Question 1

A combined experiment is performed which involves tossing a balanced coin and rolling a balanced die.
(a) i) Describe the sample space and provide the table of the joined probability distribution of the combined experiment. Tabulate the marginal probabilities of the individual experiment.

Samplespace: Ω_c ={H,T}
Ω_D={1,2,3,4,5,6}

	1	2	3	4	5	6	$f_X(x)$
T	1/12	1/12	1/12	1/12	1/12	1/12	1/2
H	1/12	1/12	1/12	1/12	1/12	1/12	1/2
$f_Y(y)$	1/6	1/6	1/6	1/6	1/6	1/6	

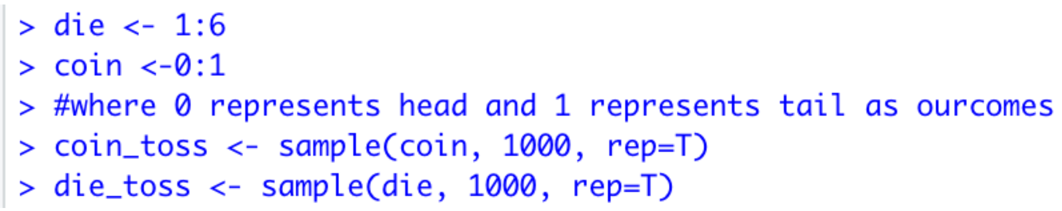
ii) If we define X to be equal to 1 if tossing a head and 0 if tossing a tail, and define Y to be equal to the number on the rolled die, calculate the expected value and variance of X, the expected value and variance of Y and the covariance between X and Y.

X=1|H X=0|T Y=the number on the rolled die
X and Y are discrete random variables with a probability mass function f_X and f_Y and range S_X={0,1}, S_Y={1,2,3,4,5,6}, respectively. Therefore we can calculate their expected values, variance and covariance in the following way:

$E[X]=\sum x_i f_X(x_i)$ $E[Y]=\sum y_i f_Y(y_i)$ $Var[X]=E[X^2]-(E[X])^2$ $Var[Y]=E[Y^2]-(E[Y])^2$ Cov[XY]= E[X,Y]-E[X]* E[Y] ~here we have 2 independent events, so E[X,Y]=E[X]*E[Y]

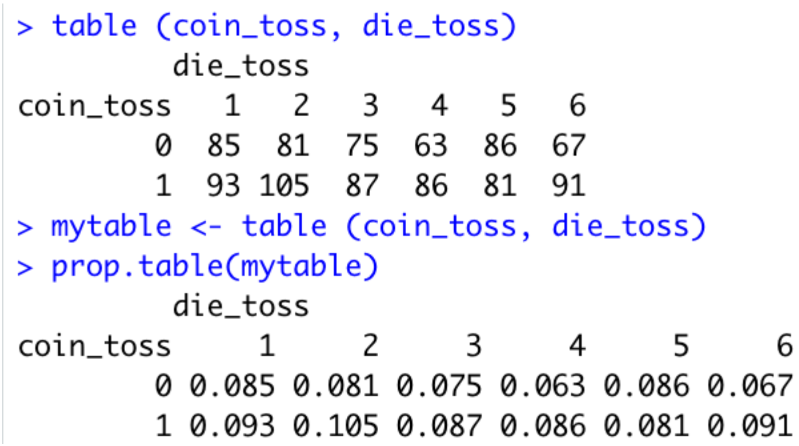
$E[X]= (0+1)/2=0.5$ $E[Y]=(1+2+3+4+5+6)/6=7/2$ Cov[XY]=0
$E[X^2]=(0^2+1^2)/2=0.5$ $E[Y^2]= (1^2+2^2+3^2+4^2+5^2+6^2)/6=91/6$
$Var[X]=0.5-0.25=0.25$ $Var[Y]=7/2-(91/6)^2=2.92$

(b) Use R to perform this combined 'experiment' 1000 times. Note that your solution should provide all 1000 combined outcomes of the coin toss and the roll of dice as two vectors each of size 1000, one containing 0 and 1 (to represent the tossing of the coin) and one containing numbers from 1 to 6 to represent the rolling of the dice.



```
> die <- 1:6
> coin <-0:1
> #where 0 represents head and 1 represents tail as outcomes
> coin_toss <- sample(coin, 1000, rep=T)
> die_toss <- sample(die, 1000, rep=T)
```

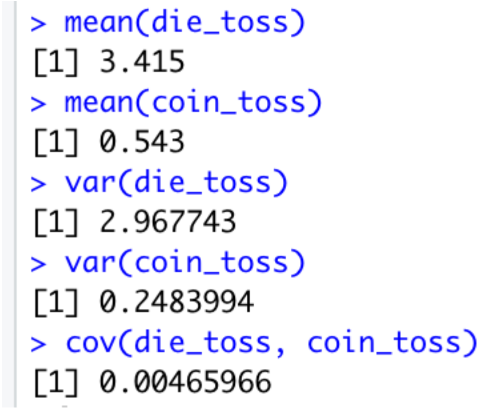
i) Produce a contingency table to show frequencies in the categories of one variable broken down by the categories of the other variable. Then use appropriate percentages to calculate the observed joined probability distribution. Compare your results with the theoretically expected results presented under point a).



```
> table (coin_toss, die_toss)
         die_toss
coin_toss   1   2   3   4   5   6
        0  85  81  75  63  86  67
        1  93 105  87  86  81  91
> mytable <- table (coin_toss, die_toss)
> prop.table(mytable)
         die_toss
coin_toss     1     2     3     4     5     6
        0 0.085 0.081 0.075 0.063 0.086 0.067
        1 0.093 0.105 0.087 0.086 0.081 0.091
```

Our obtained values for the observed joint probability distributions are very close to the theoretical 0.083 (or 1/12), that we have calculated in part a). Despite this, we can see an expected deviation from this value when sampling randomly. For example, we can see that the j.p.d of tossing a head and 2 is 0.105.

ii) Calculate the observed mean and variance of each of the two variables as well as the covariance. Compare your results with the theoretical results obtained under point a).



```
> mean(die_toss)
[1] 3.415
> mean(coin_toss)
[1] 0.543
> var(die_toss)
[1] 2.967743
> var(coin_toss)
[1] 0.2483994
> cov(die_toss, coin_toss)
[1] 0.00465966
```

Our obtained values here are very close to our theoretical values again. We compare the theoretical expected value of tossing an unbiased die (3.5) to our practical value of 3.415. The theoretical expected value of tossing a coin was 0.5 and we got a value of 0.543. Our variance was 0.25 for a coin toss in part a), now it is 0.248. Similarly, the variance for tossing the die was 2.92, and our new value is 2.97. And lastly, the theoretical covariance we calculated was 0, to which our practical value of 0.0057 is very close.

