

Homework 2

Zehua Wu

10/13/2018

1.

An article in The Engineer (“Redesign of Suspect Wiring”, June 1990) reported the results of an investigation into wiring errors on commercial 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.

Answer:

Since in the sample of 1600 aircrafts, 8 were found to have wiring errors, we have

$$p = \frac{8}{1600} = 0.005$$

By definition, the 99% confidence interval is defined as:

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

Here our *alpha* is equal to 0.1, and that gives us $z_{\alpha/2} \approx 2.576$. Hence,

$$0.005 - 2.576 \times \sqrt{\frac{0.005(1-0.005)}{1600}} \leq p \leq 0.005 + 2.576 \times \sqrt{\frac{0.005(1-0.005)}{1600}}$$

$$0.005 - 0.0045 \leq p \leq 0.005 + 0.0045$$

$$0.00045 \leq p \leq 0.0095$$

Thus the 99% confidence interval on the proportion of aircraft that have such wiring errors is $[0.00045, 0.0095]$.

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?

Answer:

Basically we want the margin of error E to be at most 0.008.

Since $E = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$, we have

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

By substituting we get

$$n = \left(\frac{2.576}{0.008} \right)^2 (0.005)(1 - 0.005) \approx 516$$

So the sample size should be 516.

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 ?

Answer:

Since we know that $\left(\frac{z_{\alpha/2}}{E} \right)^2$ will always remain the same, our n really depends solely on $p(1-p)$. Further, we can say that when our n satisfies the largest E , it will satisfy every other E . Thus, we can find the n when $p(1-p)$ is at its maximum, and that gives us:

$$n = \left(\frac{2.576}{0.008} \right)^2 (0.5)(1 - 0.5) \approx 25921$$

Thus, when a preliminary estimate p is not given, to be safe, we can take a sample of 25921.

d.

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

Answer:

We can see in this case that when preliminary information is given, we can easily get the required sample size, and it is rather small. However, when it is not given, we have to first determine the maximum value of $p(1-p)$, and the corresponding sample size is a lot bigger.

2.

The proportion of residents in Phoenix favoring the building of toll roads to complete the freeway system is believed to be $p = 0.2$. 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 proposition is $p = 0.3$.

Answer:

By definition, the probability of type I error is $P(\text{Reject } H_0 | p = 0.3)$. In this case we will reject H_0 when our sample shows that \hat{x} is no bigger than 1 (i.e. $\hat{x} = 0$ or 1). Thus, the probability can be written as:

$$\alpha = P(\hat{x} \leq 1 | p = 0.3) = P(\hat{x} = 0 | p = 0.3) + P(\hat{x} = 1 | p = 0.3)$$

Since our sample follows a binomial distribution with $n = 10$, we get:

$$\alpha = \binom{10}{0} (0.3)^0 (0.7)^{10} + \binom{10}{1} (0.3)^1 (0.7)^9 \approx 0.1493$$

The probability of type I error is approximately 0.1493.

b.

Find the probability of committing a type II error with this procedure if $p = 0.2$.

Answer:

By definition, the probability of type II error is $P(\text{Fail to Reject } H_0 | p = 0.2)$. In this case, we fail to reject H_0 whenever $\hat{x} > 1$; thus, the probability is given by $P(\hat{x} > 1 | p = 0.2)$.

$$\beta = P(\hat{x} > 1 | p = 0.2) = 1 - P(\hat{x} \leq 1 | p = 0.2)$$

Again we have:

$$P(\hat{x} \leq 1 | p = 0.2) = P(\hat{x} = 0 | p = 0.2) + P(\hat{x} = 1 | p = 0.2)$$

Hence,

$$\beta = 1 - \binom{10}{0} (0.2)^0 (0.8)^{10} + \binom{10}{1} (0.2)^1 (0.8)^9 \approx 0.6241$$

c.

What is the power of this procedure if the true proportion is $p = 0.2$?

Answer:

By definition, the power of a test is $1 - \beta$, and in this case, it is approximately 0.3758.

3.

Assume that X_1, \dots, X_9 are i.i.d. having Bernoulli distribution with parameter p . Suppose that we wish to test the hypotheses

$$H_0 : p = 0.4$$

$$H_1 : p \neq 0.4$$

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 as possible to 0.1 without being larger than 0.1.

Answer:

```
rm(list=ls())
result<-matrix(NA, nrow=100, ncol=3)
n=1
for(i in 0:9){
  for(j in 0:9){
    p=pbinom(q=i, size=9, prob=0.4, lower.tail=T)+pbinom(q=j, size=9, prob=0.4, lower.tail=F)+
      dbinom(x=j, size=9, prob=0.4)
    result[n,1]=i
    result[n,2]=j
    result[n,3]=p
    n=n+1
  }
}
result<-result[order(result[,3]),]
result[which(result[,3]<0.1),]

##      [,1] [,2]      [,3]
## [1,]    0    9 0.01033984
## [2,]    0    8 0.01387878
## [3,]    0    7 0.03511245
## [4,]    1    9 0.07080602
## [5,]    1    8 0.07434496
## [6,]    1    7 0.09557862
```

So we can see that the values of c_1 and c_2 that make $P(Y \leq c_1 | p = 0.4) + P(Y \geq c_2 | p = 0.4)$ as close to 0.1 as possible are 1 and 7.

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 ?

Answer:

If we reject the null hypothesis when either $Y \leq c_1$ or $Y \geq c_2$, the size of the test is

$$P(Y \leq 1 | p = 0.4) + P(Y \geq 7 | p = 0.4) \approx 0.096$$

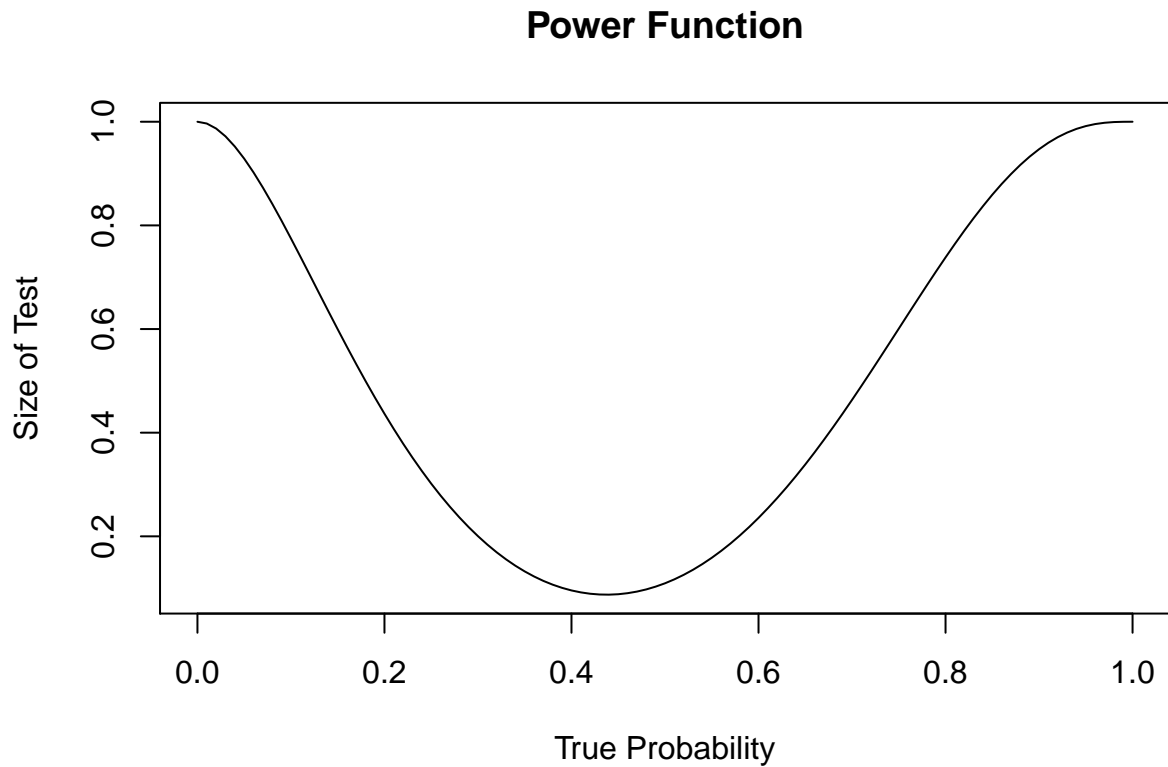
c.

