,000.
2. Product 2 has a 60% chance of producing a successful prototype. If a prototype is created, it has an 90% chance of making it to market. If the product makes it to market, the amount of sales it will produce is modeled as a uniform distribution between $2,000,000 and $8,000,000. However, if successful, there is a 30% chance that the company will be successfully sued for patent infringement, in which case, the company will lose $1 and $3,000,000 (according to a uniform distribution).

a. Calculate the expected value of the alternatives and determine whether you should give the go-ahead for the development of product 1 or product 2

```
p_a <- 0.3*0.8
p_b <- 0.6*0.9
exp_a <- 0.3*0.8*(9000000) -500000
exp_b <- 0.6*0.9*(5000000 -0.3*2000000) - 500000
paste(exp_a,",", exp_b, "are expected values for option a and b respectively")
```

```
## [1] "1660000 , 1876000 are expected values for option a and b respectively"
```

```
paste("In respect to expected values calculated, option b seems more likely to succeed")
```

```
## [1] "In respect to expected values calculated, option b seems more likely to succeed"
```

b. Create a simulation for two products, with 1,000,000 runs. Confirm that the average value of the simulation is about equal to what you calculated for the alternatives in (a). For percent of the simulations do you at least make your money back for the development?

```
expected_value_a <- mean(sample(x=c(0,1),size = 1000000,replace = TRUE,
                                prob = c(0.76,0.24))*
                           rexp(n = 1000000,rate=1/9000000))-500000

expected_value_b <- mean(sample(x=c(0,1),size = 1000000,replace = TRUE,
                                prob = c(0.46,0.54))*runif(n =1000000,min=2000000,max=8000000) -
                           sample(x=c(0,1),size=1000000,replace=TRUE,prob=c(0.838,0.162))*
                           runif(n = 1000000,min = 1000000,max = 3000000)) - 500000

paste(round(expected_value_a,2), round(expected_value_b,2),"are expected values for option a and b resp
```

```
## [1] "1662672.63 1878732.04 are expected values for option a and b respectively"
```

```
percent_a <- mean(sample(x=c(0,1),size = 1000000,replace = TRUE,prob = c(0.76,0.24))*
                           rexp(n = 1000000,rate=1/9000000) > 500000)

percent_b <- mean((sample(x=c(0,1),size = 1000000,replace = TRUE,prob = c(0.46,0.54))*
```

```
                            runif(n =1000000,min=2000000,max=8000000) -
                            sample(x=c(0,1),size=1000000,replace=TRUE,prob=c(0.838,0.162))*
                            runif(n = 1000000,min = 1000000,max = 3000000)) > 500000)

paste(round(percent_a,2), round(percent_b,2),"are the times product 1 and 2 made more money than the co
```

```
## [1] "0.23 0.53 are the times product 1 and 2 made more money than the cost of development"
```

c. For what percent of simulations do you make more money for Product 1 vs. Product 2?

```
pa_more_pb = mean(sample(x=c(0,1),size = 1000000,replace = TRUE,prob = c(0.76,0.24))*
                            rexp(n = 1000000,rate=1/9000000)>
                  (sample(x=c(0,1),size = 1000000,replace = TRUE,prob = c(0.46,0.54))*
                            runif(n =1000000,min=2000000,max=8000000) -
                            sample(x=c(0,1),size=1000000,replace=TRUE,prob=c(0.838,0.162))*
                            runif(n = 1000000,min = 1000000,max = 3000000)))

paste(round(pa_more_pb,3), "is the times product 1 made more money than product 2")
```

```
## [1] "0.249 is the times product 1 made more money than product 2"
```

d. Suppose the company needs the revenue to pay off a $7 million loan. Which alternative is more likely to produce more than $7,000,000 in revenue?

```
percent_a <- mean(sample(x=c(0,1),size = 1000000,replace = TRUE,prob = c(0.76,0.24))*
                            rexp(n = 1000000,rate=1/9000000) > 7000000)

percent_b <- mean((sample(x=c(0,1),size = 1000000,replace = TRUE,prob = c(0.46,0.54))*
                            runif(n =1000000,min=2000000,max=8000000) -
                            sample(x=c(0,1),size=1000000,replace=TRUE,prob=c(0.838,0.162))*
                            runif(n = 1000000,min = 1000000,max = 3000000)) > 7000000)

paste(round(percent_a,3),round(percent_b,3),"are times product 1 and 2 made more than 7 million dollars
```

```
## [1] "0.111 0.076 are times product 1 and 2 made more than 7 million dollars of revenue"
```

```
paste("Thus alternative 1 that is product one is more likely to pay off the loan")
```

```
## [1] "Thus alternative 1 that is product one is more likely to pay off the loan"
```