Question 2

Let X and Y be two independent random variables following a Poisson distribution with parameters 2 and 4 respectively, i.e.,
$X\sim Poisson(2)\ \ and\ \ Y\sim Poisson(4)$
Let $Z=X+Y$.

i) Specify the distribution of Z, the expected value and the variance.

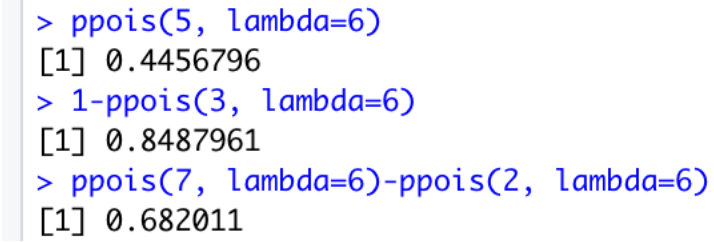
X and Y are independent random variables. The moment-generating function can be used to show other distributions are invariant to linear combinations of independent random variables.

If X_i~Poisson(l1), then for i=1,2,3…n, S_n~Poisson(l1+l2+l3…ln)

Therefore Z~Poisson(2+4)= P(6)

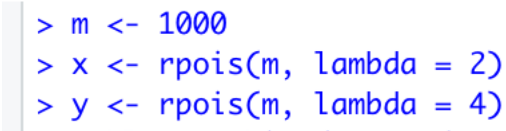
E[Z]=lambda=6   Var[Z]=lambda=6

ii) Use an appropriate R function to calculate:
$P(Z\leq5)$, $P(Z>3)$ $P(2\leq Z\leq7)$



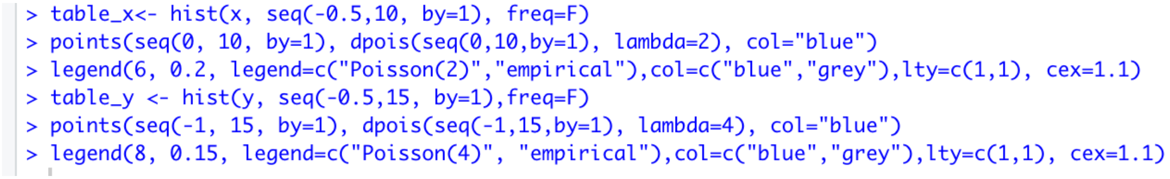
```
> ppois(5, lambda=6)
[1] 0.4456796
> 1-ppois(3, lambda=6)
[1] 0.8487961
> ppois(7, lambda=6)-ppois(2, lambda=6)
[1] 0.682011
```

b)
i) Sample 1000 values from the Poisson distribution $Poisson(2)$, and store the values in a vector x, then sample another 1000 values from the Poisson distribution $Poisson(4)$ and store the values in a vector y.

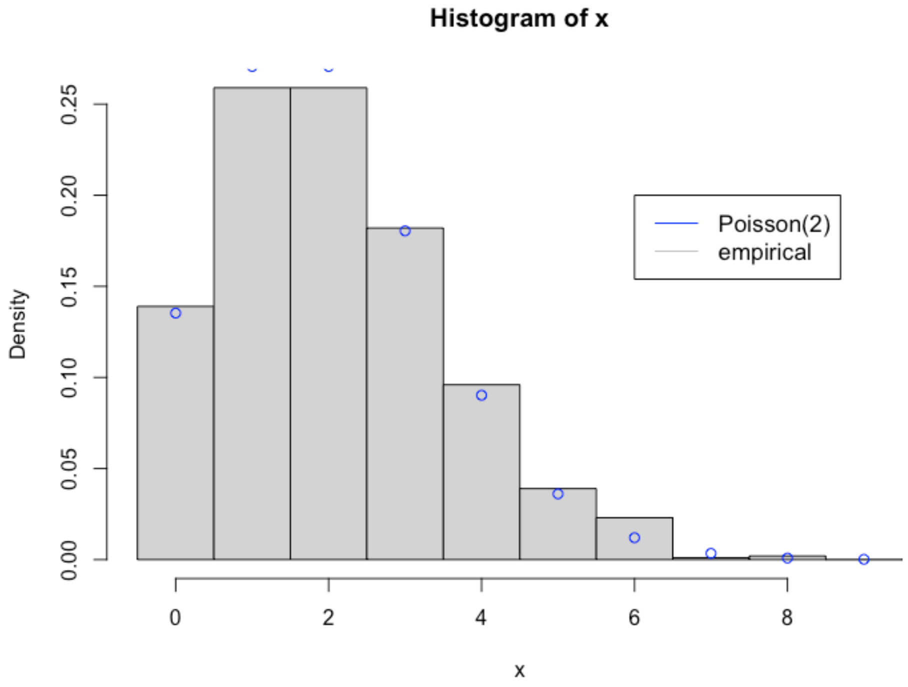


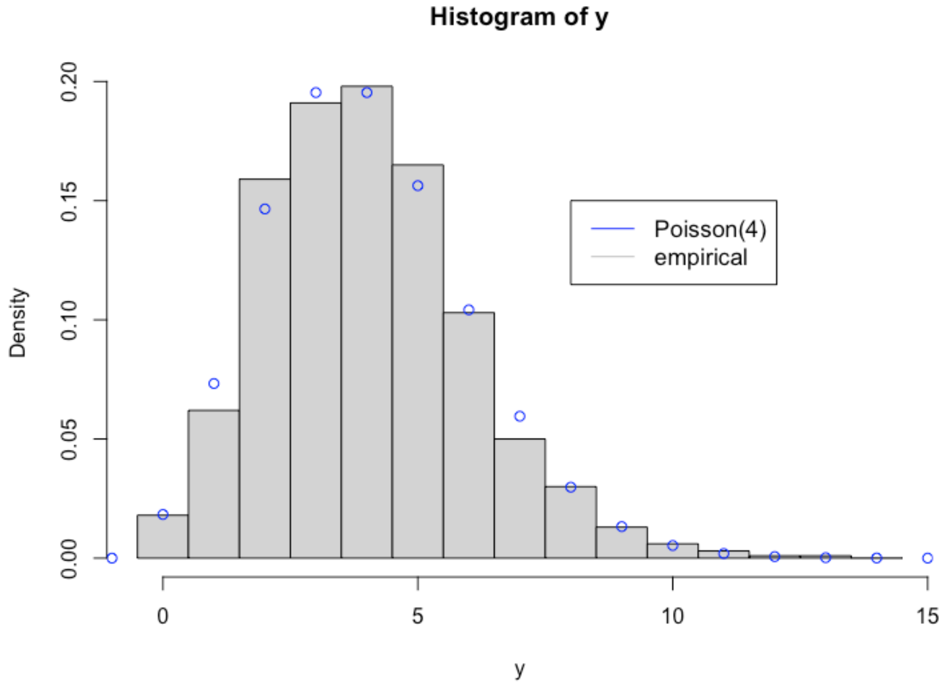
```
> m <- 1000
> x <- rpois(m, lambda = 2)
> y <- rpois(m, lambda = 4)
```