Draw a graph of the power function δ_c .

Answer:

```
rm(list=ls())
size<-vector()
for(i in seq(from=0, to=1, by=0.01)){
  size<-append(size, pbinom(q=1, size=9, prob=i, lower.tail=T)+
    pbinom(q=7, size=9, prob=i, lower.tail=F)+dbinom(x=7, size=9, prob=i))
}
```

```
n<-seq(from=0, to=1, by=0.01)
plot(n, size, type="l", ylab="Size of Test", xlab="True Probability", main="Power Function")
```



4.

The file 'Prob4_data.txt' contains 30 observations from a Gamma distribution with unknown parameters α and β .

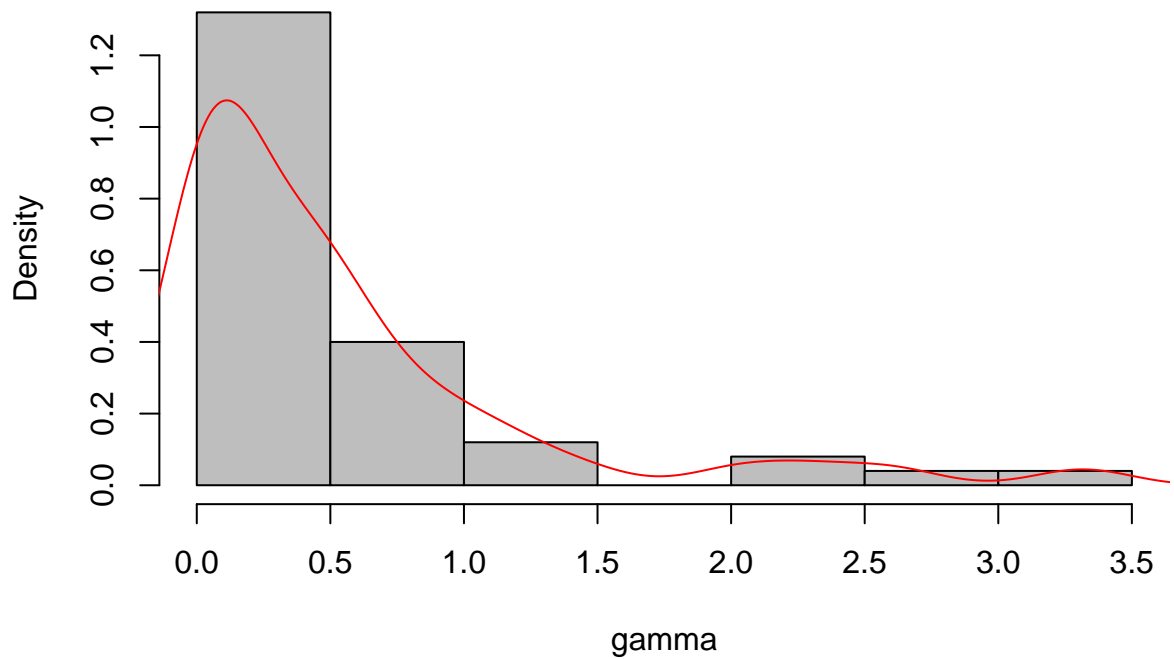
a.

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

Answer:

```
rm(list=ls())
setwd("/Users/patrick/Desktop/Econ 403A/Assignments/Homework 2")
data<-read.table("Prob4_data.txt")
gamma<-as.numeric(data$V1)
hist(gamma, col="gray", freq=F)
lines(density(gamma), col="red")
```

Histogram of gamma



b.

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

Answer:

```
n=length(gamma)
mean=mean(gamma)
sum=sum(gamma^2)
alpha=mean^2/((1/n)*sum-mean^2)
beta=mean/((1/n)*sum-mean^2)
print(alpha)
```

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

```
print(beta)
```

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

We get $\hat{\alpha} \approx 0.58$ and $\hat{\beta} \approx 1.08$.

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).

Answer:

```
library(boot)
fun1<-function(data, indices){
```

```

sample=data[indices]
n=length(sample)
sample_mean=mean(sample)
sample_alpha=sample_mean^2/((1/n)*sum(sample^2)-sample_mean^2)
return(sample_alpha)
}
alpha_boot<-boot(gamma, fun1, R=1000)
alpha_mean=mean(alpha_boot$t)
alpha_se=sd(alpha_boot$t)
alpha_ci<-boot.ci(boot.out=alpha_boot, type=c("norm","basic","perc","bca"))
cat("The Bootstrap mean of the shape parameter is", alpha_mean, "\n")

## The Bootstrap mean of the shape parameter is 0.628761
cat("The Bootstrap standard error of the shape parameter is", alpha_se, "\n")

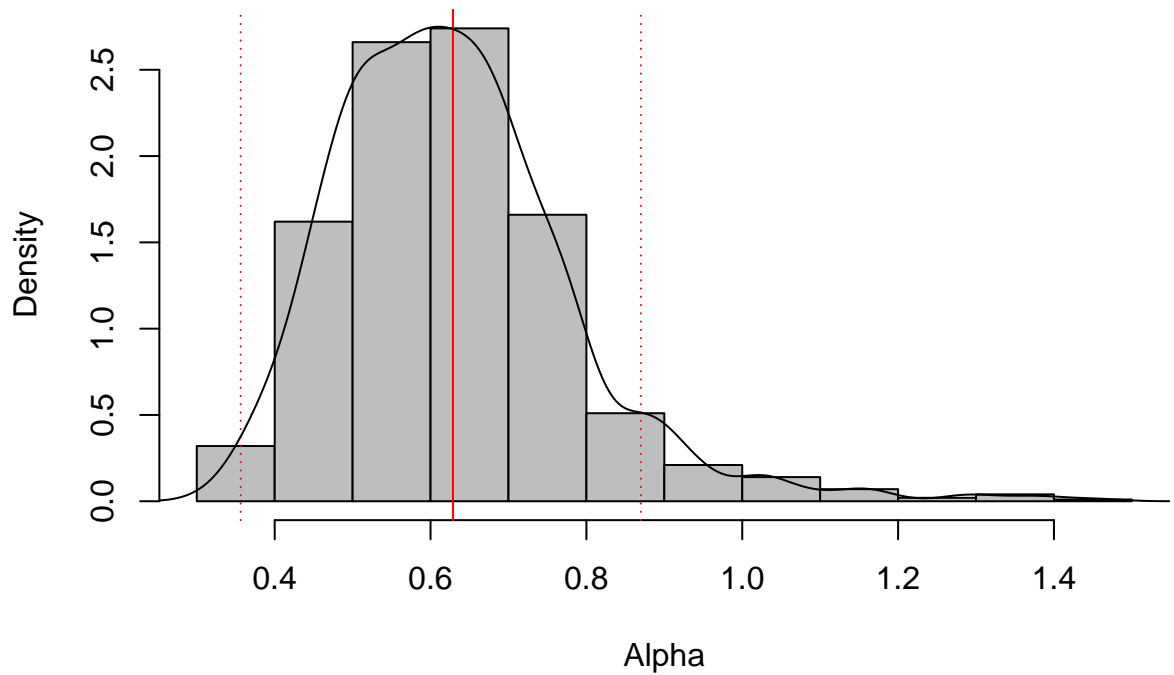
## The Bootstrap standard error of the shape parameter is 0.1547864
print(alpha_ci)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = alpha_boot, type = c("norm", "basic", "perc",
##      "bca"))
##
## Intervals :
## Level      Normal      Basic
## 95%   ( 0.2289, 0.8357 ) ( 0.1433, 0.7702 )
##
## Level      Percentile      BCa
## 95%   ( 0.3909, 1.0178 ) ( 0.3563, 0.8699 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable

alpha_ci1=alpha_ci$bca[,4:5]
hist(alpha_boot$t[,1], main="Bootstrap Values of Alpha", xlab="Alpha",
      col="grey", freq=F)
abline(v=alpha_ci1, col="red", lty="dotted")
abline(v=alpha_mean, col="red")
lines(density(alpha_boot$t))

