Chapter 7

Dr Wayne Stewart

Estimation Using Confidence Intervals

OBJECTIVE

> *To explain the basic concepts of statistical estimation; to present some estimators and to illustrate their use in practical sampling situations involving one or two samples*

CONTENTS

7.1 Point Estimators and Their Properties
7.2 Finding Point Estimators: Classical Methods of Estimation
7.3 Finding Interval Estimators: The Pivotal Method
7.4 Estimation of a Population Mean
7.5 Estimation of the Difference Between Two Population Means: Independent Samples
7.6 Estimation of the Difference Between Two Population Means: Matched Pairs
7.7 Estimation of a Population Proportion
7.8 Estimation of the Difference Between Two Population Proportions
7.9 Estimation of a Population Variance
7.10 Estimation of the Ratio of Two Population Variances
7.11 Choosing the Sample Size
7.12 Alternative Interval Estimation Methods: Bootstrapping and Bayesian Methods (*Optional*)

Calculate n

STATISTICS IN ACTION

Bursting Strength of PET Beverage Bottles

Polyethylene terephthalate (PET) bottles are used for carbonated beverages. At a certain facility, PET bottles are produced by inserting injection molded pre-forms into a stretch blow machine with 24 cavities. Each machine can produce 440 bottles per minute. A critical property of PET bottles is their bursting strength — the pressure at which bottles filled with water burst when pressurized.

In the *Journal of Data Science* (May 2003), researchers measured and analyzed the bursting strength of PET bottles made from two different mold designs – an old design and a new design. The new mold design reduces the time to change molds in the blow machine, thus reducing the downtime of the machine. This advantage, however, will be negated if the new design has problems with low bursting strength. Consequently, an analysis was performed to compare bursting strengths of PET bottles produced from the two mold designs.

The data for the analysis was obtained by testing one bottle per cavity per day produced over a period of 32 days for each design. Since the machine has 24 cavities, there were a total of 32 × 24 = 768 PET bottles tested for each design. Each bottle was filled with water and pressurized until it burst and the resulting pressure (in pounds per square inch) was recorded. These bursting strengths are saved in the **PETBOTTLE** file, described in Table SIA7.1. The researchers showed that there were no significant trends in bursting strength over time (days) and that there were no "cavity effects" (i.e., no significant bursting strength differences among the 24 cavities within each design). Thus, the data for all cavities and all days were pooled in order to compare the two mold designs.

PETBOTTLE

TABLE SIA7.1:

Variable Name	Description	Data Type
DESIGN	Mold design (OLD or NEW)	Qualitative
DAY	Day number	Quantitative
CAVITY	Cavity number	Quantitative
STRENGTH	Bursting strength (psi)	Quantitative

In the Statistics in *Action Revisited* at the end of this chapter, we demonstrate how the methods outlined in this chapter can be used to compare bursting strengths of PET bottles produced from the two mold designs.

Definition 7.1

A **point estimator** is a random variable that represents a numerical estimate of a population parameter based on the measurements contained in a sample. The single number that results from the calculation is called a **point estimate**.

> **Definition 7.2**
>
> An **interval estimator** is a pair of random variables obtained from the sample data used to form an interval that estimates a population parameter.