ii) Produce a relative frequency histogram of x and compare it with the underlying theoretical probability mass function. Also produce a relative frequency histogram for y and compare it with the underlying theoretical p.m.f. Then produce an R plot which contains the two distributions, use this plot to compare the location and spread of the two distributions.

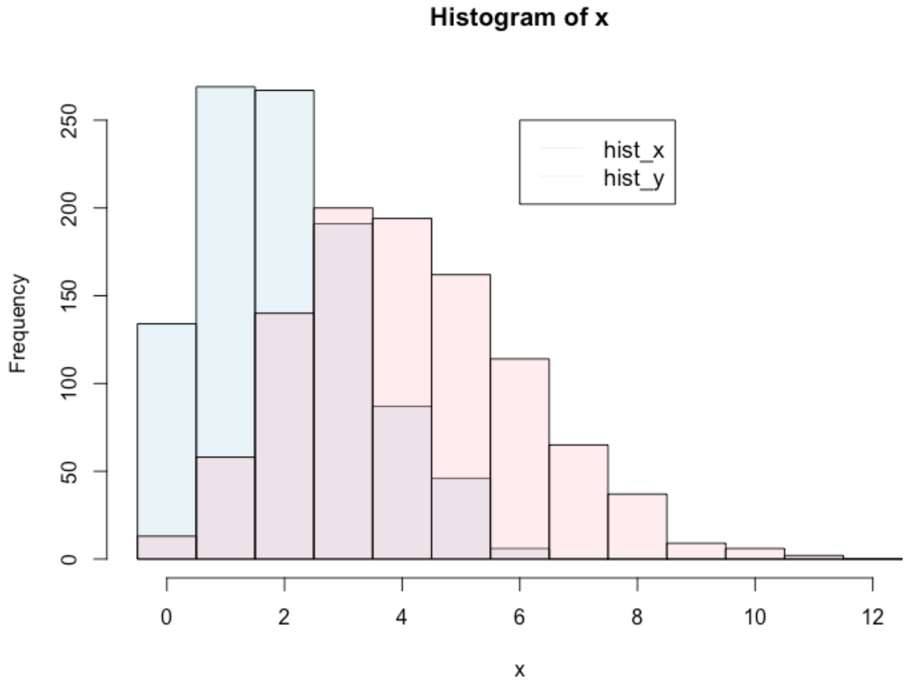


```
> table_x<- hist(x, seq(-0.5,10, by=1), freq=F)
> points(seq(0, 10, by=1), dpois(seq(0,10,by=1), lambda=2), col="blue")
> legend(6, 0.2, legend=c("Poisson(2)","empirical"),col=c("blue","grey"),lty=c(1,1), cex=1.1)
> table_y <- hist(y, seq(-0.5,15, by=1),freq=F)
> points(seq(-1, 15, by=1), dpois(seq(-1,15,by=1), lambda=4), col="blue")
> legend(8, 0.15, legend=c("Poisson(4)", "empirical"),col=c("blue","grey"),lty=c(1,1), cex=1.1)
```



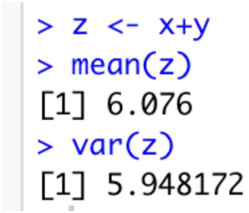


```
> hist_x<-hist(x,breaks=seq(-0.5,13, by=1))
> hist_y<-hist(y,breaks=seq(-0.5,13,by=1))
> c1 <- rgb(173,216,230,max = 255, alpha = 80, names = "lt.blue")
> c2 <- rgb(255,192,203, max = 255, alpha = 80, names = "lt.pink")
> plot(hist_x, col=c1)
> plot(hist_y,col=c2, add=TRUE)
> legend(6,250, legend=c("hist_x", "hist_y"), col=c(c1,c2), lty=c(1,1), cex=1.1)
```



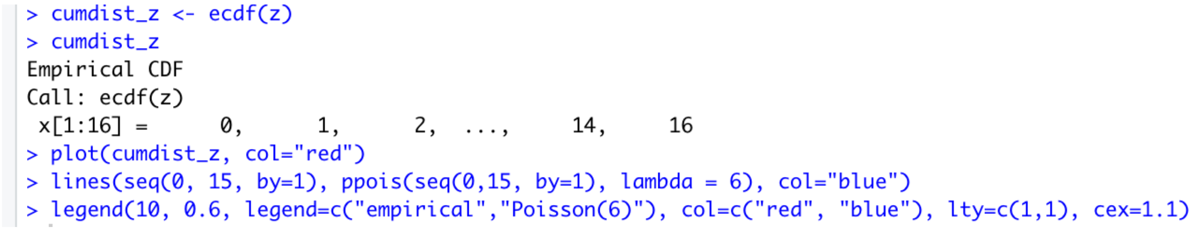
We can see that the spread of our two histograms differ, hist_x has a range from 0 to 6 and hist_y has a range 0 to 13. Their locations also differ hist_x obtains its highest values at around 1 and two, however, hist_y obtains them at around 3 and 4.

iii) Calculate the sum of the two vectors and store that in a new vector z. Calculate the mean and the variance of z and compare it to the true mean and variance provided under point a)

