Econ 403A: Homework 2

Pujan Thakrar

October 16, 2018

Question 1:

An article in The Engineer ("Redesign for Suspect Wiring," June 1990) reported the results of an investigation into wiring errors on commercial transport aircraft that may produce faulty information to the flight crew. Such a wiring error may have been responsible for the crash of a British Midland Airways aircraft in January 1989 by causing the pilot to shut down the wrong engine. Of 1600 randomly selected aircraft, eight were found to have wiring errors that could display incorrect information to the flight crew.

(a) Find a 99% confidence interval on the proportion of aircraft that have such wiring errors.

We use the following formula:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

We get:

$$0.005 \pm 2.58\sqrt{\frac{0.005(1-0.005)}{1600}}$$

which gives us the 99% confidence interval:

$$0.00045 \leq p \leq 0.00955$$

(b) Suppose we use the information in this example to provide a preliminary estimate of p. How large a sample would be required to produce an estimate of p that we are 99% confident differs from the true value by at most 0.008?

We use the following formula:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If we rearrange, we get:

$$n = (\frac{z_{\alpha/2}}{E})^2\hat{p}(1-\hat{p})$$

Substituting the values that we have we get:

$$n = (\frac{2.58}{0.008})^2(0.005)(1-0.005)$$

Solving for this, we get that a sample size of 518 would be required in order to produce an estimate of p that we are 99% confident differs from the true value by at most 0.008.

(c) Suppose we did not have a preliminary estimate of p. How large a sample would be required if we wanted to be at least 99% confident that the sample proportion differs from the true proportion by at most 0.008 regardless of the true value of p?

In this case we assume the worst case scenario, which would be a $\hat{p} = \frac{1}{2}$.

We substitute this in the equation we used in part b and get the following:

$$n = (\frac{2.58}{0.008})^2(0.5)(0.5)$$

In this case the sample size we would need is approximately 26002.

(d) Comment on the usefulness of preliminary information in computing the needed sample size.

The sample size was significantly larger in part c. This shows that it is really important to have preliminary information as it can save a lot of time, money and resources in real life.

Question 2

The proportion of residents in Phoenix favoring the building of toll roads to complete the freeway system is believed to be p = 0.3. If a random sample of 10 residents shows that 1 or fewer favor this proposal, we will conclude that p < 0.3.

(a) Find the probability of type I error if the true proportion is p = 0.3.

We have a binomial distribution so we use the following equation to solve for the probability of a type 1 error (probability that we reject the null if it is true):

$$\begin{aligned} \alpha &= P(X \leq 1|p = 0.3) \\ &= P(X = 0|p = 0.3) + P(X = 1|p = 0.3) \\ &= \binom{10}{0} 0.3^0(0.7)^{10} + \binom{10}{1} 0.3^1(0.7)^9 \\ &= 0.1493 \end{aligned}$$

Therefore, the probability of a type one error is approximately 0.1493.

(b) Find the probability of committing a typeII error with this procedure if p = 0.2.

This is similar to what we did in part a, although this time we use the following equation:

$$\begin{aligned} \beta &= P(X \geq 1|p = 0.2) \\ &= 1 - P(X \leq 1|p = 0.2) \\ &= 1 - [P(X = 0|p = 0.2) + P(X = 1|p = 0.2)] \\ &= 1 - [\binom{10}{0} 0.2^0(0.8)^{10} + \binom{10}{1} 0.2^1(0.8)^9] \\ &= 0.6242 \end{aligned}$$

Therefore the probability of a type two error is approximately 0.6242 given $p = 0.2$.

(c) What is the power of this procedure if the true proportion is p = 0.2?

The power of the test is given by the equation: $1 - \beta$

In this case:

$$1 - \beta = 1 - 0.6242 = 0.3758$$

Question 3

Assume that X1, . . . , X9 are i.i.d. having Bernoulli distribution with parameter p. Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0 &: p = 0.4 \\ H_1 &: p \neq 0.4 \end{aligned}$$

Let $Y = \sum_{i=1}^{9} X_i$

(a) Find c_1 and c_2 such that $P(Y \leq c_1|p = 0.4) + P(Y \geq c_2|p = 0.4)$ is as close to 0.1 without being larger than 0.1.

In order to do this we need to test out several values.

Trying $c_1 \geq 2$ we get $P(Y \leq c_1|p = 0.4) \geq 0.23$ and therefore $c_1 \leq 1$.

Following this logic, we also test for $c_2 \leq 5$. This gives us $P(Y \geq c_2|p = 0.4) \geq 0.26$. Therefore, $c_2 \geq 6$. I used the following code to test for several values of both c_1 and c_2.

```
c1 =1
c2 = 7
p=0.4
sum1=0
sum2 =0

for(k in 0:c1){                                    #P(Y<= c_1|p=0.4)
  a=choose(9,k)*(0.4^k)*(0.6^(9-k))
  sum1=sum1+a
}
for(l in 0:c2-1){                                  #P(Y>= c_2|p=0.4)
    b=choose(9,l)*(0.4^l)*(0.6^(9-l))
    sum2=sum2+b
}

x=sum1+1-sum2                          #P(Y<= c_1|p=0.4)+P(Y>= c_2|p=0.4)
```

Here is a table with some of the values I find:

c1	c2	P
1	6	0.1699
0	6	0.1094
1	7	0.0956

In the end, the values 1 and 7 for c_1 and c_2 respectively yield the closest probability to 0.1 without being larger.

