1. (20 points)
X_1,…., X_5 are independent random samples from a distribution with mean 5 and standard deviation 3. Complete the following:
(a) For the sample mean $\overline{X} = \frac{1}{5}\sum_{i=1}^{5} Xi$, find its mean E ($\overline{X}$) and standard deviation sd ($\overline{X}$) .
(b) Can you find the P (2<$\overline{X}$<5.1) approximately? If yes, what is your estimate for P (2<$\overline{X}$<5.1) ? If no, why not?

Answer: a)
$\frac{1}{n}\sum_{i=1}^{n} \mu = \mu$ so mean is 5

$(\frac{1}{n})^2 \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$ = sd/n = 9/5 = 1.8

Rscript: b)
We cannot find the answer as the sample size is too low to use Central limit theorem. Though if we use CLT we will get the answer but this answer cannot be trusted.

```
pnorm(5.1,mean=5,sd=(9/5))-pnorm(2,mean=5,sd=(9/5))
```

Answer: b)
0.4743617

2. (20 points)
Suppose that for certain microRNA of size 20 the probability of a purine is binomially distributed with probability 0.7. Say there are 100 such microRNAs, each independent of the other. Let Y denote the average number of purine in these microRNAs. Find the probability that Y is great than 15. Please give a theoretical calculation, do NOT use Monte Carlo simulation to approximate. Show all the steps and formulas in your calculation.

Rscript:
Mean = np = 20*0.7 = 14
Variance = np(1-p) =4.2

```
pnorm(20,mean=14,sd=sqrt(4.2)/sqrt(100))-
pnorm(15,mean=14,sd=sqrt(4.2)/sqrt(100))
```

Answer:
5.317746e-07

explanation.
We can use pnorm function to find the approximate probability. We are finding probability of size 20 for 100 samples and subtracting it with probability of size 15 for 100 samples giving the probability of size above 15.

3. (20 points) Two genes' expression values follow a bivariate normal distribution. Let X and Y denote their expression values respectively. Also assume that X has mean 9 and variance 3; Y has mean 10 and variance 5; and the covariance between X and Y is 2.

In a trial, 50 independent measurements of the expression values of the two genes are collected, and denoted as (X_1,Y_1), ..., (X_{50},Y_{50}) . We wish to find the probability P ($\overline{X}$+0.5<$\overline{Y}$), that is, the probability that the sample mean for the second gene exceeds the sample mean of the first gene by more than 0.5.

Conduct a Monte Carlo simulation to approximate this probability, providing a 95% confidence interval for your estimation. Submit your R script for the Monte Carlo simulation, and a brief summary of the actual simulation results.

(Extra bonus: Provide a theoretical calculation of this probability. While the formula has not been given in the course lecture, it can be calculated from a bivariate normal distribution. You do not have to do this theoretical calculation if have no idea. You will get extra bonus points for doing it correctly.)

Rscript:

```
install.packages("mvtnorm")
library("mvtnorm")
ci<-0
py<-0
for (i in 1:100000){
xy<-rmvnorm(50,mean=c(9,10),sigma = matrix(c(3,sqrt(2),sqrt(2),5),nrow=2))
xbar[i]<-mean(xy[,1])
ybar[i]<-mean(xy[,2])

ci[i]<-(xbar[i]+0.5)<ybar[i]
if(((xbar[i]+0.5)<ybar[i]) == TRUE){
  py<-py+1
}
}
mean(ci) + c(-1,1)*1.96*sqrt(var(ci)/100000)
prob<- py/100000
prob
```

Answer:
Confidence interval
0.9394991 0.9424209

P ($\overline{X}$+0.5<$\overline{Y}$) = 0.94096

4. (20 points)
Assume there are three independent random variables X1 ~chisq (df =10) , X2 ~ Gamma(α=1, β=2), X3 ~ t-distribution with m=3 degrees of freedom. Define a new random variable Y as $Y = \sqrt{X1}X2 + 4(X3)^2$.