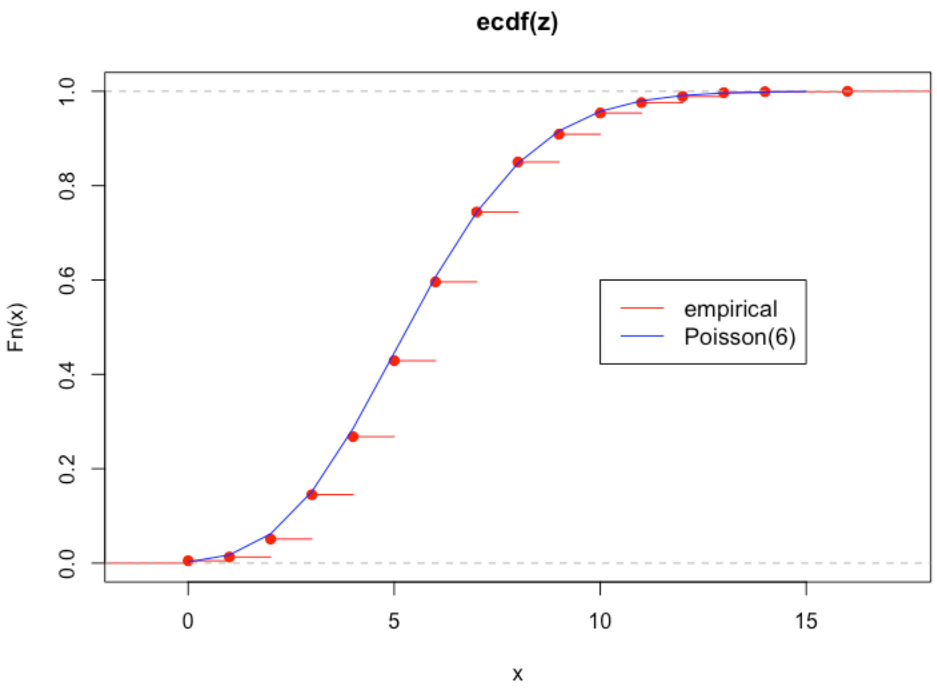


```
> z <- x+y
> mean(z)
[1] 6.076
> var(z)
[1] 5.948172
```

iv) Plot the empirical cumulative distribution function of z and compare it with the theoretical cumulative distribution What does the plot suggest?

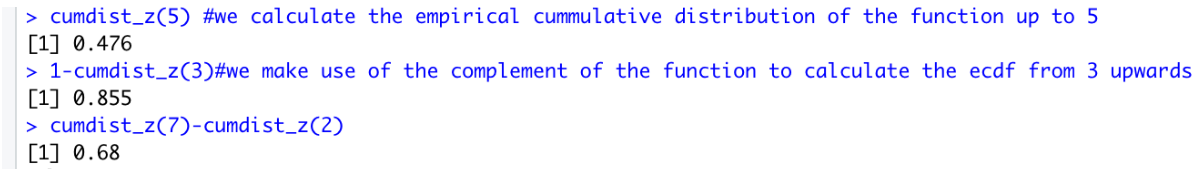


```
> cumdist_z <- ecdf(z)
> cumdist_z
Empirical CDF
Call: ecdf(z)
 x[1:16] =      0,      1,      2,  ...,      14,      16
> plot(cumdist_z, col="red")
> lines(seq(0, 15, by=1), ppois(seq(0,15, by=1), lambda = 6), col="blue")
> legend(10, 0.6, legend=c("empirical","Poisson(6)"), col=c("red", "blue"), lty=c(1,1), cex=1.1)
```



The plot suggests that the theoretical and empirical values of our Poisson(6) distribution are very close to each other, therefore we can predict our values accurately with theoretical calculations.

v) Use the empirical cumulative function to calculate $P(2 \leq z \leq 7)$ from the sample z, compare this to the theoretical probability obtained under point a).



```
> cumdist_z(5) #we calculate the empirical cummulative distribution of the function up to 5
[1] 0.476
> 1-cumdist_z(3)#we make use of the complement of the function to calculate the ecdf from 3 upwards
[1] 0.855
> cumdist_z(7)-cumdist_z(2)
[1] 0.68
```

Our values are very close to our predicted values yet again.

Question 3

Let $X1,X2,\ldots,Xn$ be n independent and identically distributed random variables following the standardised normal distribution:

$X1,X2,\ldots,Xn\sim N(0,1)$

Let $S=X12+X22+\cdots+Xn2$

a)

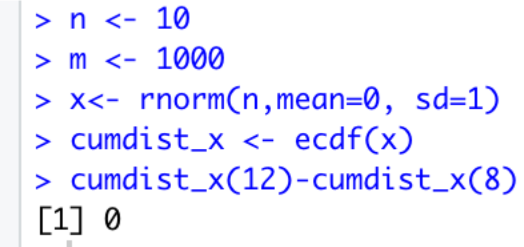
i) Specify the distribution of S, the expected value and the variance.

These are independent, identically distributed variables following a standardised normal distribution. X_i~N(0,1)

S is a sum of squared random variables which follow N(0,1), S~chi^2(n) because of the relationship between the normal and chi-squared distributions.

ii) For n=10, use an appropriate R function to calculate:

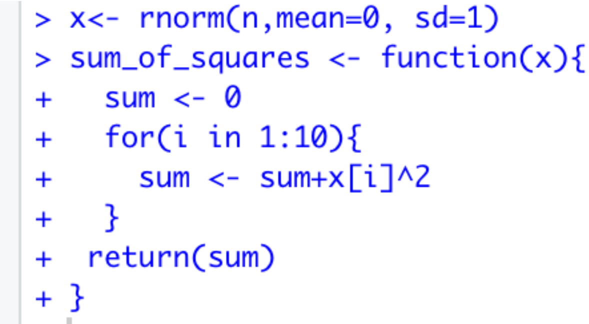
$$P(8\leq S\leq 12)$$



```
> n <- 10
> m <- 1000
> x<- rnorm(n,mean=0, sd=1)
> cumdist_x <- ecdf(x)
> cumdist_x(12)-cumdist_x(8)
[1] 0
```

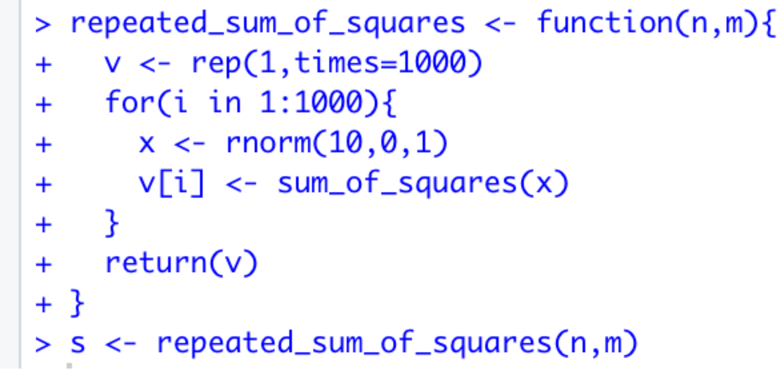
b) Let's consider n=10 in the above example.

i) Use R to sample n=10 values from the standardised normal distribution $N(0,1)$, store the values in a vector x of size 10 and then calculate the sum of squared values of x.