Question 3

Define a multivariate normal distribution with mean value = (-1, 2), var(X) = 1, var(Y) = 2, and the correlation of X and Y = -0.3. Plot this distribution, by

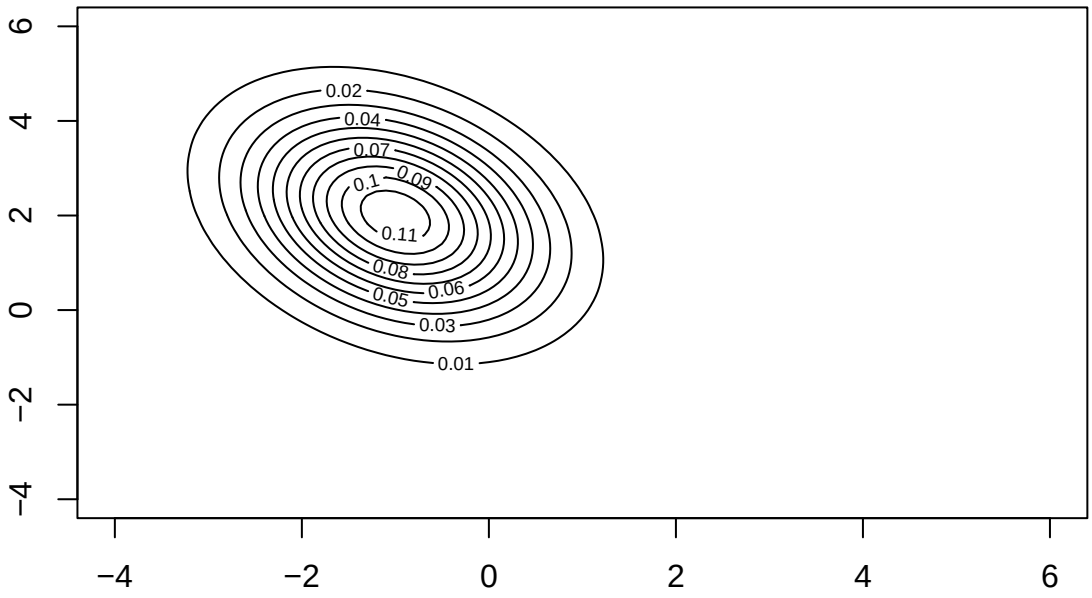
a. Creating a contour plot of the pdf, and

```
library(mvtnorm)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

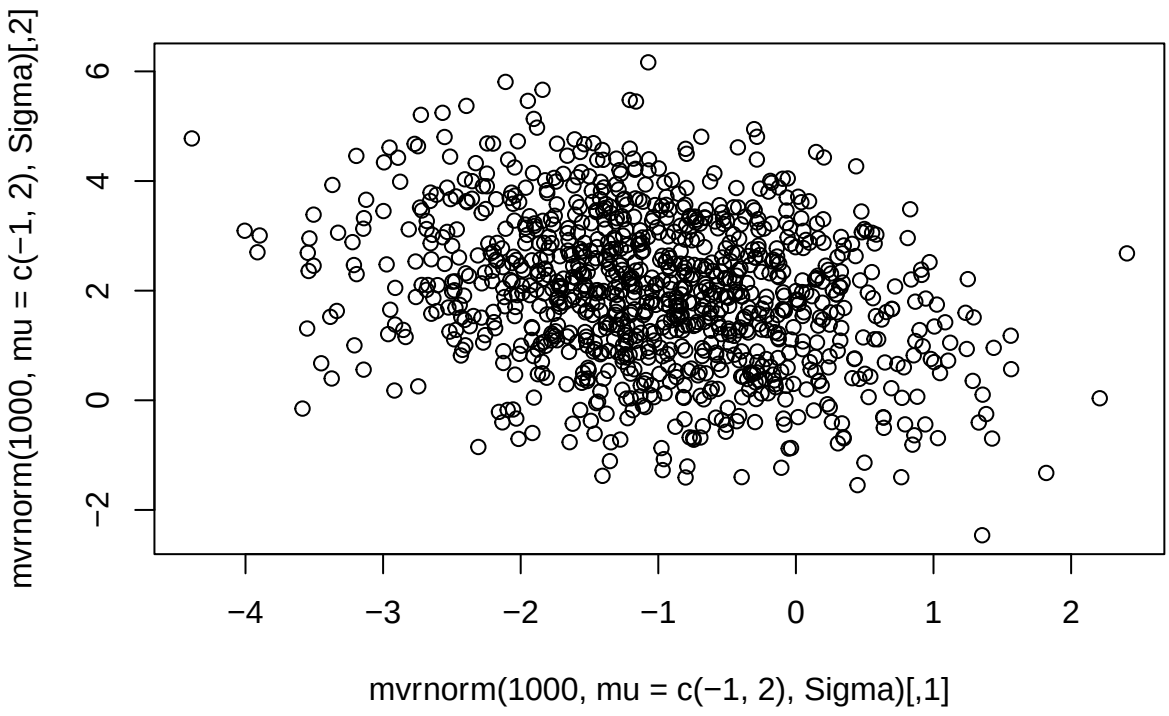
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
min <- -4
max <- 6
var_x <- 1
var_y <- 2
cor_xy <- -0.3
Sigma <- cbind(c(var_x,cor_xy*sqrt(var_x)*sqrt(var_y)),
               c(cor_xy*sqrt(var_x)*sqrt(var_y),var_y))
x <- seq(min, max, by = 0.1)
y <- seq(min, max, by = 0.1)
z <- matrix(nrow=length(x), ncol=length(y))
co_df <- data.frame('x', 'y', 'z')
for (i in 1:length(x)){
  for (j in 1:length(y)){
    z[i,j] <- dmvnorm(c(x[i],y[j]),c(-1,2),Sigma)
  }
}
contour(x, y, z)
```