Note: I did not attempt negative numbers for c_1 and c_2

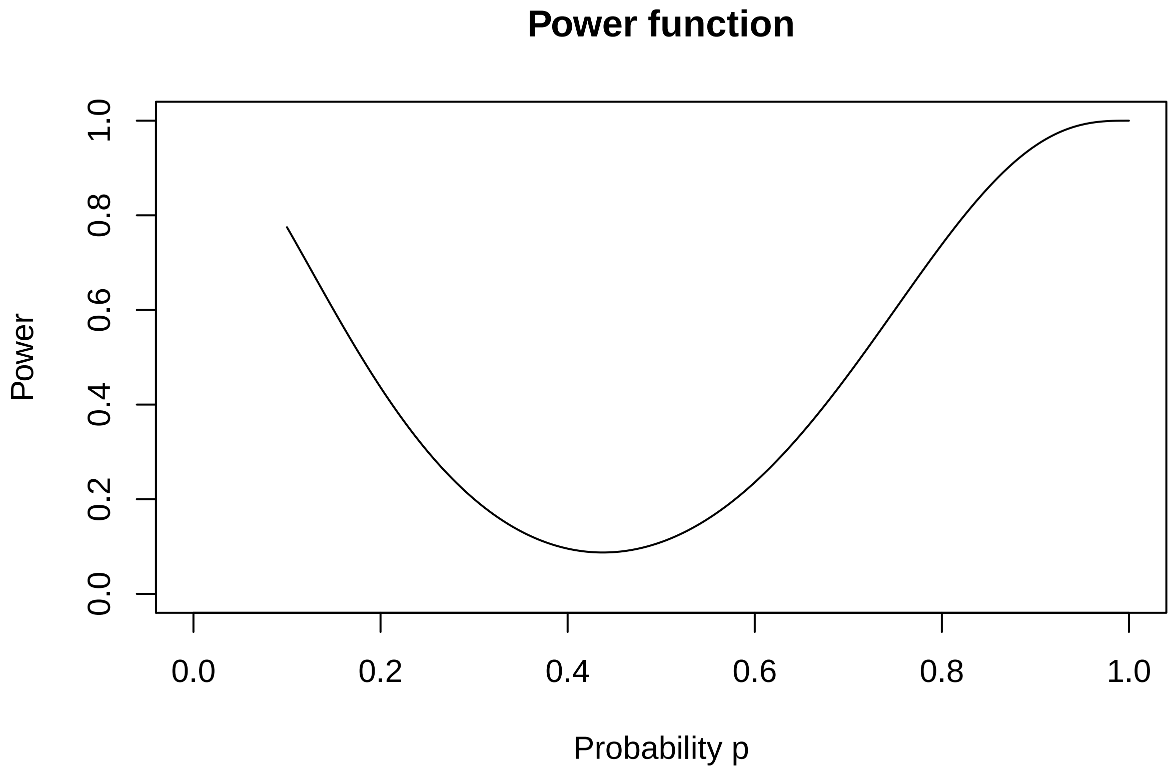
(b) Let δ be the test that rejects H_0 if either $Y \leq c_1$ or $Y \geq c_2$. What is the size of the test δ_c?

From the table in part a, you can see that the calculated P is equal to the size of the test. Therefore, the size of the test is 0.0956.

(c) Draw a graph of the power function δ_c.

To do this: I use the following code:

```
c1 =1                                         #test for values of c1 and c2
c2 = 7
del=c()                                             #vector to store deltas
  s = seq(0.1,1,by=0.001)
  c = 1

  for(p in s){               #saame loop as before except changing value of p
    sum1=0
    sum2 =0

    for(k in 0:c1){
      a=choose(9,k)*(p^k)*((1-p)^(9-k))
      sum1=sum1+a
    }
    for(l in 0:c2-1){
      b=choose(9,l)*(p^l)*((1-p)^(9-l))
      sum2=sum2+b
    }

    del[c]=sum1+1-sum2
    c=c+1

  }

plot(s,del,type="l",xlim=c(0:1),ylim=c(0:1),xlab="Probability p",ylab="Power",
     main="Power function")
```