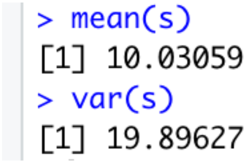
```
> x<- rnorm(n,mean=0, sd=1)
> sum_of_squares <- function(x){
+   sum <- 0
+   for(i in 1:10){
+     sum <- sum+x[i]^2
+   }
+  return(sum)
+ }
```

ii) Repeat step i) above m=1000 times and save the result in a vector s that should contain 1000 values.



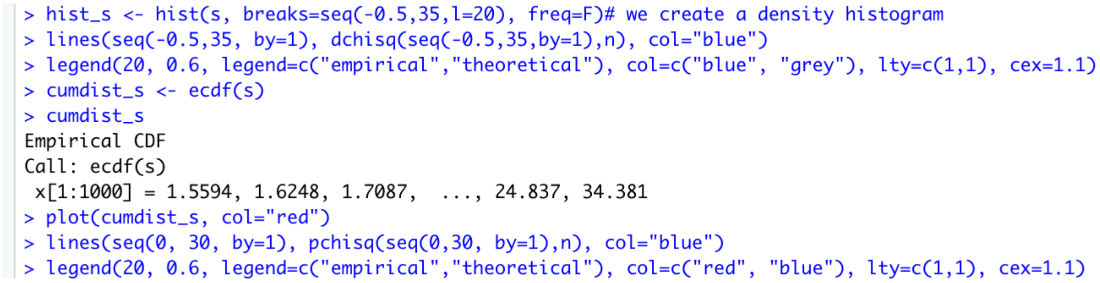
```
> repeated_sum_of_squares <- function(n,m){
+   v <- rep(1,times=1000)
+   for(i in 1:1000){
+     x <- rnorm(10,0,1)
+     v[i] <- sum_of_squares(x)
+   }
+   return(v)
+ }
> s <- repeated_sum_of_squares(n,m)
```

iii) Calculate the mean and the variance of s and compare it against the theoretical mean and variance obtained under point a).

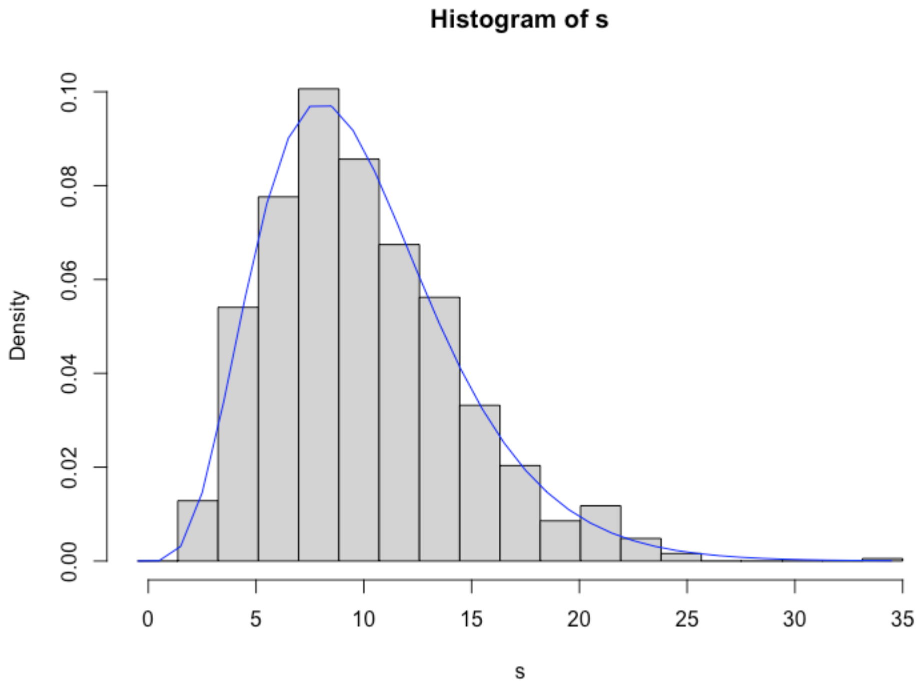


```
> mean(s)
[1] 10.03059
> var(s)
[1] 19.89627
```

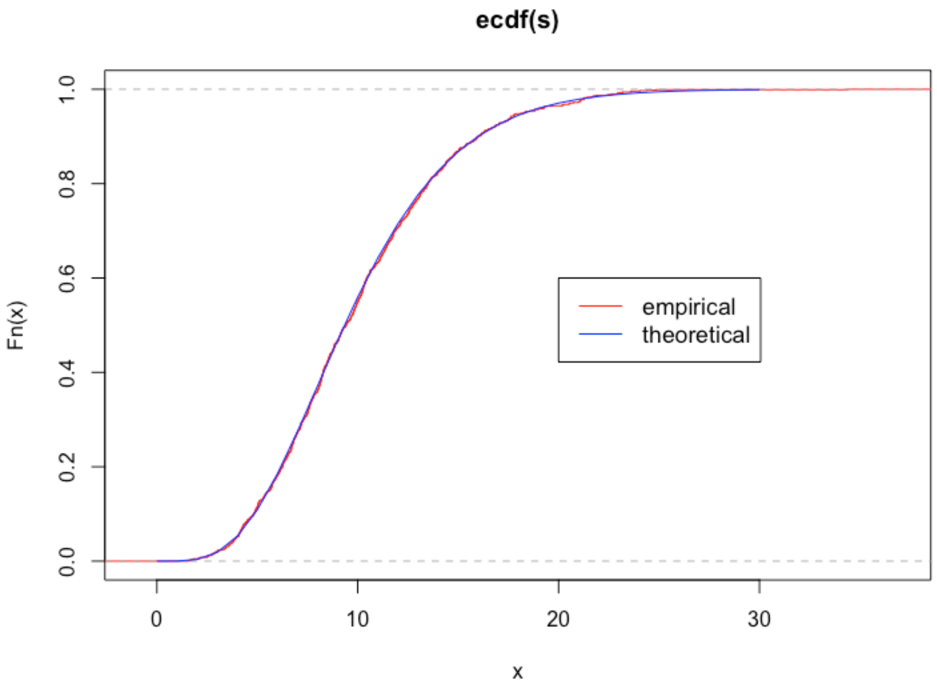
iv) Plot the empirical density histogram of s and compared it to the theoretical density function. Similarly, plot the cumulative distribution function of s and compare it with the theoretical cumulative distribution function. What does the plot suggest?