```

Bootstrap Values of Alpha



```
library(boot)
fun2<-function(data, indices){
  sample=data[indices]
  n=length(sample)
  sample_mean=mean(sample)
  sample_beta=sample_mean/((1/n)*sum(sample^2)-sample_mean^2)
  return(sample_beta)
}
beta_boot<-boot(gamma, fun2, R=1000)
beta_mean=mean(beta_boot$t)
beta_se=sd(beta_boot$t)
beta_ci<-boot.ci(boot.out=beta_boot, type=c("norm","basic","perc","bca"))
cat("The Bootstrap mean of the rate parameter is", beta_mean, "\n")

## The Bootstrap mean of the rate parameter is 1.196985
cat("The Bootstrap standard error of the rate parameter is", beta_se, "\n")

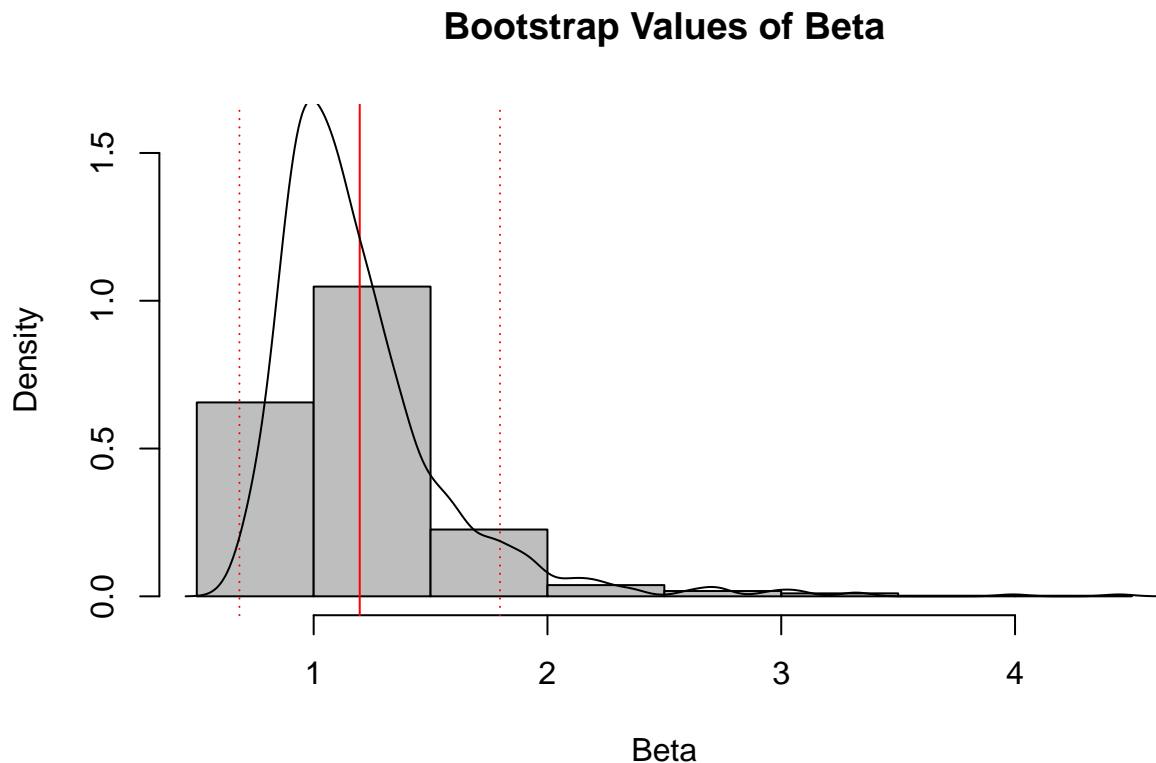
## The Bootstrap standard error of the rate parameter is 0.3874862
print(beta_ci)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = beta_boot, type = c("norm", "basic", "perc",
##    "bca"))
##
## Intervals :
## Level      Normal          Basic
```



```
## 95%   ( 0.207,  1.726 )   (-0.041,  1.415 )
##
## Level      Percentile      BCa
## 95%   ( 0.749,  2.204 )   ( 0.683,  1.797 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

```
beta_ci1=beta_ci$bca[,4:5]
hist(beta_boot$t[,1], main="Bootstrap Values of Beta", xlab="Beta",
      col="grey", ylim=c(0, 1.6), freq=F)
abline(v=beta_ci1, col="red", lty="dotted")
abline(v=beta_mean, col="red")
lines(density(beta_boot$t))
```



We can see that our bootstrap means are larger than MOM estimates, but the former's confidence intervals generated by the bootstrap process contain the latter.

5.

A 1992 article in the *Journal of the American Medical Association* ("A Critical Appraisal of 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 temperature 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.

Answer:

```
rm(list=ls())
library(stats)
library(distrEx)
library(MASS)
temp<-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)
t.test(temp, mu=98.6, conf.level=0.95)
```