b. Simulate the distribution 1,000 times and plot the simulated points.

```
plot(mvrnorm(1000, mu = c(-1,2), Sigma))
```



c. Next, using a larger simulation of the same distribution (1,000,000 values), estimate

(i) E(X)
(ii) E(Y)
(iii) Var(X) and Var(Y)
(iv) E(X + Y) and Var(X + Y). Compare the result to what the results of the formulas for linear combinations of random variables.
(v) E(X|Y = 3)

```
xc <- mvrnorm(1000000, mu = c(-1,2), Sigma)
expected_x <- mean(xc[,1])
expected_y <- mean(xc[,2])
var_x <- var(xc[,1])
var_y <- var(xc[,2])

expected_x_y <- mean(xc[,1] + xc[,2])
var_x_y <- var(xc[,1] + xc[,2])

cond_x <- c()
for(i in 1:1000000){
  if(xc[i,2] > 2.9 && xc[i,2] < 3.1){
    cond_x = append(cond_x, xc[i,1])
  }
}
cond_xm <- mean(cond_x)
```

```
paste("i. E(X)",round(expected_x,3))
```

```
## [1] "i. E(X) -0.999"
```

```
paste("ii. E(Y)",round(expected_y,3))
```

```
## [1] "ii. E(Y) 2"
```

```
paste("iii. Var(X), Var(Y)",round(var_x,3), round(var_y,3))
```

```
## [1] "iii. Var(X), Var(Y) 1.001 2.002"
```

```
paste("iv. E(X+Y), Var(X+Y)",round(expected_x_y,3), round(var_x_y,3))
```

```
## [1] "iv. E(X+Y), Var(X+Y) 1.001 2.156"
```

The expected value seem to follow the rule for linear combinations of random variable, however there is change in the value of variance

```
paste("v. E(X|Y = 3)", round(cond_xm,3))
```

```
## [1] "v. E(X|Y = 3) -1.209"
```

Question 4

4. Suppose we have an exponential probability distribution with parameter lamda = 1/3, so that its expected value is 3. We will perform some simulation experiments to determine the behavior of the sample mean.

a. Generate 2,000 simulations which each constitute 100 outcomes of this distribution. Calculate the mean values for the 2,000 experiments.

```
xa = lapply(1:2000, function(x) mean(rexp(100,1/3)))
xa_mean = c()
for (i in 1:2000){
  xa_mean[i] = xa[[i]]
}
```

b. Next, calculate the mean value and variance for the 2,000 sample means.

```
xb_mean = mean(xa_mean)
xb_var = var(xa_mean)
paste(round(xb_mean,3),round(xb_var,3))
```

```
## [1] "3.002 0.096"
```