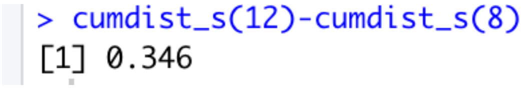
```
> hist_s <- hist(s, breaks=seq(-0.5,35,l=20), freq=F)# we create a density histogram
> lines(seq(-0.5,35, by=1), dchisq(seq(-0.5,35,by=1),n), col="blue")
> legend(20, 0.6, legend=c("empirical","theoretical"), col=c("blue", "grey"), lty=c(1,1), cex=1.1)
> cumdist_s <- ecdf(s)
> cumdist_s
Empirical CDF
Call: ecdf(s)
 x[1:1000] = 1.5594, 1.6248, 1.7087,  ..., 24.837, 34.381
> plot(cumdist_s, col="red")
> lines(seq(0, 30, by=1), pchisq(seq(0,30, by=1),n), col="blue")
> legend(20, 0.6, legend=c("empirical","theoretical"), col=c("red", "blue"), lty=c(1,1), cex=1.1)
```



The empirical data follows the theoretical data closely.



v) Use the empirical cumulative function to calculate $P(8 \le s \le 12)$ from the sample s, compare this to the theoretical probability obtained under point a).

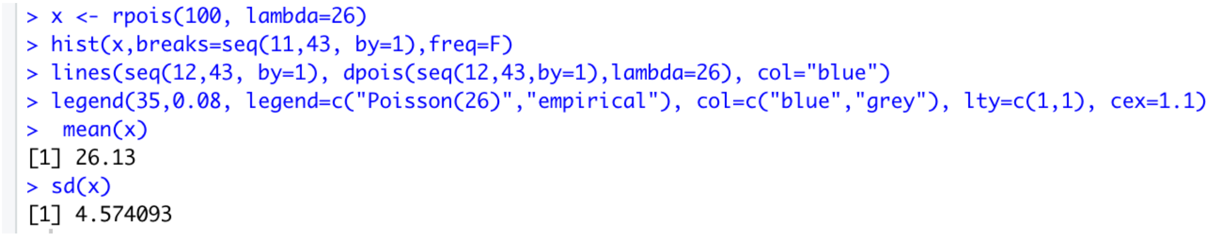


```
> cumdist_s(12)-cumdist_s(8)
[1] 0.346
```

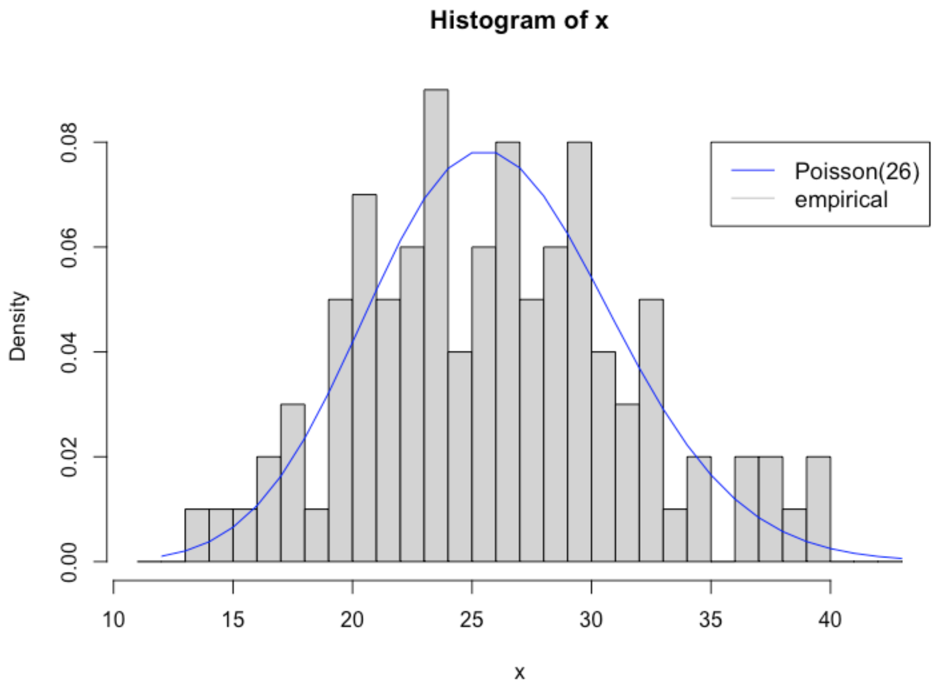
The theoretical value we reached when calculating the empirical cumulative function was 0, now comparing that to our new value, 0.346, and looking at our graphs, we can conclude that our predictions are very close to the values we reached with random sampling.

Question 4

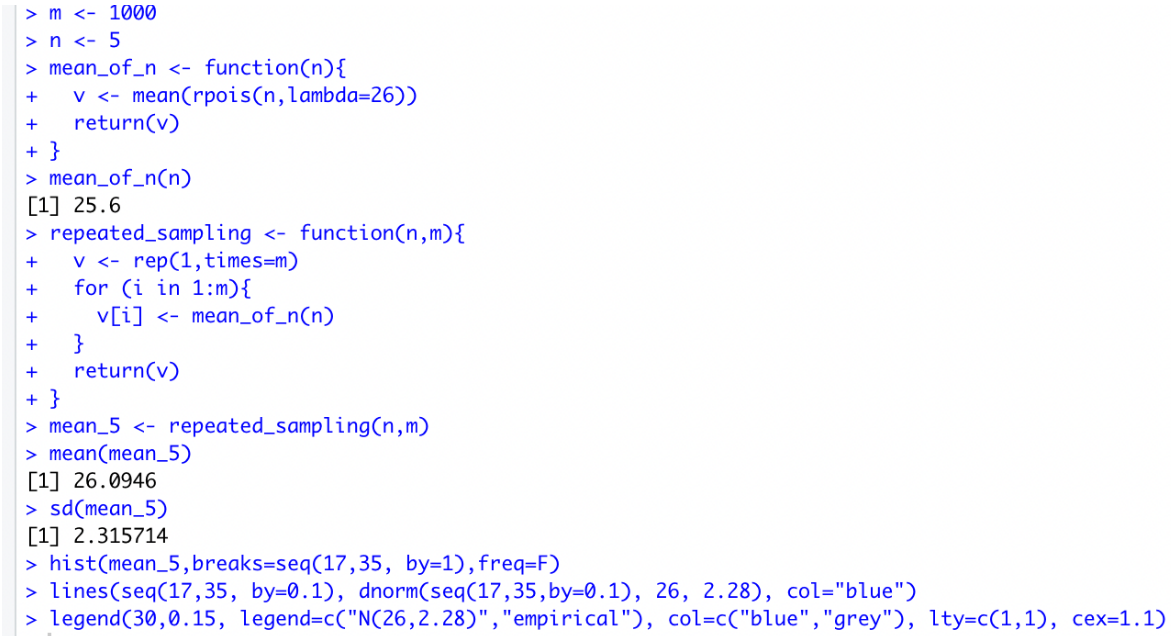
a) Select a distribution of your choice from the family of binomial, Poisson, negative-binomial, exponential or chi-squared distribution, with appropriate parameters of your choice. This should be the original distribution that you are sampling from. Use R to present an empirical and theoretical plot of this distribution, providing the mean and the standard deviation of the chosen distribution.