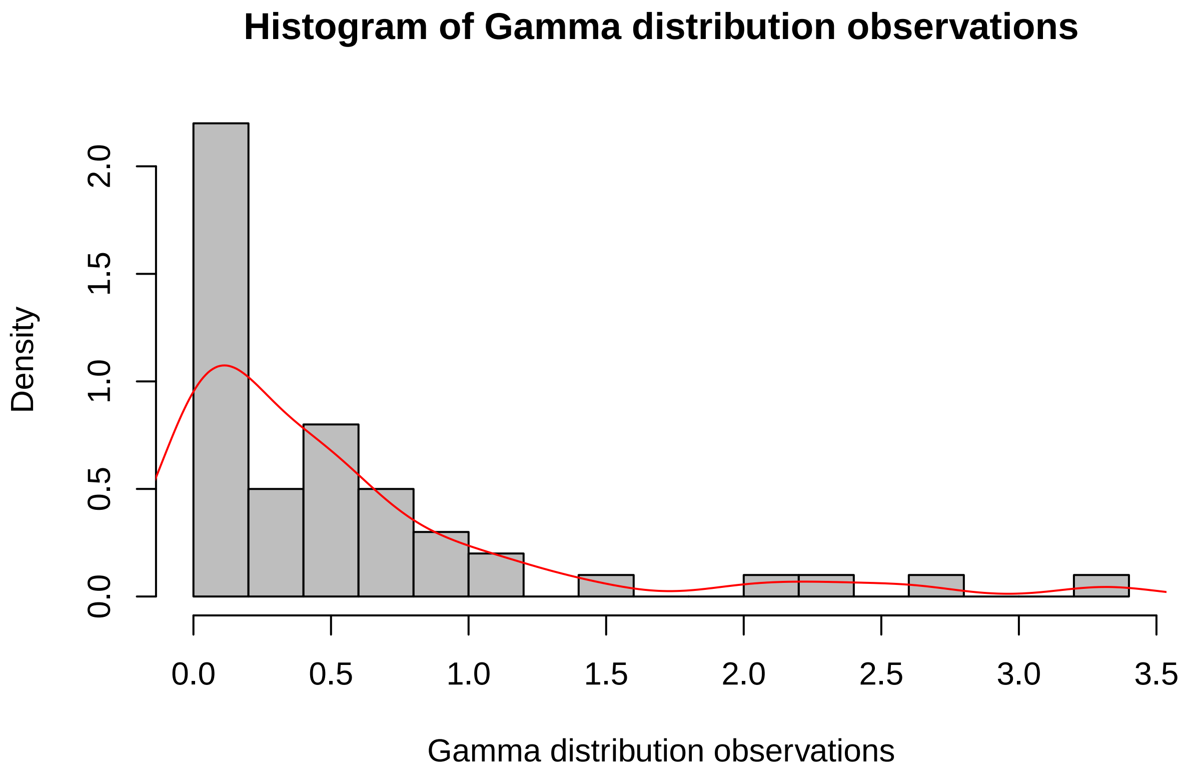
Question 4

The file ‘Prob4_data.txt’ contains 50 observations from a Gamma distribution with unknown parameters α and β.

(a) Plot a histogram of the data and overlay the respective density curve.

```
data<-read.table("Prob4_data.txt")
data2<-data.frame(data)
data3<-data.matrix(data2,rownames.force = NA)
hist(data3,20, prob=TRUE,col="grey", xlab="Gamma distribution observations",
     main="Histogram of Gamma distribution observations")
```

```
lines(density(data3),col="red")
```



(b) Compute the Method of Moments estimates of the parameters, $\hat{\alpha}$ and $\hat{\beta}$.

```
mean(data3)                                                   #find the mean
```

```
## [1] 0.53662
```

```
var(data3)                                                #find the variance
```

```
##          V1
## V1 0.506144
```

```
beta=var(data3)/mean(data3)                                  #solve for beta
alpha = mean(data3)*beta                                    #solve for alpha

summation =0                                   #calculate the summation term
for(i in 1:50){
  x=(data3[i])^2-(mean(data3))^2
  summation=summation+x
}




alpha_hat = (mean(data3))^2/((1/50)*summation)#using MoM solve for params
beta_hat =mean(data3)/((1/50)*summation)

alpha_hat
```

```
## [1] 0.5805419
```

```
beta_hat
```

```
## [1] 1.081849
```

(c) Generate 1000 new samples from your data and compute the Bootstrap Mean, standard errors and 95% confidence intervals of the parameters and compare them against your results from part (b).

```
alphas =0  #create empty vectors to store alphas an betas from each sample
betas = 0

for (j in 1:1000){  #loop will get samples and store parameters in vectors
  samp =sample(data3,50, replace=TRUE)
  for(i in 1:50){
    x=(samp[i])^2-(mean(samp))^2
    summation=summation+x
  }

  alpha_hat = (mean(samp))^2/((1/50)*summation)
  beta_hat =mean(samp)/((1/50)*summation)

  alphas[j]=alpha_hat
  betas[j]=beta_hat

}

#bootstrapped mean
boot_mean_alpha =mean(alphas)
boot_mean_beta=mean(betas)
boot_mean_alpha
```

```
## [1] 0.003877236
```

```
boot_mean_beta
```

```
## [1] 0.007127329
```

```
#bootstrapped standard error
boot_se_alpha = sqrt(var(alphas))
boot_se_beta = sqrt(var(betas))
boot_se_alpha
```

```
## [1] 0.01541408
```

```
boot_se_beta
```

```
## [1] 0.02743868
```

```
#bootstrapped confidence interval

tstat = abs(qt(0.025,49))

boot_ci_alpha=c(boot_mean_alpha-tstat*boot_se_alpha,boot_mean_alpha+tstat*boot_se_alpha)
boot_ci_beta=c(boot_mean_beta-tstat*boot_se_beta,boot_mean_beta+tstat*boot_se_beta)

boot_ci_alpha
```

```
## [1] -0.02709852  0.03485300
```

```
boot_ci_beta
```

```
## [1] -0.04801275  0.06226741
```

My answers differ quite a bit from our results in part b. Generating the 1000 samples probably gives a more accurate result.

Question 5

A 1992 article in the Journal of the *American Medical Association* ("A Critical Appraisal of the 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich") reported body temperature, gender, and heart rate for a number of subjects. The body temperatures for the 25 female subjects. The body temperatures for 25 female subjects follow:97.8, 97.2, 97.4, 97.6, 97.8, 97.9, 98.0, 98.0, 98.0, 98.1, 98.2, 98.3, 98.3, 98.4, 98.4, 98.4, 98.5, 98.6, 98.6, 98.7, 98.8, 98.8, 98.9, 98.9, and 99.0.

(a) Test the hypothesis $H_0 : \mu = 98.6$ versus $H_1 : \mu \neq 98.6$ using $\alpha = 0.05$. Find the p-value.

Using R, I calculated the mean and standard deviation of the data:

mean = 98.264 standard deviation = 0.4820788.

Using the following t-stat formula:

$$t_c = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

I get $t_c = -3.484907$.