```
##
## One Sample t-test
##
## data: temp
## t = -3.4849, df = 24, p-value = 0.001912
## alternative hypothesis: true mean is not equal to 98.6
## 95 percent confidence interval:
## 98.06501 98.46299
## sample estimates:
## mean of x
## 98.264
```

Since our p-value is 0.001912, a lot smaller than α , we should reject H_0 .

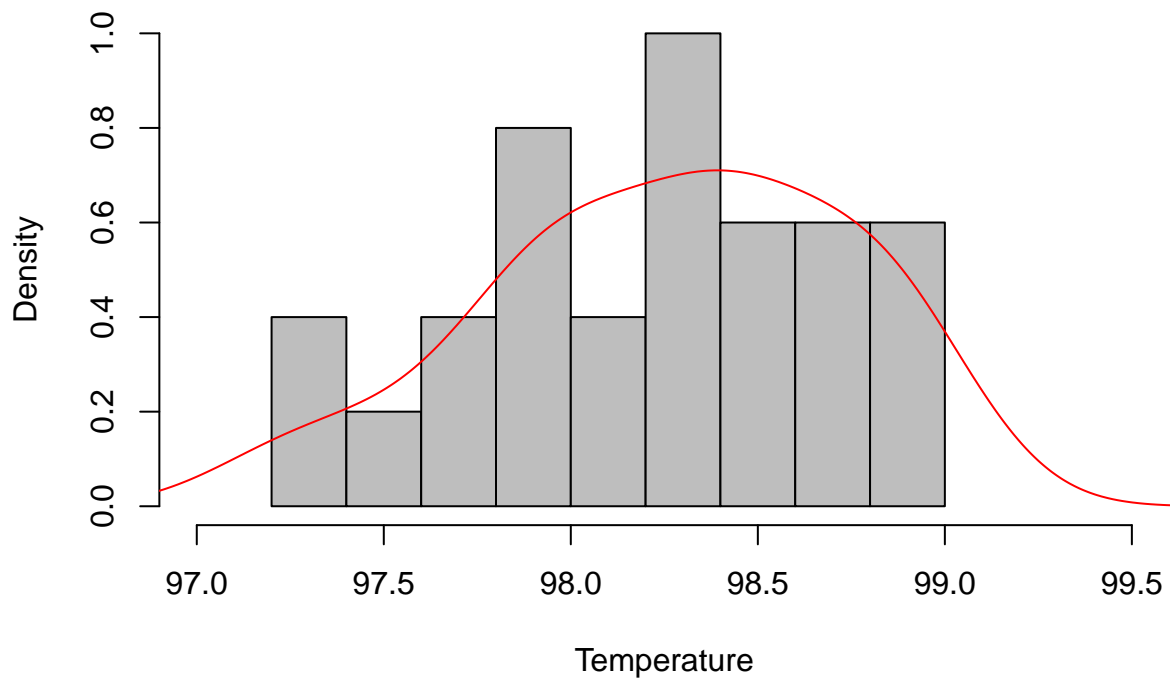
b.

Check the assumption that female body temperature is normally distributed.

Answer:

```
hist(temp, main="Female Body Temperatures", xlab="Temperature", freq=F,
      col="grey", breaks=9, xlim=c(97, 99.5))
lines(density(temp), col="red")
```

Female Body Temperatures



```
normalfit<-fitdistr(temp, "normal")
normal<-rnorm(n=25, mean=normalfit[["estimate"]][["mean"]], sd=normalfit[["estimate"]][["sd"]])
ks.test(temp, normal)
```

```
##
## Two-sample Kolmogorov-Smirnov test
##
## data: temp and normal
## D = 0.24, p-value = 0.4676
## alternative hypothesis: two-sided
```

A large p-value suggests that female body temperature does follow a normal distribution.

c.

Compute the power of the test if the true mean female body temperature is as low as 98.0.

Answer:

```
library(pwr)
d=(98.6-98)/sd(temp)
pwr.t.test(n=25, d, sig.level=0.05, type="one.sample", alternative="two.sided")
```

```
##
## One-sample t test power calculation
##
## n = 25
## d = 1.24461
## sig.level = 0.05
## power = 0.9999683
## alternative = two.sided
```

So the power of the test is approximately 99.9%, which means that the probability of rejecting the null hypothesis when it is false is almost 100%.

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?

Answer:

```
d=(98.6-98.2)/sd(temp)
pwr.t.test(d=d, sig.level=0.05, type="one.sample", alternative="two.sided", power=0.9)
```

```
##
##      One-sample t test power calculation
##
##              n = 17.29754
##              d = 0.8297398
##      sig.level = 0.05
##              power = 0.9
##      alternative = two.sided
```

So we can see that we need approximately a sample size of 18 in order for the power of the test to be at least 0.9.

e.

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

Answer:

By definition, 95% confidence intervals in the long run have a 95% ‘success rate’ of capturing the real population mean; or in other words, among all the confidence intervals established using different samples, only 5% of them do not capture population mean. In this case, if the two-sided confidence interval does not contain our hypothesis, it is very likely that our hypothesis is wrong.

6.

Suppose a sample of size 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 committing a Type I error.

Answer:

By definition, the probability of committing a Type I error is given as $P(\text{Reject } H_0 | H_0 \text{ is true})$. In this case, it can be written as $P(y \geq 3.2 | \lambda = 1)$, which is

$$\int_{3.2}^{\infty} e^{-y} dy$$

```
rm(list=ls())
exp1<-function(y){
  return(dexp(y, rate=1))
}
integrate(exp1, lower=3.2, upper="infinity")
```

```
## 0.0407622 with absolute error < 2.3e-06
```

b.

Calculate the probability of committing a Type II error when $\lambda = 4/3$.

Answer:

By definition, the probability of committing a Type II error is given as $P(\text{Fail to reject } H_0 | H_0 \text{ is false})$. In this case, it can be written as $P(y \leq 3.2 | \lambda = \frac{4}{3})$, which is

$$\int_0^{3.2} \frac{3}{4} e^{-\frac{3}{4}y} dy$$

```
exp2<-function(y){
  return(dexp(y, rate=3/4))
}
integrate(exp2, lower=0, upper=3.2)
```

```
## 0.909282 with absolute error < 1e-14
```

c.

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

Answer:

```
alpha<-vector()
exp1<-function(y){
  return(dexp(y, rate=1))
}
for(l in seq(from=0, to=4, by=0.1)){
  alpha<-append(alpha, integrate(exp1, lower=l, upper="infinity")[[ "value" ]])
}