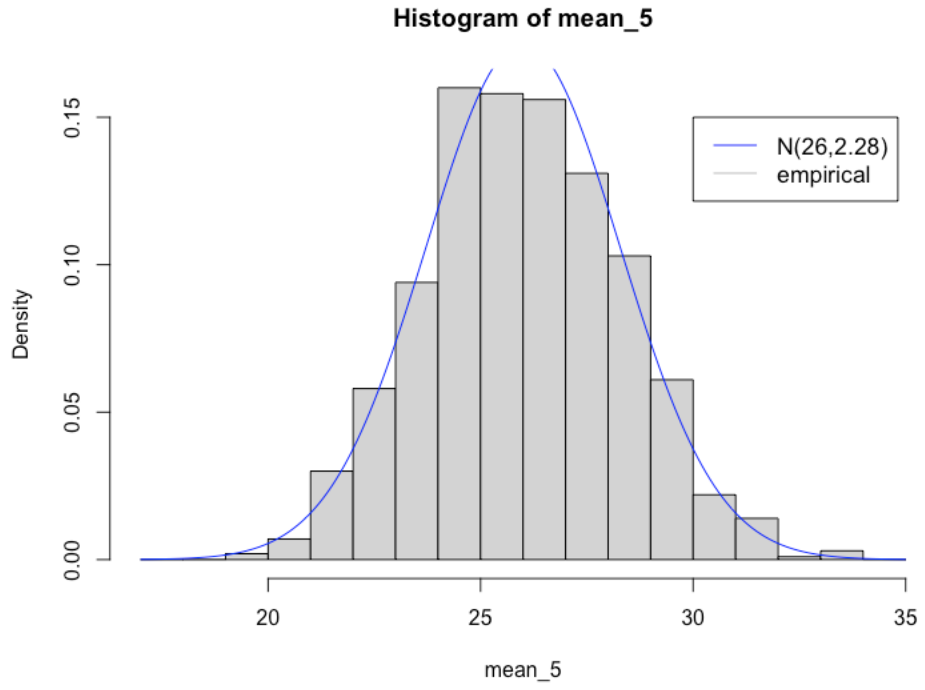
```
> x <- rpois(100, lambda=26)
> hist(x,breaks=seq(11,43, by=1),freq=F)
> lines(seq(12,43, by=1), dpois(seq(12,43,by=1),lambda=26), col="blue")
> legend(35,0.08, legend=c("Poisson(26)","empirical"), col=c("blue","grey"), lty=c(1,1), cex=1.1)
>  mean(x)
[1] 26.13
> sd(x)
[1] 4.574093
```



b) Sample from the original distributions m=1000 samples of size n=5, then calculate for each of them, the corresponding sample mean and store the results in a vector mean_5 of size m=1000. Calculate the mean and the standard deviation of the sample mean and compare them to the mean and standard deviation of the original distribution. Plot the relative frequency histogram of the empirical distribution and compare it to the associated normal distribution as informed by the central limit theorem.

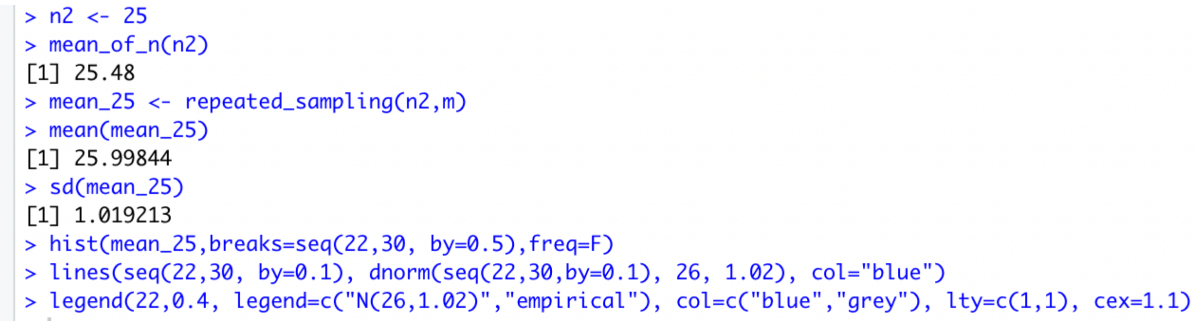


```
> m <- 1000
> n <- 5
> mean_of_n <- function(n){
+   v <- mean(rpois(n,lambda=26))
+   return(v)
+ }
> mean_of_n(n)
[1] 25.6
> repeated_sampling <- function(n,m){
+   v <- rep(1,times=m)
+   for (i in 1:m){
+     v[i] <- mean_of_n(n)
+   }
+   return(v)
+ }
> mean_5 <- repeated_sampling(n,m)
> mean(mean_5)
[1] 26.0946
> sd(mean_5)
[1] 2.315714
> hist(mean_5,breaks=seq(17,35, by=1),freq=F)
> lines(seq(17,35, by=0.1), dnorm(seq(17,35,by=0.1), 26, 2.28), col="blue")
> legend(30,0.15, legend=c("N(26,2.28)","empirical"), col=c("blue","grey"), lty=c(1,1), cex=1.1)
```