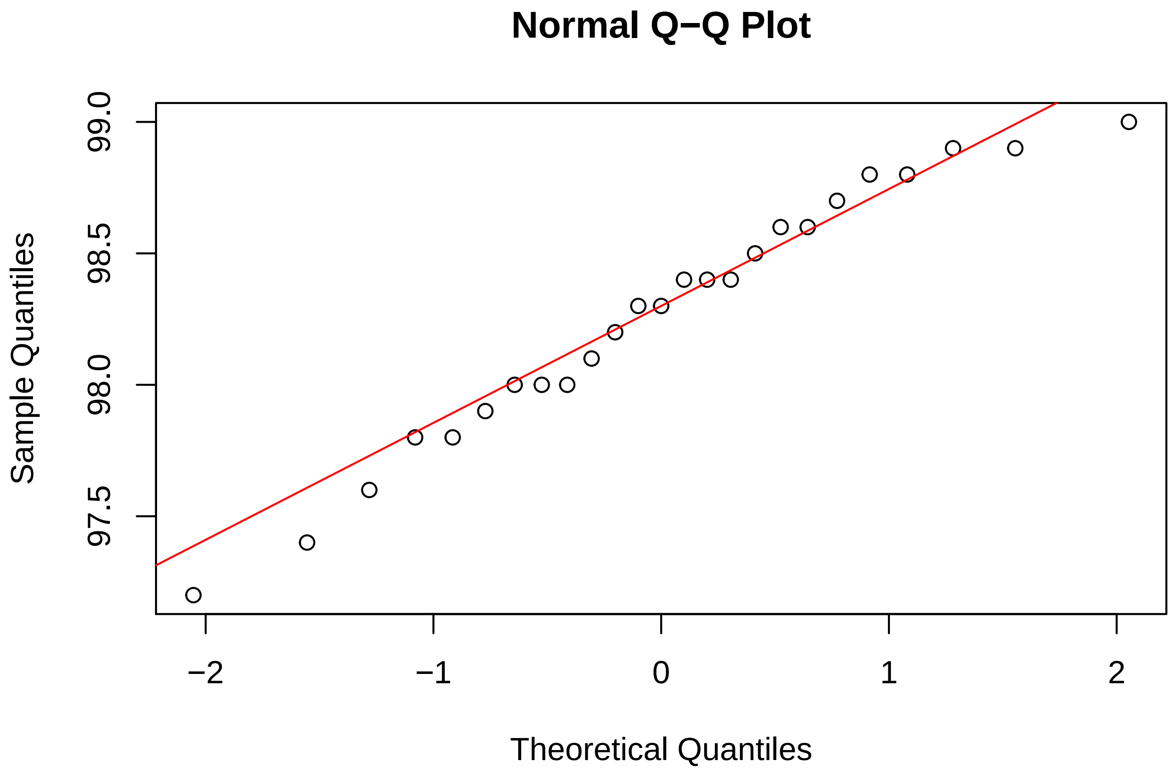
Since $|t_c| > t_{24,0.025}$ we reject the null at 5%. Given the t-statistic, I looked into the t-table and the $0.001 < p - value < 0.002$.

(b) Check the assumption that female body temperature is normally distributed.

In order to check if the data was normally distributed, I did the following checks.

First, I plotted the data in a Normal Q-Q Plot.

```
temps<-c( 97.8, 97.2, 97.4, 97.6, 97.8, 97.9, 98.0, 98.0, 98.0, 98.1, 98.2, 98.3, 98.3,
          98.4, 98.4, 98.4, 98.5, 98.6, 98.6, 98.7,98.8, 98.8, 98.9, 98.9, 99.0)

qqnorm(temps)
qqline(temps,col="red")
```



From this we can see that the data is very close to the line, meaning that it is probably normally distributed,

Then, I did a Shapiro-Wilk normality test:

```
shapiro.test(temps)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  temps
## W = 0.96746, p-value = 0.5817
```

This gives us a p-value $>$ 0.05, which means we fail to reject the null hypothesis that the population is normally distributed.

And as a final check. I compare the mean and median of the data:

```
mean(temps)
```

```
## [1] 98.264
```

```
median(temps)
```

```
## [1] 98.3
```

As you can see, the mean and median are very close which is a characteristic of a normally distributed dataset.

Therefore I can conclude that the data is normally distributed.

(c) Compute the power of the test if the true mean femal body temperature is as low as 98.0.

```
n = length(temps)
s = sd(temps)
SE = s/sqrt(n)

alpha=0.05
mu0=98
I=c(alpha/2,1-alpha/2)
q=mu0+qt(I,df=n-1)*SE
q
```

```
## [1] 97.80101 98.19899
```

```
mu=98.6
p=pt((q-mu)/SE,df=n-1)
p
```

```
## [1] 8.409417e-09 1.760860e-04
```

```
x=diff(p)                      #this is the probability of a type II error

1-x                                         #this is the power of the test
```

```
## [1] 0.9998239
```

Therefore the power of the test approximately 1.

(d) What sample size would be required to detect a true mean female body temperature as low as 98.2 if we wanted the power of the test to be at least 0.9?

```
n_new = ((1.96+1.28)^2)*var(temps)/(0.4)^2
n_new
```

```
## [1] 15.24776
```

Therefore, a sample size of 16 would be required to detect a true mean female body temperature as low as 98.2.

(e) Explain how question in part a could be answered by constructing a two-sided confidence interval on the mean female body temperature.

If we construct the following confidence interval:

```
alpha=0.05
mu0=98.264
I=c(alpha/2,1-alpha/2)
q=mu0+qt(I,df=n-1)*SE
q
```

```
## [1] 98.06501 98.46299
```

We can see that 98.6 does not fall within the acceptance range and therefore we reject the null.