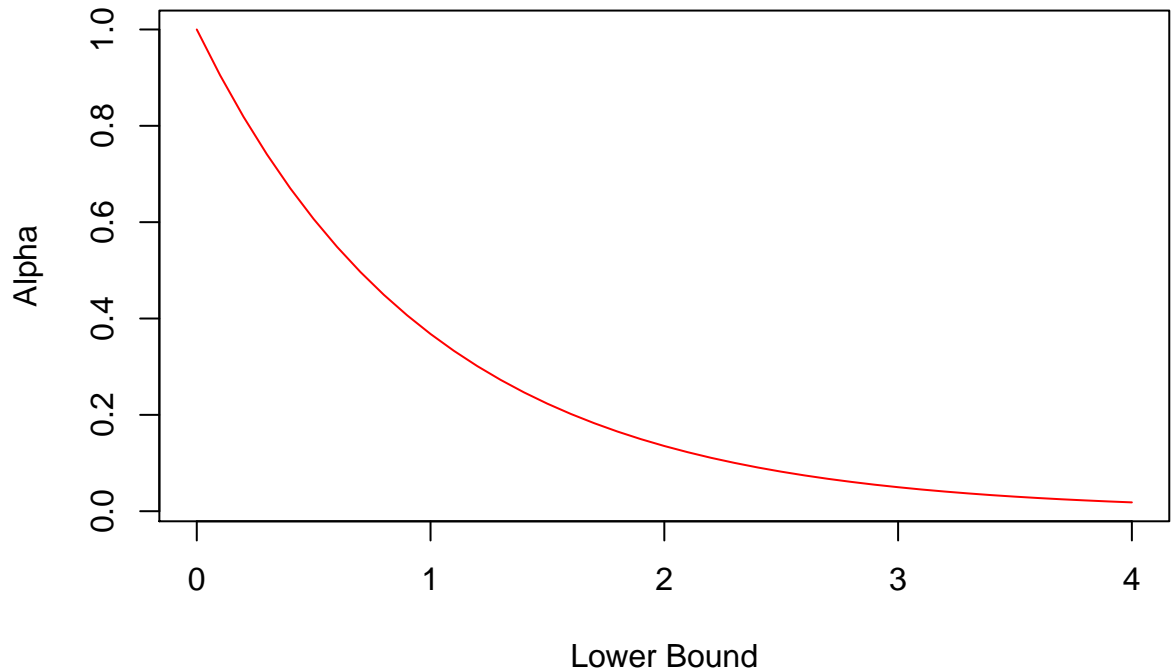
beta<-vector()
exp2<-function(y){
  return(dexp(y, rate=3/4))
}
```

```

}
for(u in seq(from=0, to=4, by=0.1)){
  beta<-append(beta, integrate(exp2, lower=0, upper=u)[["value"]])
}

t=seq(from=0, to=4, by=0.1)
plot(t, alpha, type="l", col="red", ylab="Alpha", xlab="Lower Bound")

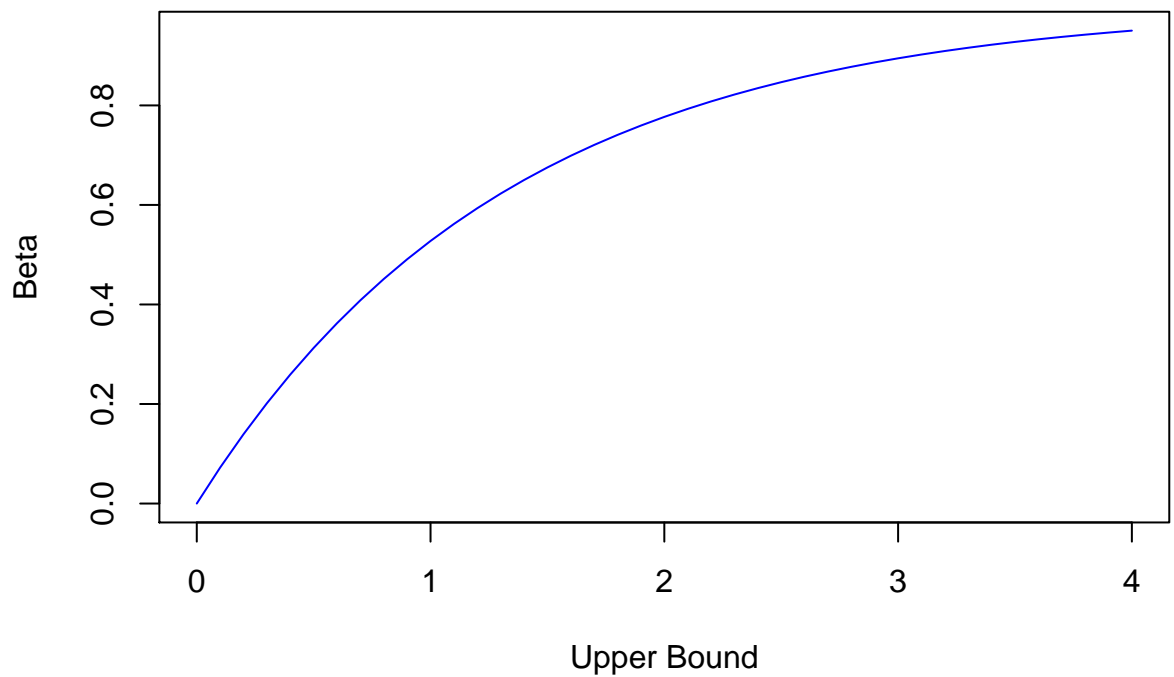
```



```

plot(t, beta, type="l", col="blue", ylab="Beta", xlab="Upper Bound")

```



7. Evans & Rosenthal: 5.4.12

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

a.

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

Answer:

```
sample1<-sample(1:10000, size=500, replace=F)
print(sample1)
```

```
## [1] 255 3944 2821 2295 9040 6239 1274 3581 6251 223 5109 4661 8537 1767
## [15] 9454 7198 5747 2839 2393 2119 9614 4086 8035 1900 8802 5422 8257 8201
## [29] 8294 4650 6187 7007 9712 2898 6671 7205 9180 9428 2639 4430 6125 9096
## [43] 8990 9575 43 2436 3199 3963 5363 2984 3931 2741 3680 4203 250 1532
## [57] 731 6719 9873 7211 3018 9138 2433 7875 7464 8157 8444 1826 535 8591
## [71] 5256 8505 1764 3055 8600 8259 959 4120 6436 584 7137 8485 6420 9141
## [85] 511 7855 1887 4516 9183 1210 684 5226 3238 6642 6040 9278 9977 4877
## [99] 6876 9509 7360 6385 6862 5766 8408 2126 6657 895 287 4044 585 4764
## [113] 8118 5698 2547 5651 6888 4800 3257 376 3857 8849 7809 2942 5893 7443
## [127] 8985 9043 1922 4247 229 8938 3663 8108 5793 5588 5544 2900 3236 1965
## [141] 1477 5387 8665 9989 1588 7296 8835 9326 3702 6740 1252 5148 180 2240
## [155] 338 6782 2999 7754 3124 3356 4379 294 9272 3516 4059 6006 3041 6863
## [169] 5773 1874 8718 132 4945 1797 6805 3785 907 8701 9357 8508 2829 4208
## [183] 5921 9961 1810 3699 1791 4855 9181 5768 9252 2696 8644 6383 3089 5443
## [197] 6102 9531 7478 2510 8695 7702 1121 1462 2124 7285 3869 4695 2517 8823
## [211] 967 1688 4993 1205 6072 9880 9417 1926 8324 2855 2207 4612 8191 3506
## [225] 6197 4854 1542 144 9387 59 3180 5075 8738 4005 3514 2694 2059 6564
## [239] 5839 4174 3529 2638 1572 7086 2463 6809 6918 6802 5669 6059 4895 7460
## [253] 4658 3968 7965 3773 6578 7936 8228 5803 6091 1696 7167 2816 8059 9721
## [267] 6926 7076 2594 2434 1723 7445 6730 8323 5195 7113 7905 3329 3864 836
## [281] 2552 9236 26 6785 5498 1909 8951 1169 1177 2943 4707 9672 1892 4884
## [295] 4769 7851 748 6948 2716 429 3688 3661 224 5929 5398 622 6535 8796
## [309] 5464 9795 3318 9474 7364 2342 5423 7044 1682 5814 1554 1227 3739 5537
## [323] 5289 4603 7847 5912 2764 7102 2586 925 1907 7001 4515 6050 7341 6914
## [337] 1155 5271 2311 2396 6677 5044 9878 2667 6329 4654 813 2770 6106 114
## [351] 5758 4026 783 8578 3406 3956 6427 7977 5691 5318 7689 3964 3546 304
## [365] 8136 6324 942 7925 9650 8009 8479 1356 9142 5551 2534 3408 2227 1754
## [379] 1903 4156 9548 2187 3556 7524 6451 100 7585 2662 5240 1755 9163 8698
## [393] 8645 2482 1864 3198 5091 9885 7870 8231 7509 2557 7539 7884 170 6327
## [407] 5073 549 5567 106 408 7259 4760 2098 3677 9725 1769 9470 9788 7074
## [421] 5858 2235 2989 9979 2472 9965 2683 1657 8384 222 6314 7733 5333 428
## [435] 1766 2180 7710 1316 773 6262 9087 6581 6254 3168 5862 1744 8788 7516
## [449] 9413 9257 8580 2425 3004 3297 9699 6660 7930 9828 1326 6127 9723 2133
## [463] 9954 9502 6768 4576 6337 708 5625 5996 1334 5524 7749 3069 7325 7660
## [477] 9786 9243 9444 6153 809 396 5079 7521 884 8774 51 2275 4722 372
## [491] 4449 6489 7144 3222 2471 7166 149 1418 6605 3915
```

b.

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

Answer:

```
sample2<-sample(1:10000, size=500, replace=T)
print(sample2)
```

```
## [1] 7394 1596 4549 6719 6947 1632 4859 7812 7792 6575 4671 5131 6540 9936
## [15] 9054 7367 4011 7129 9908 8516 9666 9901 3377 9213 315 4625 4878 8110
## [29] 9385 9502 1517 1344 1414 7409 9033 1544 3444 3443 7793 4578 557 71
## [43] 6924 1331 9056 7312 973 6941 8927 6824 7166 2139 3041 4553 123 1072
## [57] 631 4539 9170 557 6184 7556 3281 2715 5147 2274 5015 9536 3670 3625
## [71] 278 9038 9910 6760 4484 9226 4132 7406 187 5041 7848 3552 7112 8096
## [85] 2822 6972 5543 7236 4885 4380 9434 6713 8298 1488 1650 8083 9210 5110
## [99] 4604 226 1388 6577 2990 8192 7638 897 5353 8784 8976 7104 3500 2088
## [113] 9316 3850 6142 4879 407 9657 3320 2074 6119 1532 5541 428 3391 6115
## [127] 8153 1276 2428 6005 2972 1513 5390 7749 2638 3047 7793 2341 8019 455
## [141] 9783 1009 6048 5745 8304 828 1516 5794 2761 3994 4423 5364 7746 3741
## [155] 4343 4642 2530 4366 8060 7233 4738 9829 1811 2245 7988 7932 2072 6700
## [169] 5868 815 9194 5933 7187 4973 1339 8542 9186 8091 6968 3258 271 1308
## [183] 8218 3107 8097 1658 7175 9753 440 6156 2641 4168 4239 728 3378 1050
## [197] 1982 4430 4724 6251 1053 3393 72 4595 9178 9946 5999 6748 9308 5865
## [211] 2428 8211 3504 5438 8631 9990 4670 9813 9421 1632 6030 8665 2927 8224
## [225] 713 5837 1639 4957 6706 461 9735 719 6278 6530 3204 8386 9905 8745
## [239] 6651 2070 6676 1721 7704 6222 294 2380 1639 6114 8050 4291 6398 714
## [253] 216 9546 6102 9147 7436 255 3732 6894 4013 3286 5860 5241 1534 3824
## [267] 3654 6375 8413 355 4785 8606 7368 2993 9268 3467 5350 7812 9717 6056
## [281] 4463 2955 8921 9938 7645 2661 2780 5539 7932 576 1201 4553 6951 4326
## [295] 1317 8607 7909 1952 3711 696 9516 9642 8597 2623 1938 6325 5213 1971
## [309] 8177 3131 6496 1117 3644 8523 1367 4580 8065 9415 3534 3338 6724 990
## [323] 4529 129 254 4117 9151 4771 5787 5105 9403 135 4862 7290 6003 9990
## [337] 1947 4406 919 4813 7801 9938 7420 2551 6922 3770 1189 8529 532 9398
## [351] 5780 3728 3790 8327 7041 2304 46 1495 631 5857 2361 4275 2502 6250
## [365] 6012 7220 2807 2159 98 3740 4832 2097 2722 9878 2139 6756 1726 4747
## [379] 415 4317 5582 3232 6708 2751 5855 9643 1036 2335 4122 9632 7292 7535
## [393] 4395 2277 6623 2593 1831 6685 1839 5589 1947 3589 2127 3757 164 4695
## [407] 998 9855 4379 2422 5241 7173 6298 7818 6348 5261 1352 9237 4726 8821
## [421] 4353 3842 1708 9981 7239 1646 5810 4429 2521 3373 9499 6380 9612 9335
## [435] 2901 6057 8382 866 5682 8446 4003 4009 9313 1185 7597 2158 5108 9417
## [449] 6907 2560 9036 2935 942 7434 5738 8496 7075 9014 7181 9053 3774 5952
## [463] 6155 646 9932 27 2371 1299 5061 3736 4654 7857 3046 3912 4599 9082
## [477] 8725 4248 1552 240 2901 1704 7014 896 7838 8589 1813 1322 9402 1914
## [491] 4828 344 4433 2757 1583 420 6204 1996 7688 6234
```

8. Evans & Rosenthal: 5.5.18

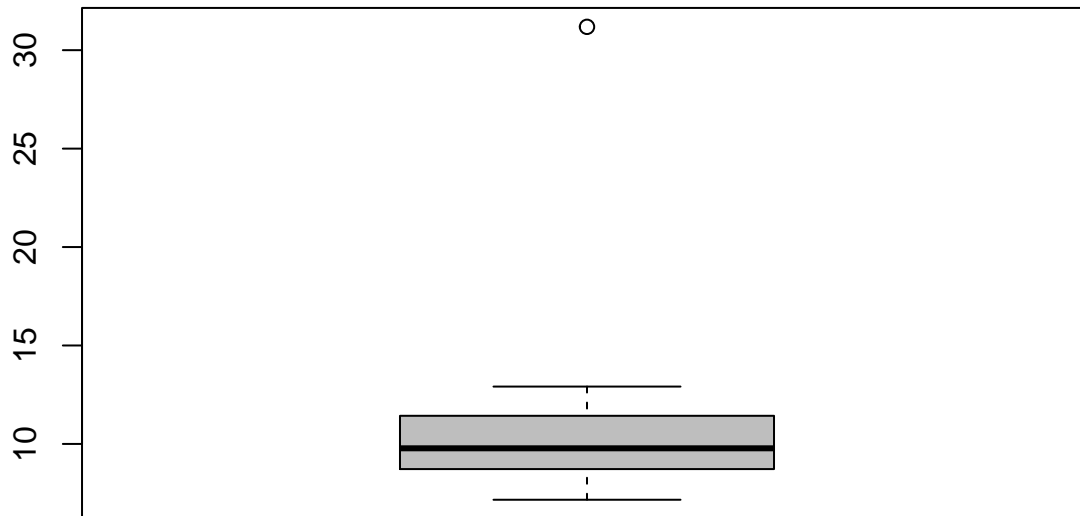
Generate a sample of 30 from an $N(10, 2)$ distribution and a sample of 1 from an $N(30, 2)$ distribution. Combine these together to make a single sample of 31.

a.

Produce a boxplot of these data.

Answer:


```
sample1<-rnorm(n=30, mean=10, sd=2)
sample2<-rnorm(n=1, mean=30, sd=2)
sample<-append(sample1, sample2)
boxplot(sample, col="grey")
```



b.

What do you notice about this plot?

Answer:

There is a single outlier in the boxplot, and it comes from the $N(30, 2)$ distribution.

c.

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

Answer:

Since this boxplot shows an outlier that actually comes from a different distribution, it suggests that we can just ignore the outlier. Thus, the characteristics that best measure the location and spread are still the mean and the variance.

9. Evans & Rosenthal: 6.2.17

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 equispaced points in $(-10, 10]$. Also plot the likelihood function. Comment on the form of likelihood intervals.

Answer:

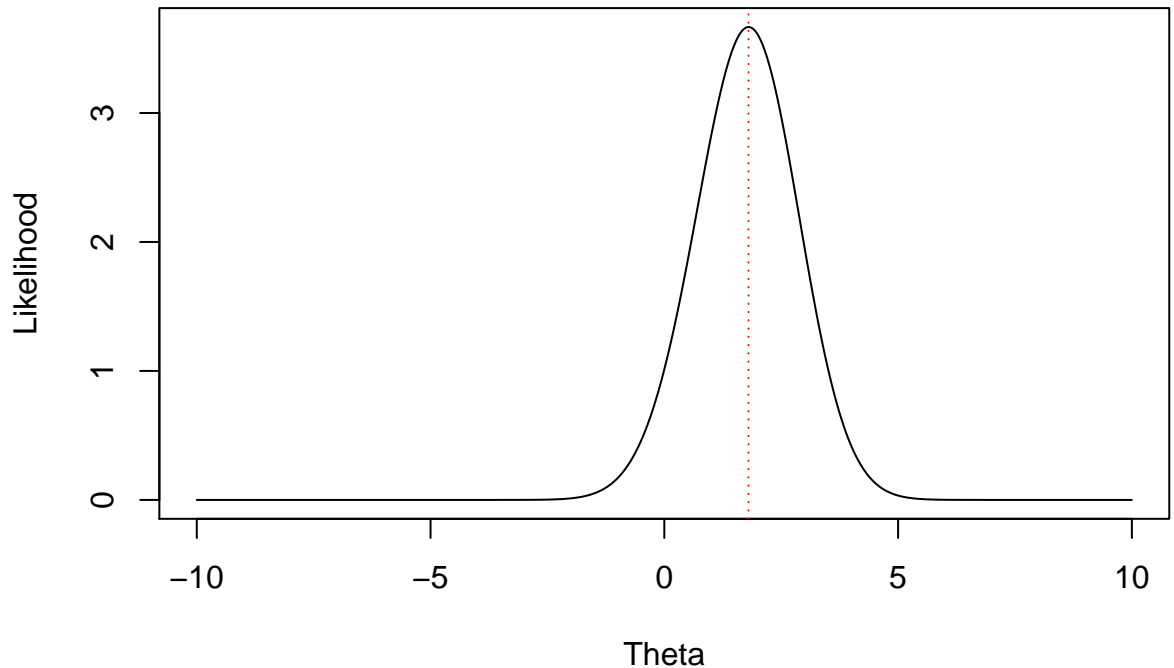
```
rm(list=ls())
y<-vector()
n<-seq(from=-10, to=10, by=0.02)
for(i in seq(from=-10, to=10, by=0.02)){
  y<-append(y, exp(-(i-1)^2/2)+3*exp(-(i-2)^2/2))
}
```

```

}
n[which(y==max(y))]

## [1] 1.8
plot(n, y, type="l", xlab="Theta", ylab="Likelihood")
abline(v=n[which(y==max(y))], lty=3, col="red")

```



10. Evans & Rosenthal: 6.2.25

Suppose the proportion of left-handed individuals in a population is 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.

Answer:

When the sample size is small relative to the population size, the number of left-handed individuals follows a binomial distribution:

$$P(X = x) = \binom{20}{x} \theta^x (1 - \theta)^{20-x}$$

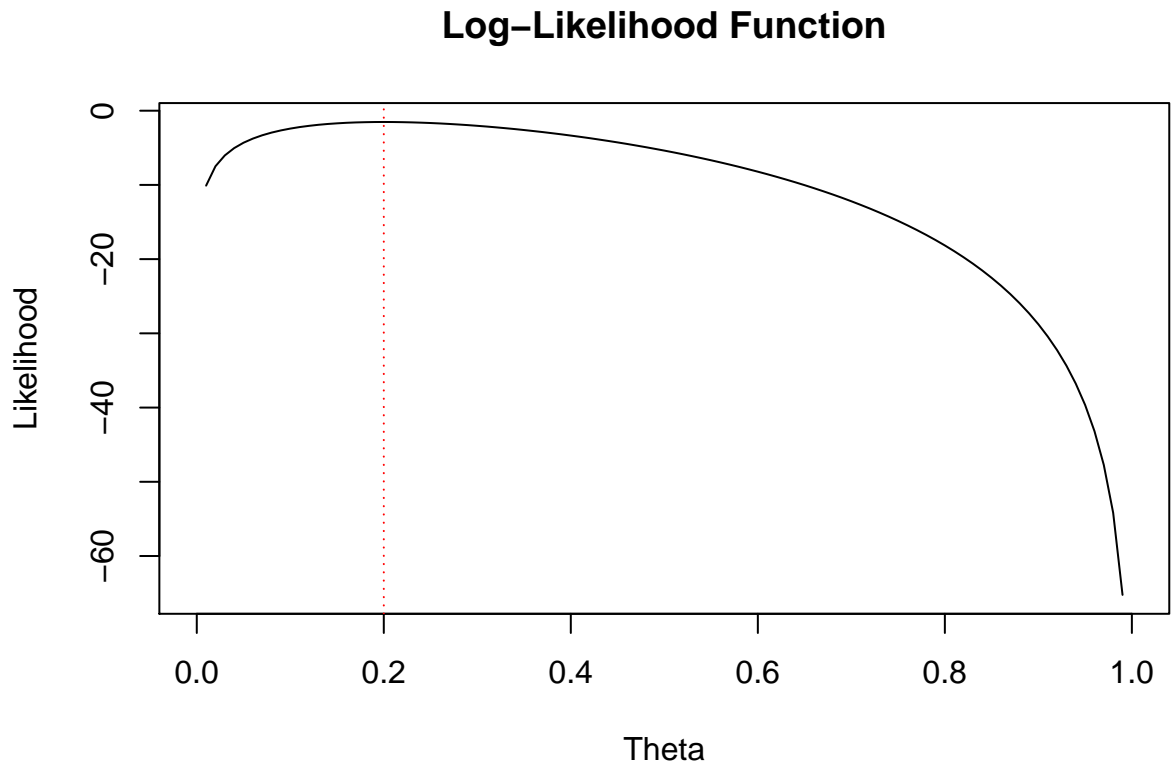
Here our x is 4; plug it in we get

$$P(X = 4) = \binom{20}{4} \theta^4 (1 - \theta)^{16}$$

Taking logarithm, we get

$$\ln P(X = 4) = \ln \binom{20}{4} + 4 \ln \theta + 16 \ln(1 - \theta)$$