Use Monte Carlo simulation to find the mean of Y. Submit your R script for the Monte Carlo simulation, and a brief summary of the actual simulation results

Rscript:

```
x1<- rchisq(n=10000,df=10)
x2<- rgamma(n=10000,shape=1,scale=2)
x3<- rt(n=10000, df=3)
y<- sqrt(x1)*x2+ 4*(x3)^2
mean(y)
```

Answer:
17.97064

Brief explanation,
To use Monte Carlo simulation I am first generating 10000 variables for all X1, X2 and X3 using there corresponding functions.
Then I define random variable Y as asked in the question and I find its mean.
Higher the number of random variables generated better the results from Monte Carlo simulations.
Further we can check our Monte Carlo simulations result by finding the 95% confidence interval by following formula:-

```
x1<-matrix(rchisq(n=10000*1000,df=10),nrow=1000)
x2<-matrix(rgamma(n=10000*1000,shape=1,scale=2),nrow=1000)
x3<-matrix(rt(n=10000*1000, df=3),nrow=1000)
y<- sqrt(x1)*x2+ 4*(x3)^2
mean.y<-apply(y,1,mean)
mean(mean.y)+ c(-1,1)*1.96*sqrt(var(mean.y)/1000)
```

result 17.95167 18.23709

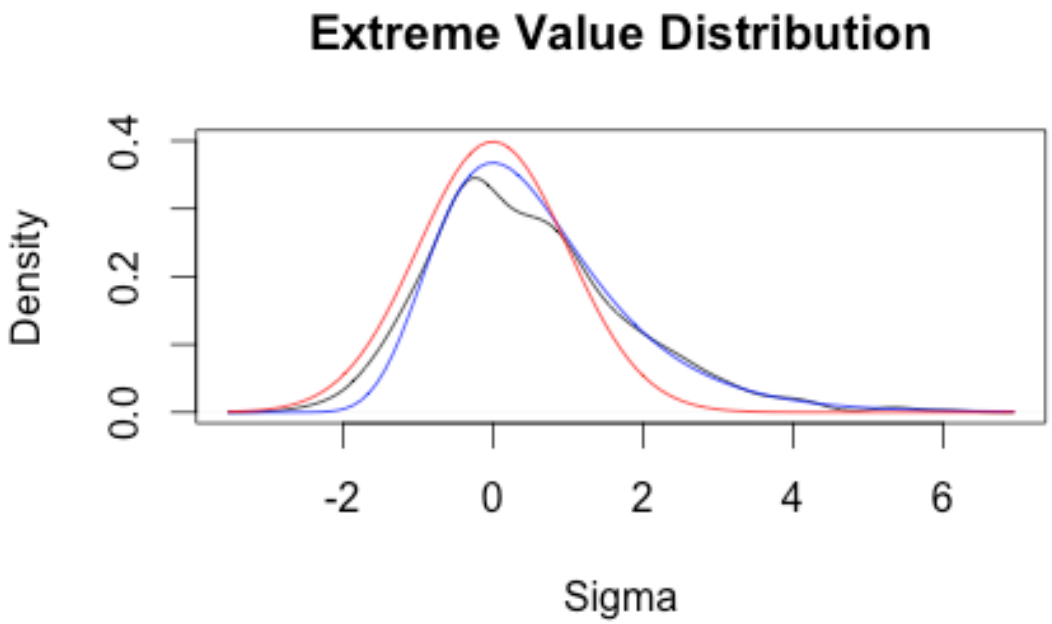
The above answer lies inside this confidence interval!.

5. (20 points)
Complete exercise 10 in Chapter 3 of Applied Statistics for Bioinformatics using R (page 45-46). Submit the plot, and a brief explanation of your observation. The problem refers to the density function of extreme value distribution in another book. You do not have to look for the other book, the density function is $f(x) = (e^{-x})e^{-e^{-x}}$.

Rscript:

```
n<-1000
an <- sqrt(2*log(n)) - 0.5*(log(log(n))+log(4*pi))*(2*log(n))^(-1/2)
bn <- (2*log(n))^(-1/2)
ex<-double()
for (i in 1:1000){
ex[i] <- (max(rnorm(n))-an)/bn
}
plot(density(ex),ylim=c(0,0.4),xlab="Sigma",ylab="Density",main="Extreme Value Distribution")
f<-function(x){exp(-x)*exp(-exp(-x))}
curve(f,range(density(ex)$x),add=TRUE,col = "blue")
curve(dnorm,add=TRUE,col = "red")
```

Answer:



From the above plot we can infer that blue line for extreme value distribution depicts the density function (black line) better than normal distribution(red line).