Question 6

Suppose a sample size of 1 is taken from the pdf $f_Y(y) = (1/\lambda)e^{-y/\lambda}, y > 0$, for the purpose of testing

$$H_0 : \lambda = 1$$

$$H_1 : \lambda > 1$$

The null hypothesis will be rejected if $y \geq 3.20$.

(a) Calculate the probability of commiting a Type I error.

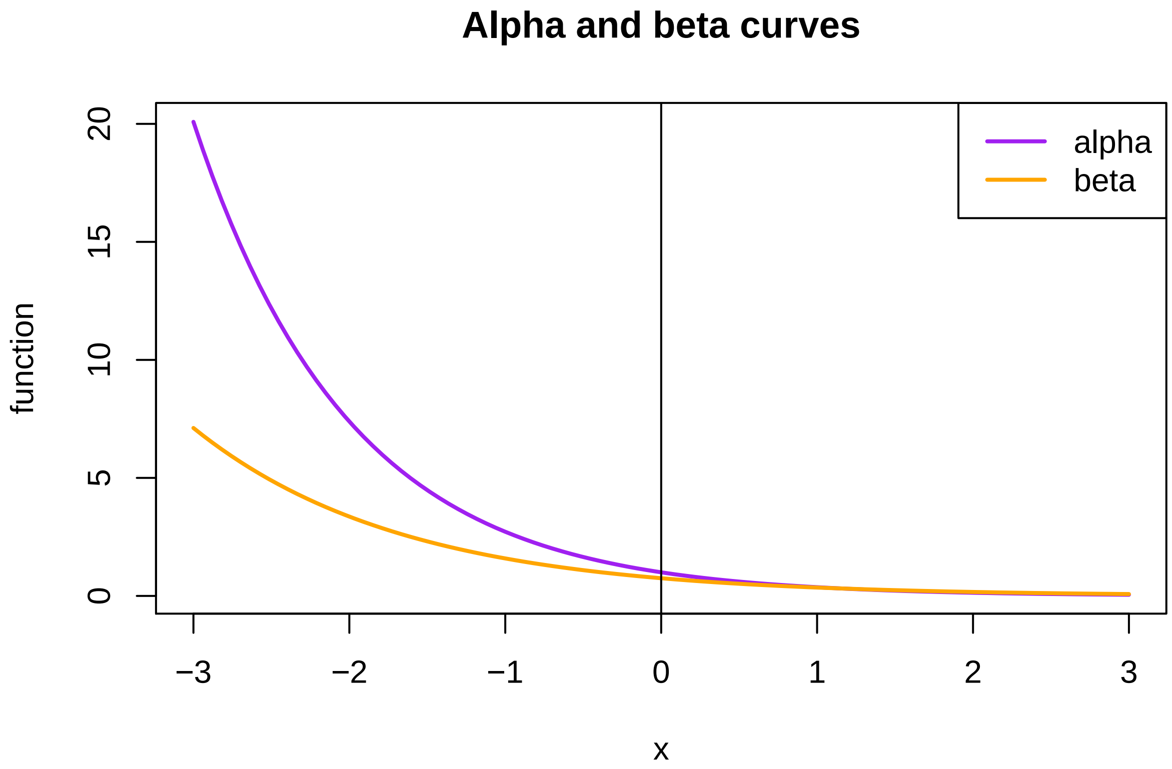
$$
\begin{aligned}
\alpha &= P(y \geq 3.20|\lambda = 1) \\
&= \int_{3.2}^{\infty} (\frac{1}{1}) e^{\frac{-y}{1}} dy \\
&= \int_{3.2}^{\infty} e^{-y} dy \\
\alpha &= 0.0407622
\end{aligned}
$$

(b) Calculate the probability of commiting a Type II error when $\lambda = 4/3$.

$$
\begin{aligned}
\beta &= P(y \leq 3.20|\lambda = \frac{4}{3}) \\
&= \int_{0}^{3.2} (\frac{3}{4}) e^{\frac{-3y}{4}} dy \\
&= \frac{3}{4} \int_{0}^{3.2} e^{\frac{-3y}{4}} dy \\
\beta &= 0.0909282
\end{aligned}
$$

(c) Draw a diagram that shows the α and β calculated in parts (a) and (b).

```r
funct <- function(x,lambda) {
  (1/lambda)*exp(-x/lambda)
}


curve(funct(x,1), from = -3, to = 3, n = 100, col = "purple", lwd = 2,
      ylab="function",main="Alpha and beta curves")
curve(funct(x,4/3), from = -3, to = 3, n = 100, col = "orange", lwd = 2, add = TRUE)
abline(v = 0)
legend("topright", legend = c("alpha", "beta"), lty = c(1,1),
       lwd = c(2,2), col = c("purple", "orange"))
```



Question 7

Suppose we have a population of 10,000 elements, each with a unique label from the set {1,2,3,. . . ,10000}.