```
rm(list=ls())
y<-vector()
theta<-seq(from=0, to=1, by=0.01)
for(i in seq(from=0, to=1, by=0.01)){
  y<-append(y,log(choose(20,4))+4*log(i)+16*log(1-i))
}
plot(theta, y, type="l", xlab="Theta", ylab="Likelihood", main="Log-Likelihood Function")
abline(v=theta[which(y==max(y))], lty=3, col="red")
```



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.)

Answer:

When the population size N is 50, the probability that among the 20 people selected 4 are left-handed is given by:

$$P(X = 4) = \frac{\binom{n}{4} \binom{50-n}{16}}{\binom{50}{20}}$$

where n is the total number of left-handed individuals in the population and $n = 50\theta$.

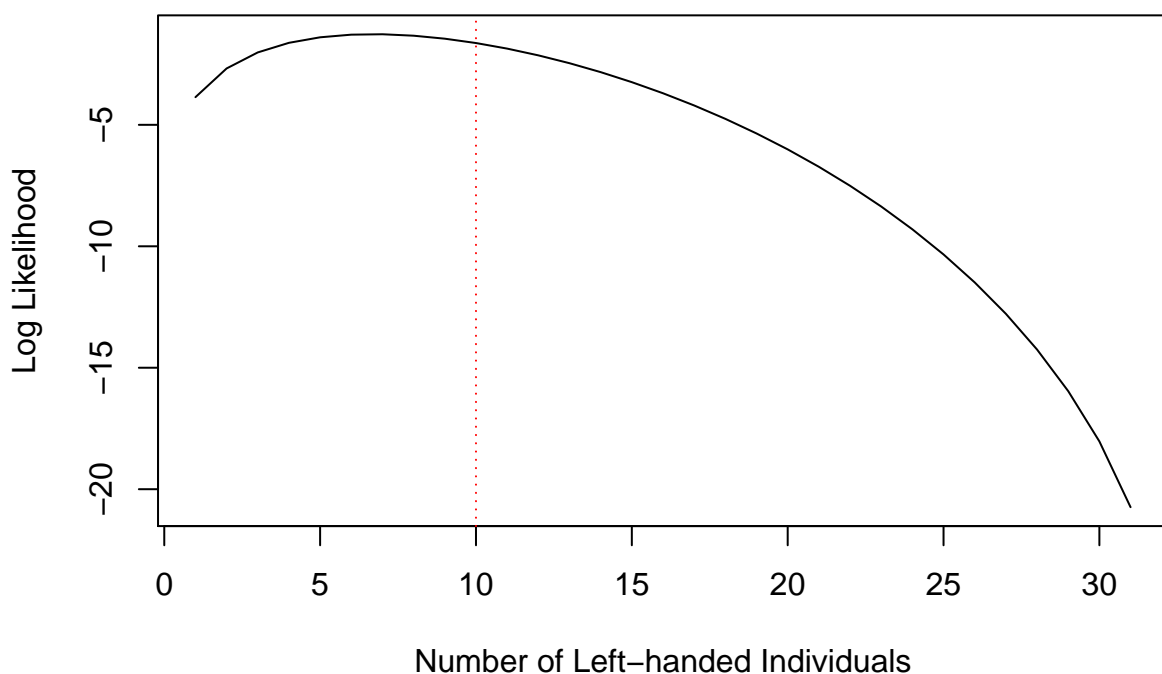
This equation tells us the $n \geq 4$ and $16 \leq 50 - n$, and hence we have $4 \leq n \leq 34$.

Take logarithm of the likelihood function, and we get:

$$\ln P(X = 4) = \ln \binom{n}{4} + \ln \binom{50-n}{16} - \ln \binom{50}{20}$$

```
x<-4:34
y=log(choose(x, 4))+log(choose(50-x, 16))-log(choose(50,20))
plot(y, type="l", xlab="Number of Left-handed Individuals",
     ylab="Log Likelihood", main="Log-Likelihood Function")
abline(v=x[which(y==max(y))], lty=3, col="red")
```

Log-Likelihood Function



Since the n that maximized the log-likelihood function is 10, our $\hat{\theta}$ should be $10/50$, which is 0.2.

11. Evans & Rosenthal: 6.3.22

Generate 10^4 samples of sizes $n = 5$ 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 this interval contains μ . Repeat this simulation with $n = 10$ and 100 and compare your results.

Answer:

```
normal_ci<-function(size=5, trial=10000){
  size=size
  trial=trial
  n=0
```

```

samples<-matrix(NA, ncol=size, nrow=trial)
ci<-matrix(NA, ncol=2, nrow=trial)
for(i in 1:10000){
  samples[i,]<-rnorm(n=size, mean=0, sd=1)
  mean=mean(samples[i,])
  sd=sd(samples[i,])
  lower=mean-sd/sqrt(size)
  upper=mean+sd/sqrt(size)
  if(0<upper&0<lower){
    n=n+1
  }
  ci[i,1]<-lower
  ci[i,2]<-upper
}
proportion=n/trial
cat("The proportion of times that this interval contains the real mean when sample size is",
    size, "is ", proportion, "\n")
}

normal_ci(5)

## The proportion of times that this interval contains the real mean when sample size is 5 is 0.6233
normal_ci(10)

## The proportion of times that this interval contains the real mean when sample size is 10 is 0.6591
normal_ci(100)

## The proportion of times that this interval contains the real mean when sample size is 100 is 0.6823
normal_ci(500)

## The proportion of times that this interval contains the real mean when sample size is 500 is 0.6874

```

12. Evans & Rosenthal: 6.4.17

For the data of Exercise 6.4.1, use the plug-in MLE to estimate $F(3)$ for an $N(\mu, \sigma^2)$ distribution. Use bootstrapping to estimate the MSE of this estimate for $m = 10^3$ and $m = 10^4$.

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

Answer:

```

rm(list=ls())
library(stats)
library(distrEx)
library(MASS)
library(boot)
data<-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(data, densfun="normal")
p<-pnorm(q=3, mean=MLE[["estimate"]][["mean"]], sd=MLE[["estimate"]][["sd"]])

```

```

fun<-function(data, indices){
  sample<-data[indices]
  MLE1<-fitdistr(sample, densfun="normal")
  return(pnorm(q=3, mean=MLE1[["estimate"]][["mean"]], sd=MLE1[["estimate"]][["sd"]]))
}
boot1<-boot(data, statistic=fun, R=1000)
MSE1<-sd(boot1$t)^2+(mean(boot1$t)-p)^2

boot2<-boot(data, statistic=fun, R=10000)
MSE2<-sd(boot2$t)^2+(mean(boot2$t)-p)^2

cat("When m=1000, the bootstrap MSE is", MSE1, "\n")

## When m=1000, the bootstrap MSE is 0.008894618
cat("When m=10000, the bootstrap MSE is", MSE2, "\n")

## When m=10000, the bootstrap MSE is 0.008869473

```