c. Repeat (a) and (b) for N = 1,000, 10,000, 100,000, and 1,000,000 outcomes (2,000 simulations of each). Note: 1,000,000 could take several minutes.

```
xa = lapply(1:2000, function(x) mean(rexp(100,1/3)))
xa0 = lapply(1:2000, function(x) mean(rexp(1000,1/3)))
xa00 = lapply(1:2000, function(x) mean(rexp(10000,1/3)))
xa000 = lapply(1:2000, function(x) mean(rexp(100000,1/3)))
xa0000 = lapply(1:2000, function(x) mean(rexp(1000000,1/3)))

xa_mean = c()
xa_mean0 = c()
xa_mean00 = c()
xa_mean000 = c()
xa_mean0000 = c()
for (i in 1:2000){
  xa_mean[i] = xa[[i]]
  xa_mean0[i] = xa0[[i]]
  xa_mean00[i] = xa00[[i]]
  xa_mean000[i] = xa000[[i]]
  xa_mean0000[i] = xa0000[[i]]
}

xb_mean = mean(xa_mean)
xb_var = var(xa_mean)

xb_mean0 = mean(xa_mean0)
xb_var0 = var(xa_mean0)

xb_mean00 = mean(xa_mean00)
```

```
xb_var00 = var(xa_mean00)

xb_mean000 = mean(xa_mean000)
xb_var000 = var(xa_mean000)

xb_mean0000 = mean(xa_mean0000)
xb_var0000 = var(xa_mean0000)
```

```
paste("For N 100 mean and variance of the sample are", round(xb_mean,3), round(xb_var,3))
```

```
## [1] "For N 100 mean and variance of the sample are 2.99 0.093"
```

```
paste("For N 1000 mean and variance of the sample are", round(xb_mean0,3), round(xb_var0,3))
```

```
## [1] "For N 1000 mean and variance of the sample are 3 0.009"
```

```
paste("For N 10000 mean and variance of the sample are", round(xb_mean00,3), round(xb_var00,3))
```

```
## [1] "For N 10000 mean and variance of the sample are 3.001 0.001"
```

```
paste("For N 100000 mean and variance of the sample are", round(xb_mean000,3), round(xb_var000,3))
```

```
## [1] "For N 100000 mean and variance of the sample are 3 0"
```

```
paste("For N 1000000 mean and variance of the sample are", round(xb_mean0000,3), round(xb_var0000,3))
```

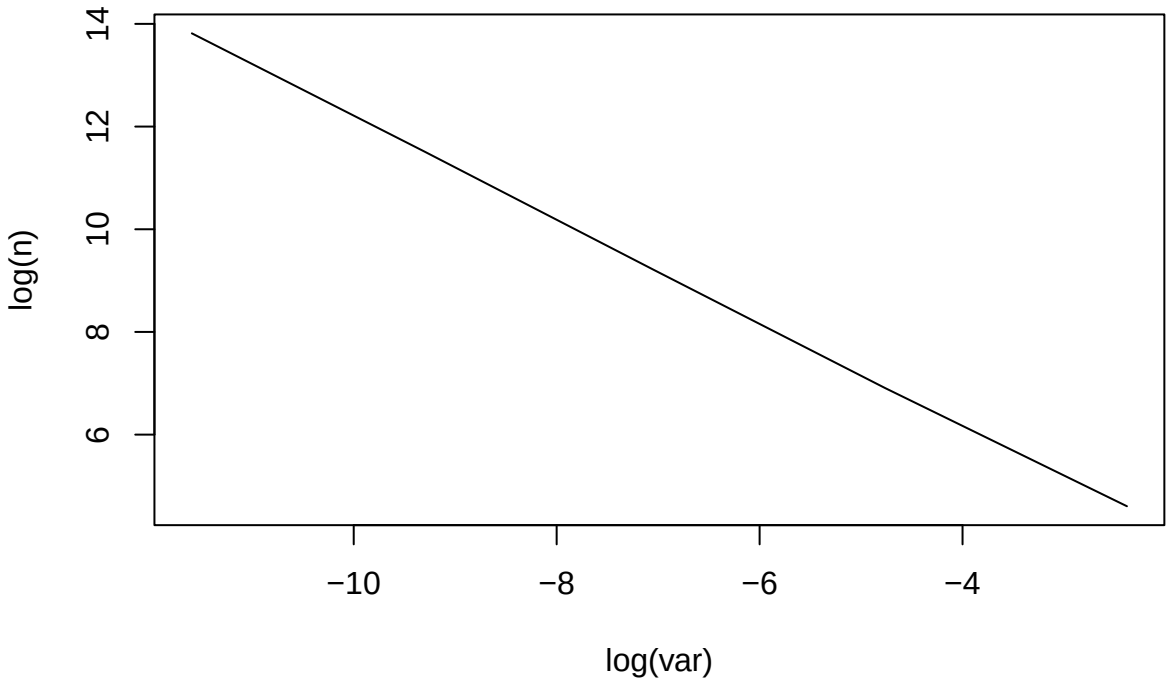
```
## [1] "For N 1000000 mean and variance of the sample are 3 0"
```