(a) Generate a sample of 500 labels from this population using simple random sampling.

```
labels<-c(1:10000)
simple=sample(labels, 500, replace=FALSE)
```

(b) Generate a sample of 500 labels from this population using i.i.d. sampling

```
iid=sample(labels, 500, replace=TRUE)
```

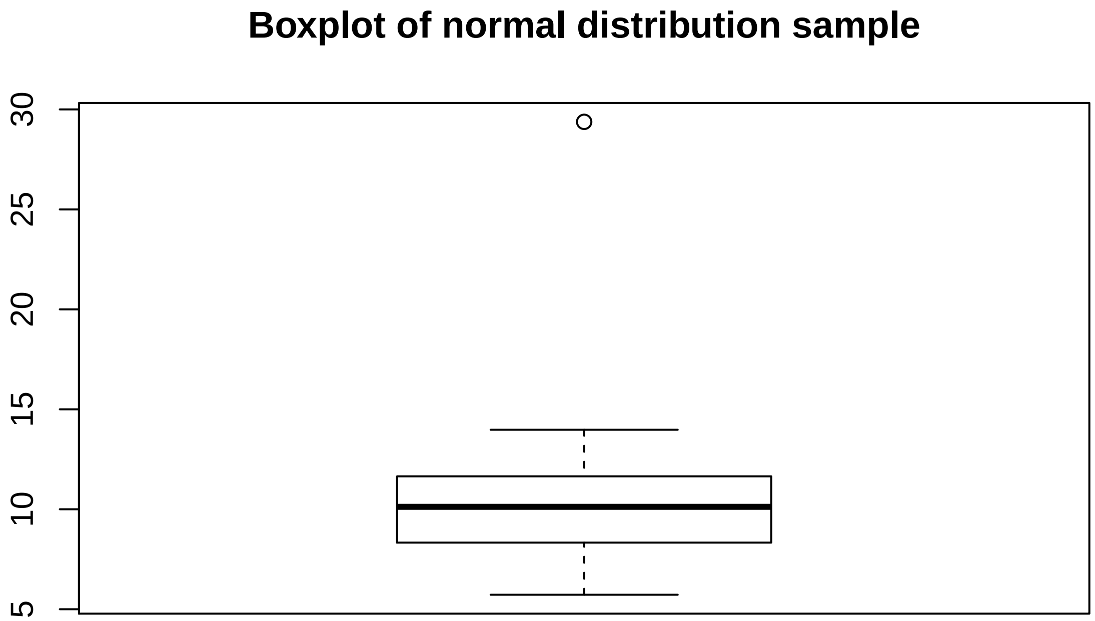
Question 8

Generate a sample of 30 from an N(10,2) distribution and a sample of 1 from an N(30,2) distribution. Combine these together to make a single sample of 31.

```
norm1<-rnorm(30,mean=10, sd=2)
norm2 <-rnorm(1,mean=30, sd=2)
norm_samp =c(norm1,norm2)
```

(a) Produce a boxplot of these data

```
boxplot(norm_samp,main="Boxplot of normal distribution sample")
```



(b) What do you notice about this plot?

I notice that the single point that was taken from the second normal distribution is represented as an outlier.

(c) Based on the boxplot, what characteristics do you think would be appropriate to measure the location and spread of the distribution? Explain why.

I think that the interquartile range would be the most appropriate measure of the location and spread of the data since it is unaffected by the outlier and therefore will give a more accurate representation of where most of the data lies.