FIGURE 7.1
Sampling distribution of a sample mean for large samples

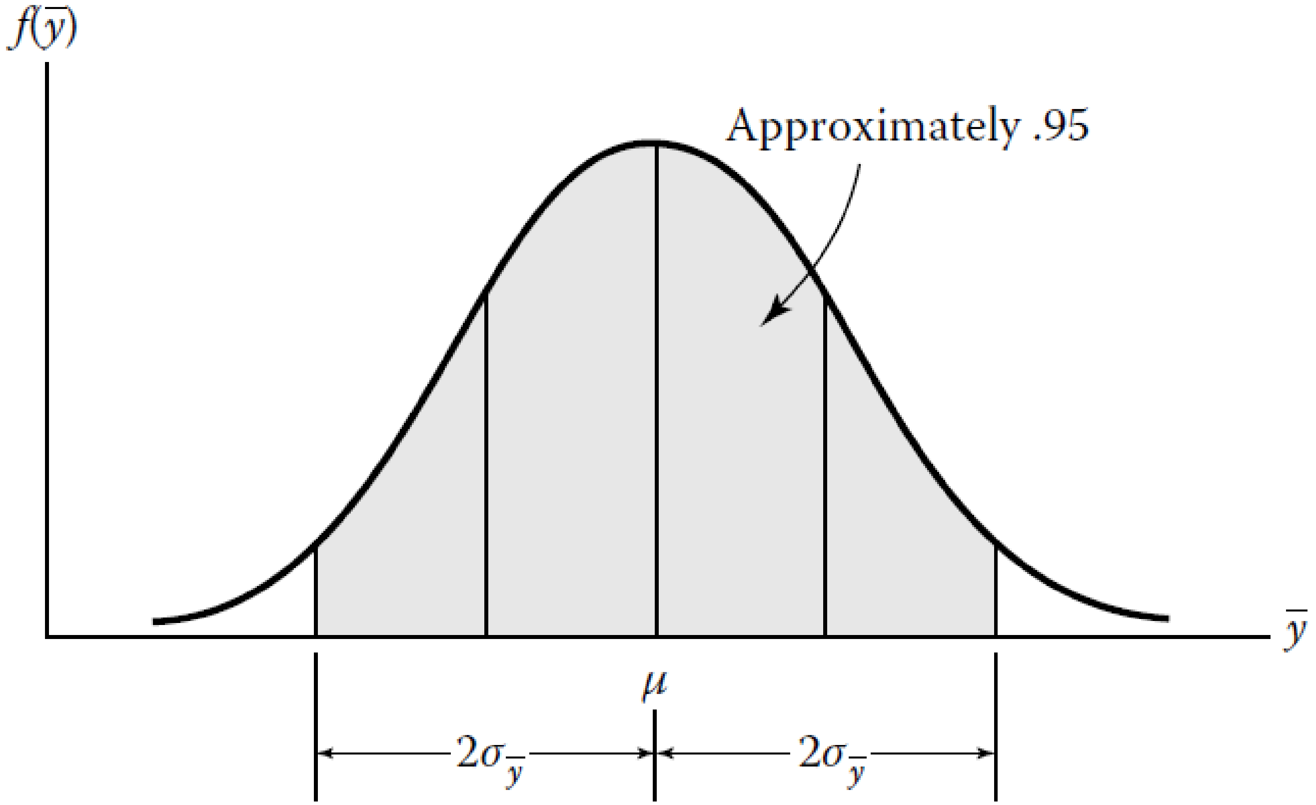
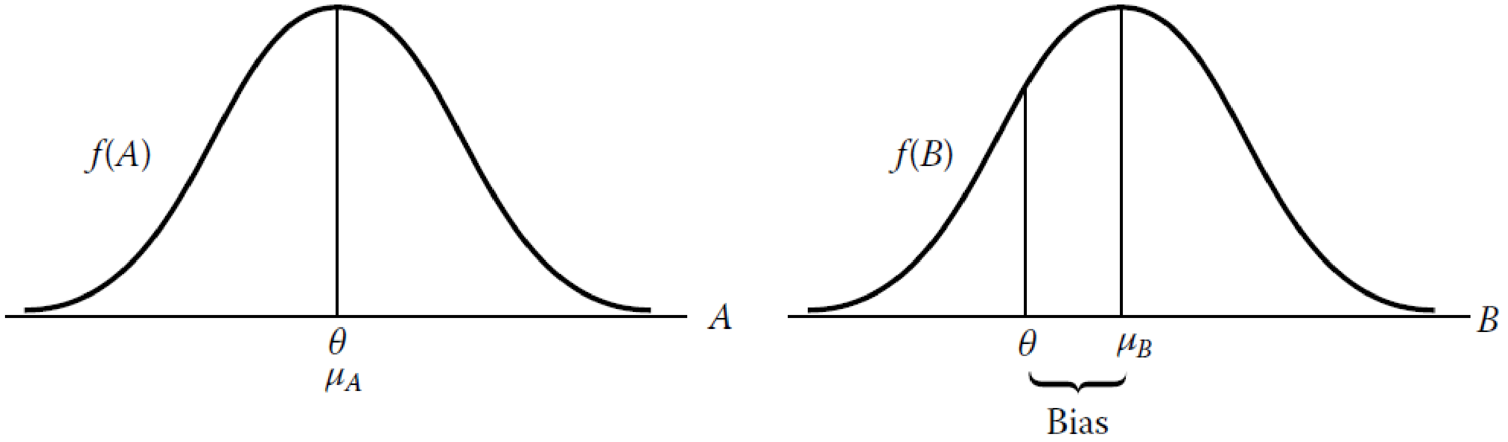


FIGURE 7.2
Sampling distributions for unbiased and biased estimators of θ



a. Estimator *A* is unbiased.

b. Estimator *B* is biased.

> **Definition 7.3**
> An estimator $\hat{\theta}$ of a parameter θ is **unbiased** if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$, the estimator is said to be **biased**.