d. Plot the log(Variance) vs. log(N). Do you get something which is close to a straight line? What is the slope? Does the relationship between N and the variance close to what you would expect with the Central Limit Theorem?

```
n = c(100,1000,10000,100000,1000000)
var = c(xb_var,xb_var0,xb_var00,xb_var000,xb_var0000)
slope = (log(n[2])- log(n[1]))/(log(var[2])- log(var[1]))
paste("We do seem have a straight line with a negative slope of 45 degrees or slope =", round(slope,3))
```

```
## [1] "We do seem have a straight line with a negative slope of 45 degrees or slope = -0.966"
```

```
plot(log(var), log(n), type = "l")
```



The relationship between N and variance is close to central limit theorem as sample size increases variance decreases.

Question 5

Take a probability density function given by p(x) = 1/x2 , where x is between 1 and infinity.

a. Find the cdf. Verify that you get 0 and 1 at the minimum and "maximum" values.

Ans. CDF of p(x) = 1/xˆ2 is = -1/x" So to get 0 and 1 at the minimum and maximum values of x, the x should belong to (-infinity, -1]. And the above specified range is not valid for a probability distribution as values of CDF are negative for the range.

b. Find the quantile function.

Ans. Quintile function is inverse of cdf thus, interchanging x and y in the above equation gives us x = -1/y, which gives y = -1/x

c. Use the quantile function to simulate the distribution. Recreate the experiment in (4) for this disb. Find the quantile function.tribution (just create the table; you don't need to plot). Do you get the same convergence behavior? What is going on

```
xc = lapply(1:2000, function(x) mean(-1/runif(100)))
xc_mean = c()
for (i in 1:2000){
  xc_mean[i] = xc[[i]]
}
```

```
xc1_mean = mean(xc_mean)
xc1_var = var(xc_mean)
paste(round(xc1_mean,3),round(xc1_var,3))
```

```
## [1] "-11.515 959.288"
```

```
xa = lapply(1:2000, function(x) mean(-1/runif(100)))
xa0 = lapply(1:2000, function(x) mean(-1/runif(1000)))
xa00 = lapply(1:2000, function(x) mean(-1/runif(10000)))
xa000 = lapply(1:2000, function(x) mean(-1/runif(100000)))
xa0000 = lapply(1:2000, function(x) mean(-1/runif(1000000)))

xa_mean = c()
xa_mean0 = c()
xa_mean00 = c()
xa_mean000 = c()
xa_mean0000 = c()
for (i in 1:2000){
  xa_mean[i] = xa[[i]]
  xa_mean0[i] = xa0[[i]]
  xa_mean00[i] = xa00[[i]]
  xa_mean000[i] = xa000[[i]]
  xa_mean0000[i] = xa0000[[i]]
}

xb_mean = mean(xa_mean)
```

```
xb_var = var(xa_mean)

xb_mean0 = mean(xa_mean0)
xb_var0 = var(xa_mean0)

xb_mean00 = mean(xa_mean00)
xb_var00 = var(xa_mean00)

xb_mean000 = mean(xa_mean000)
xb_var000 = var(xa_mean000)

xb_mean0000 = mean(xa_mean0000)
xb_var0000 = var(xa_mean0000)
```

```
paste("For N 100 mean and variance of the sample are", round(xb_mean,3), round(xb_var,3))
```

```
## [1] "For N 100 mean and variance of the sample are -14.307 5059.749"
```

```
paste("For N 1000 mean and variance of the sample are", round(xb_mean0,3), round(xb_var0,3))
```

```
## [1] "For N 1000 mean and variance of the sample are -16.963 17621.534"
```

```
paste("For N 10000 mean and variance of the sample are", round(xb_mean00,3), round(xb_var00,3))
```

```
## [1] "For N 10000 mean and variance of the sample are -16.533 1914.124"
```

```
paste("For N 100000 mean and variance of the sample are", round(xb_mean000,3), round(xb_var000,3))
```

```
## [1] "For N 100000 mean and variance of the sample are -19.833 7469.483"
```

```
paste("For N 1000000 mean and variance of the sample are", round(xb_mean0000,3), round(xb_var0000,3))
```

```
## [1] "For N 1000000 mean and variance of the sample are -21.833 1519.202"
```

Mean values of the sample seems to deviate a little however, the variance of the sample shows similar trends of convergence and produces lower values as number of samples are increases.