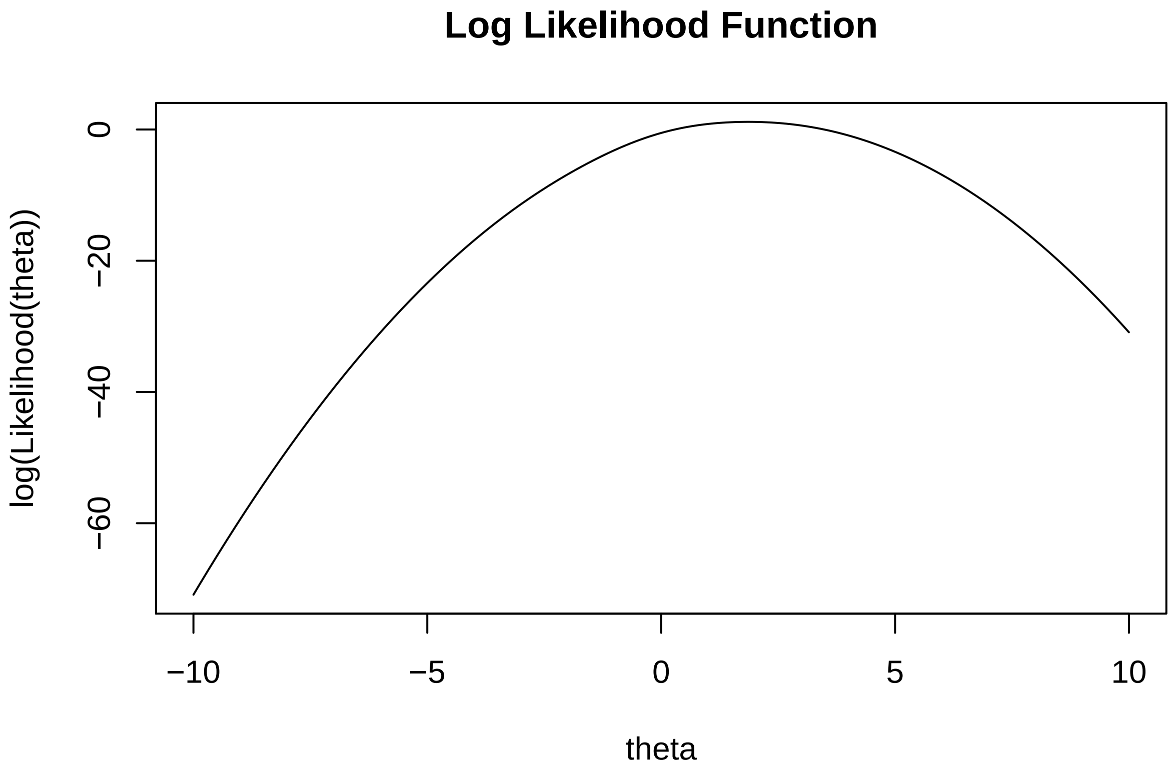
Question 9

A likelihood function is given by $\exp(-(\theta-1)^2/2)+3\exp(-(\theta-2)^2/2)$ for $\theta \in R^1$. Numerically approximate the MLE by evaluating this function at 1000 equispacd points in (-10,10]. Also plot the likelihood function.

```
Likelihood<-function(theta){
  return((exp(-(theta-1)^2)/2)+3*exp(-((theta-2)^2)/2))
}

theta<-seq(-10,10,length=1000)

MLE = theta[which(Likelihood(theta)==max(Likelihood(theta)))]
MLE
```

```
## [1] 1.871872
```

```
plot(theta, log(Likelihood(theta)), type='l', main="Log Likelihood Function")
```

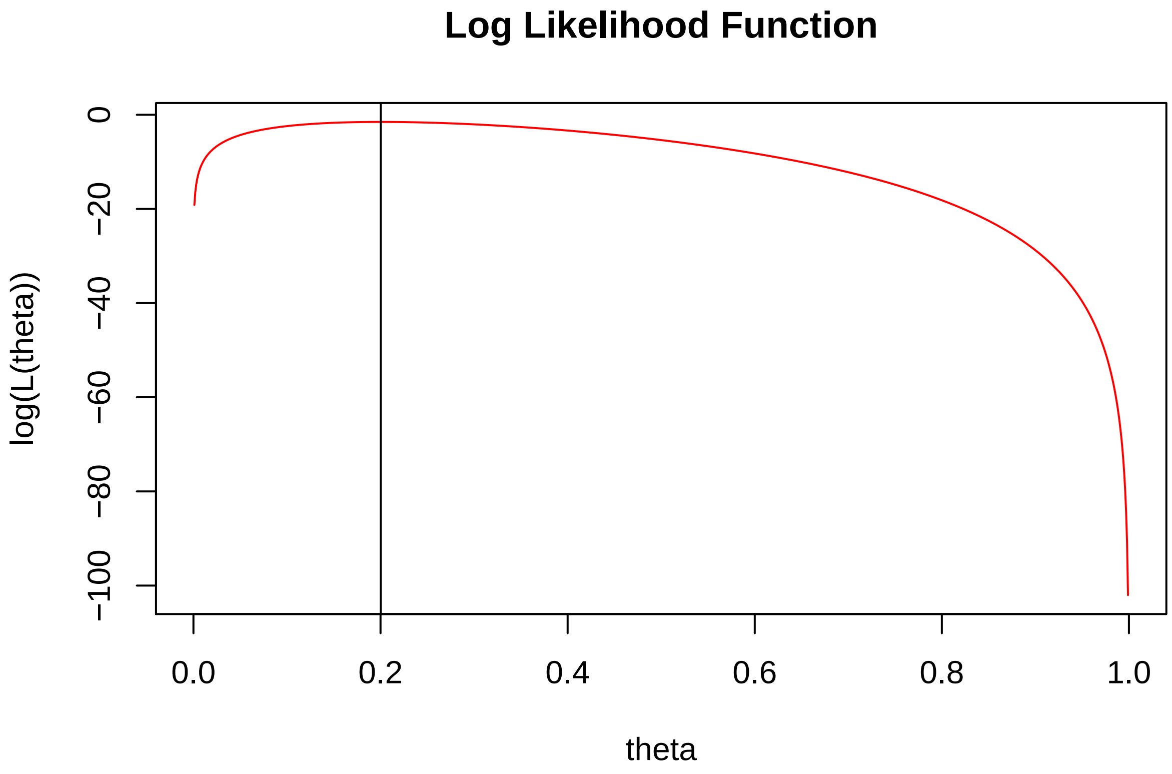


Question 10

Suppose the proportion of left-handed individuals in a population i θ. Based on a simple random sample of 20, you observe four left-handed individuals.