> **Definition 7.4**
> The **bias** $b(\boldsymbol{\theta})$ of an estimator $\hat{\theta}$ is equal to the difference between the mean $E(\hat{\theta})$ of the sampling distribution of $\hat{\theta}$ and θ, i.e.,
>
> $$b(\theta) = E(\hat{\theta}) - \theta$$

> ### Definition 7.5
>
> The **minimum variance unbiased estimator (MVUE)** of a parameter θ is the estimator $\hat{\theta}$ that has the smallest variance of all unbiased estimators.

Mean squared error for $\hat{\theta}$: $E[(\hat{\theta} - \theta)^2]$

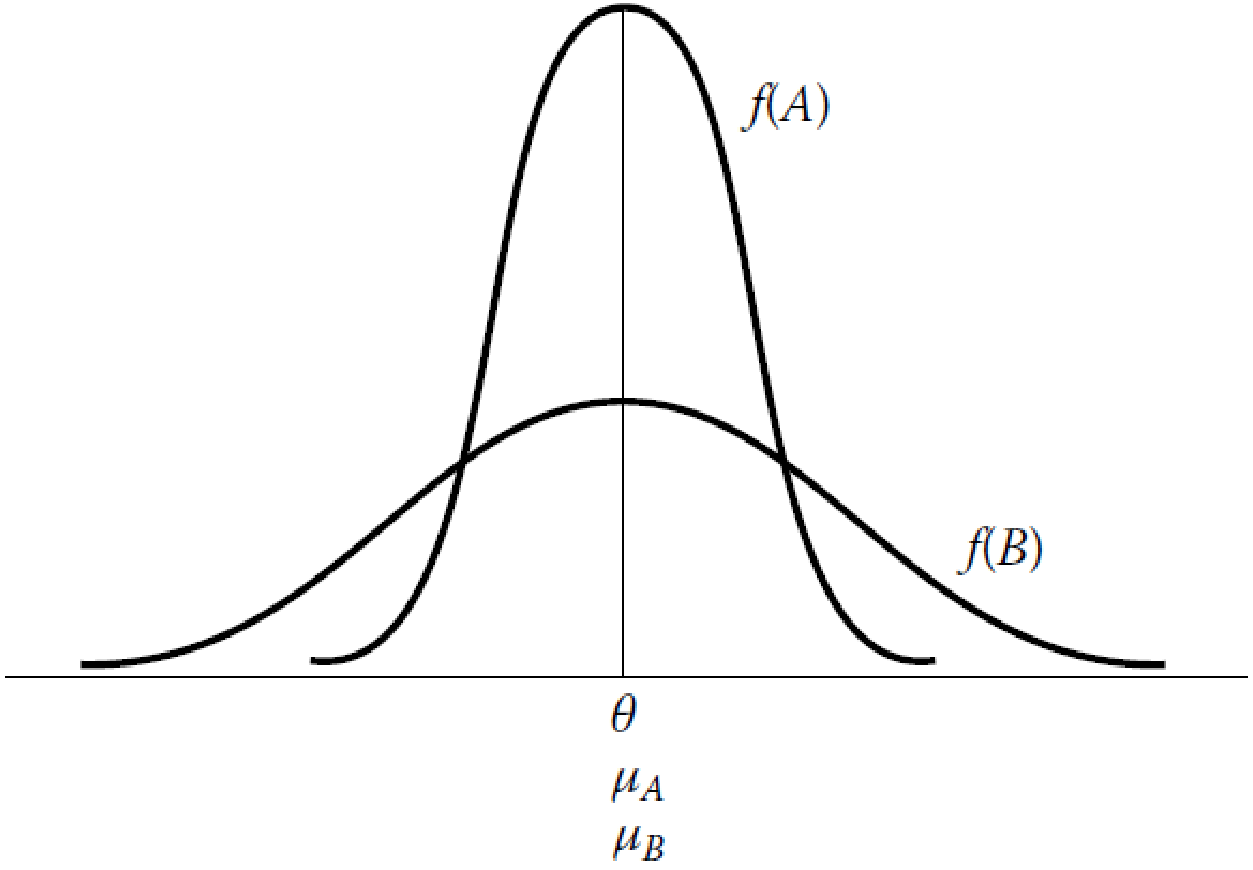
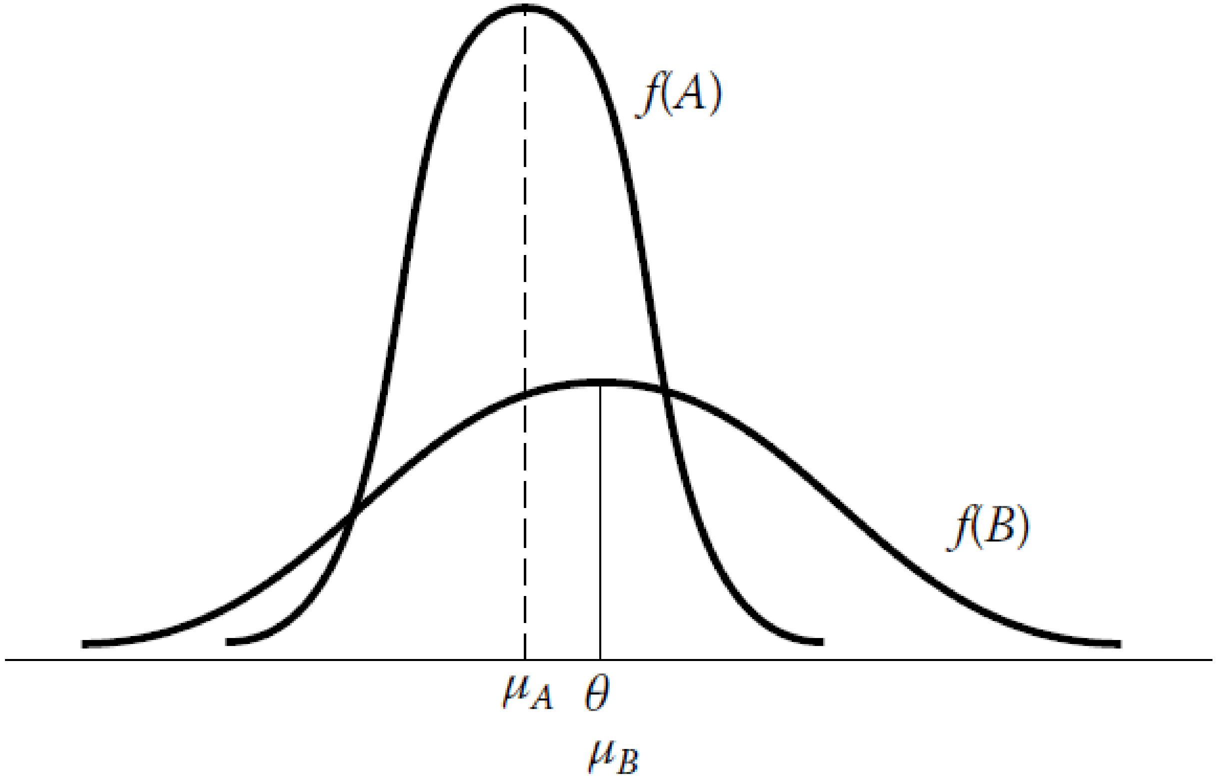