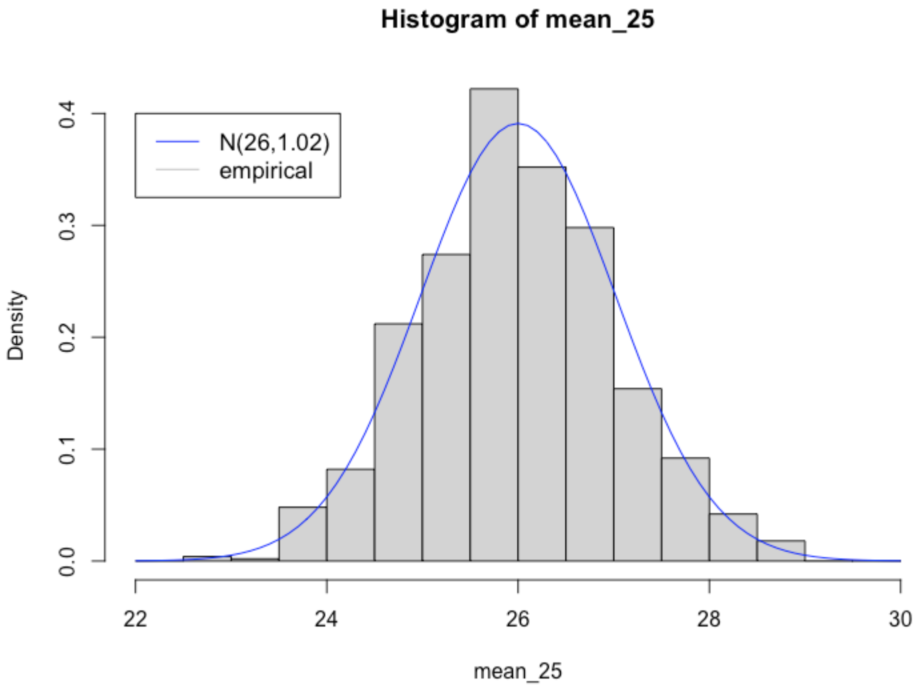


I calculated the mean and standard deviation of our normal distribution using the formula: μ=lambda, sd=√(lambda÷n), so it is reproducible and unbiased.

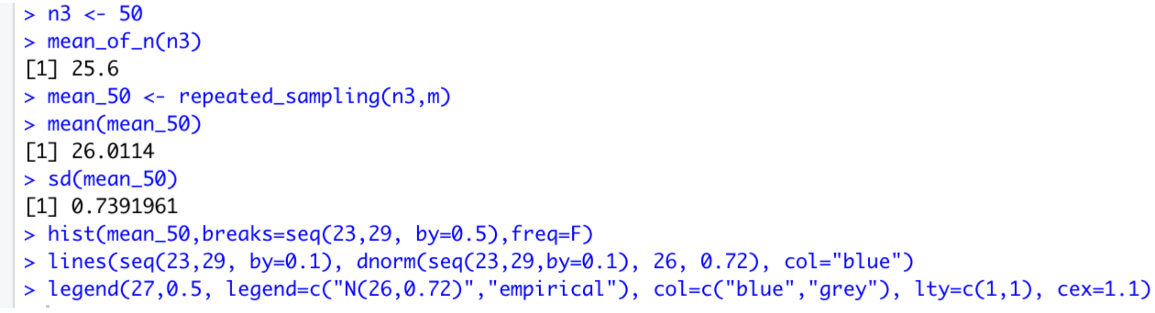
c) Repeat step b) by considering this time n=25, and provide the same outputs as under point b).



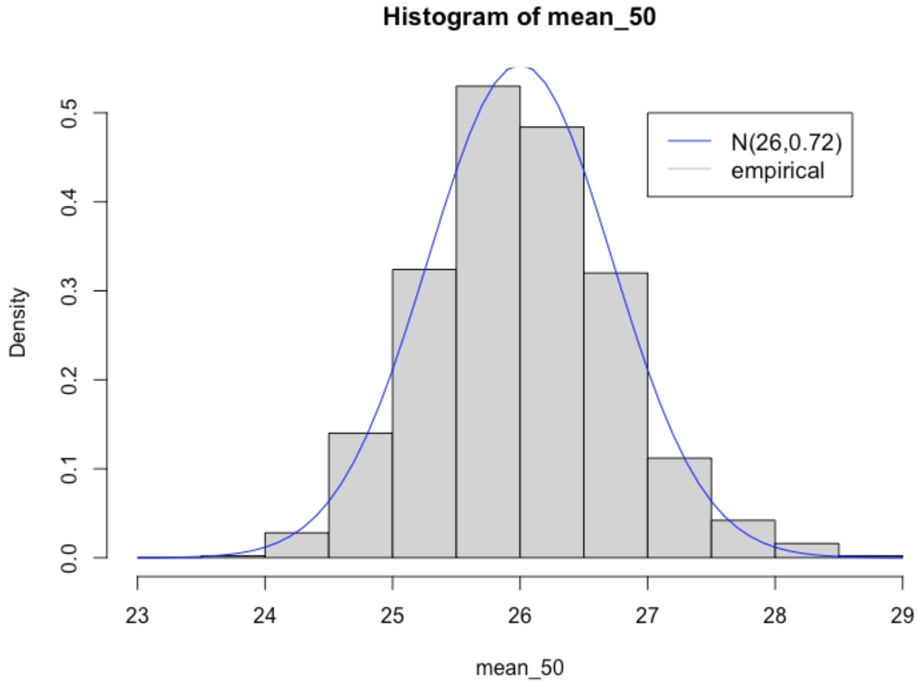
```
> n2 <- 25
> mean_of_n(n2)
[1] 25.48
> mean_25 <- repeated_sampling(n2,m)
> mean(mean_25)
[1] 25.99844
> sd(mean_25)
[1] 1.019213
> hist(mean_25,breaks=seq(22,30, by=0.5),freq=F)
> lines(seq(22,30, by=0.1), dnorm(seq(22,30,by=0.1), 26, 1.02), col="blue")
> legend(22,0.4, legend=c("N(26,1.02)","empirical"), col=c("blue","grey"), lty=c(1,1), cex=1.1)
```



d) Repeat step b) by considering this time n=50, and provide the same outputs as under point b).



```
> n3 <- 50
> mean_of_n(n3)
[1] 25.6
> mean_50 <- repeated_sampling(n3,m)
> mean(mean_50)
[1] 26.0114
> sd(mean_50)
[1] 0.7391961
> hist(mean_50,breaks=seq(23,29, by=0.5),freq=F)
> lines(seq(23,29, by=0.1), dnorm(seq(23,29,by=0.1), 26, 0.72), col="blue")
> legend(27,0.5, legend=c("N(26,0.72)","empirical"), col=c("blue","grey"), lty=c(1,1), cex=1.1)
```



e) Summarise the results under points a) to d) and formulate the conclusion.

In point a) where we were sampling randomly from the Poisson distribution with a chosen sample size 100, our mean turned out to be close to the theoretical mean of the Poisson distribution (which is always lambda) and we obtained the number 26.09. In our first graph, we can see that without considering the means of our different repeated distributions our values do not represent the normal distribution with mean µ and standard deviation σ well. However, our findings support the Central Limit Theorem which states that when we increase the sample size (which we did through exercises b-d), the spread of the means of our repeted sampling converges to a normal distribution, which we have represented visually through graphs.