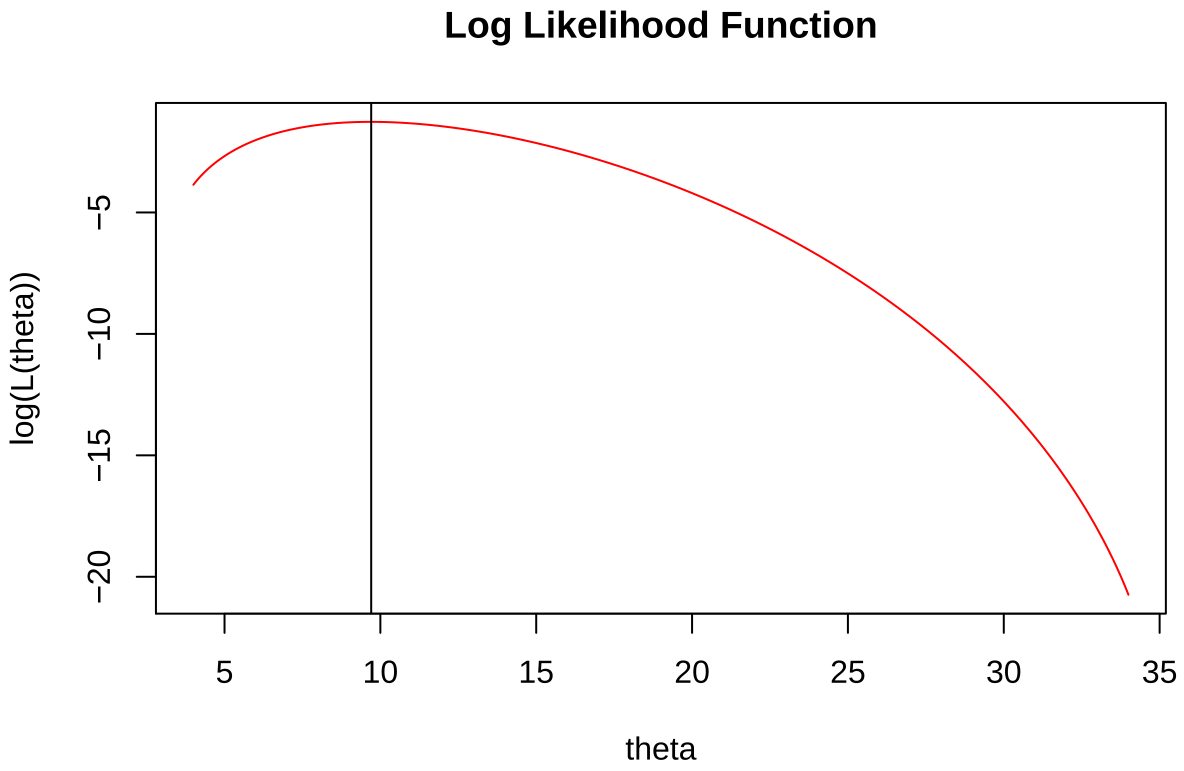
(a) Assuming the sample size is small relative to the population size, plot the log-likelihood function and determine the MLE.

```r
L<-function(theta){
  return((choose(20,4))*(theta^4)*((1-theta)^16))
}

theta<-seq(0,1,length=1000)

plot(theta, log(L(theta)),type='l', main="Log Likelihood Function", col="red")
m=theta[which(L(theta)==max(L(theta)))]
abline(v=m)
```



(b) If instead the population size is only 50, then plot the log-likelihood function and determine the MLE. (Hint: Remember that the number of left-handed individuals follows a hypergeometric distribution. This forces θ to be of the form i/50 for some integer i between 4 and 34. From a tabulation of the log-likelihood, you can obtain the MLE.)

```
L<-function(theta){
 return(choose(theta,4)*choose(50-theta,16)/choose(50,20))

}

theta<-seq(4,34,length=1000)

plot(theta, log(L(theta)),type='l', main="Log Likelihood Function", col="red")
m=theta[which(L(theta)==max(L(theta)))]
abline(v=m)
```



Question 11

Generate 10^4 samples of size n=5 from the N(0,1) distribution. For each of these samples, calculate the interval $(\bar{x} - s/\sqrt{5}, \bar{x} + s/\sqrt{5})$, where s is the sample standard deviation, and compute the proportion of times the interval contains μ. Repeat this simulation with n=10 and 100 and compare your results.

```
p=0
n=5
for (i in 1:10^4){
  samp <- rnorm(n, mean=0, sd=1)
  x<-mean(samp)
  s<-sd(samp)

  if(0>=(x-(s/sqrt(n)))&0<=(x+(s/sqrt(n)))){
    p=p+1
  }
}
proportion=p/10^4

proportion
```

```
## [1] 0.6278
```

```
p=0
n=10
for (i in 1:10^4){
  samp <- rnorm(n, mean=0, sd=1)
  x<-mean(samp)
  s<-sd(samp)

  if(0>=(x-(s/sqrt(n)))&0<=(x+(s/sqrt(n)))){
    p=p+1
  }
}
proportion=p/10^4

proportion
```

```
## [1] 0.6529
```

```
p=0
n=100
for (i in 1:10^4){
  samp <- rnorm(n, mean=0, sd=1)
  x<-mean(samp)
  s<-sd(samp)

  if(0>=(x-(s/sqrt(n)))&0<=(x+(s/sqrt(n)))){
    p=p+1
  }
}
proportion=p/10^4

proportion
```

```
## [1] 0.6742
```

From this we see that all the proportions are fairly close to each other.

Question 12

For the data of Exercise 6.4.1, use the plug-in MLE to estimate the first quartile of an N(μ,σ^2) distribution. Use bootstrapping to estimate the MSE of this estimate for m=10ˆ3 and m=10ˆ4.

```
samp<-c(3.27,-1.24,3.97,2.25,3.47,-0.09,7.45,6.20,3.74,4.12,1.42,2.75,-1.48,4.97,8.00,
        3.26,0.15,-3.64,4.88,4.55)
MLE<-fitdistr(samp,densfun = "normal")
p<-pnorm(q=3,mean=MLE[["estimate"]][["mean"]],sd=MLE[["estimate"]][["sd"]])



pvec=c()
for(i in 1:1000){
  sample<- sample(samp,length(samp),replace=TRUE)
  MLE1<-fitdistr(sample, densfun="normal")
  p1<-pnorm(q=3,mean=MLE1[["estimate"]][["mean"]],sd=MLE1[["estimate"]][["sd"]])
  pvec[i]=p1
}




MSE1 = var(pvec)+(mean(pvec)-mean(samp))^2
MSE1
```

```
## [1] 5.728513
```

```
pvec2=c()
for(i in 1:10000){
  sample<- sample(samp,length(samp),replace=TRUE)
  MLE2<-fitdistr(sample, densfun="normal")
  p2<-pnorm(q=3,mean=MLE2[["estimate"]][["mean"]],sd=MLE2[["estimate"]][["sd"]])
  pvec2[i]=p2
}

MSE2 = var(pvec2)+(mean(pvec2)-mean(samp))^2
MSE2
```

```
## [1] 5.71862
```