FIGURE 7.3

Sampling distributions for two unbiased estimators of θ with different variances

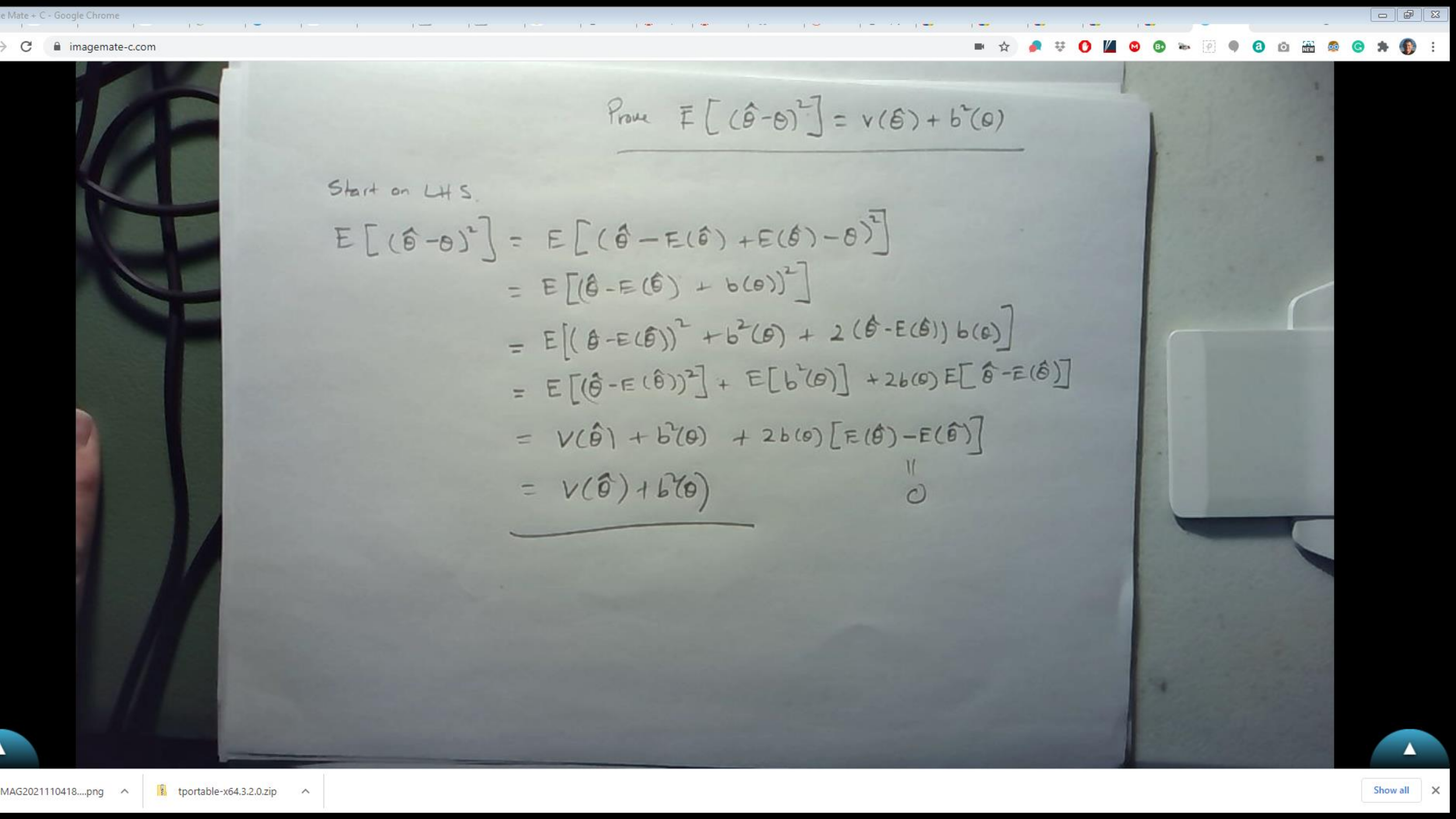
FIGURE 7.4

Sampling distributions of biased estimator *A* and MVUE *B*



It can be shown (proof omitted) that

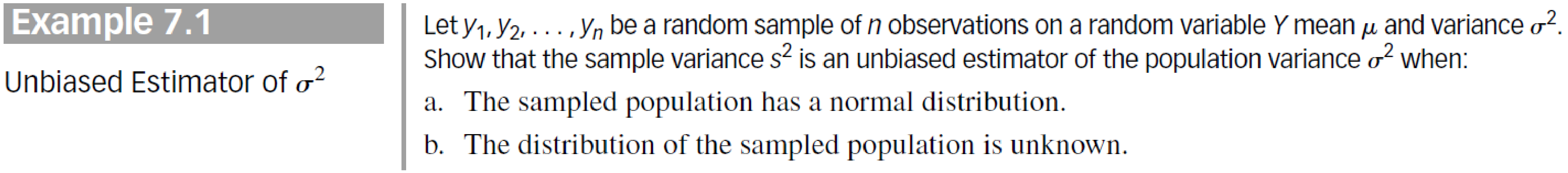
$$E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + b^2(\theta)$$



Prove $E[(\hat{\theta}-\theta)^2] = v(\hat{\theta}) + b^2(\theta)$

Start on LHS.

$$
\begin{aligned}
E[(\hat{\theta}-\theta)^2] &= E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}) + b(\theta))^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2 + b^2(\theta) + 2(\hat{\theta} - E(\hat{\theta}))\, b(\theta)] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2] + E[b^2(\theta)] + 2b(\theta) E[\hat{\theta} - E(\hat{\theta})] \\
&= V(\hat{\theta}) + b^2(\theta) + 2b(\theta) \underset{\overset{\shortparallel}{0}}{[E(\hat{\theta}) - E(\hat{\theta})]} \\
&= \underline{V(\hat{\theta}) + b^2(\theta)}
\end{aligned}
$$



Example 7.1

Unbiased Estimator of σ^2

Let $y_1, y_2, \ldots, y_n$ be a random sample of n observations on a random variable Y mean μ and variance σ^2. Show that the sample variance s^2 is an unbiased estimator of the population variance σ^2 when:

a. The sampled population has a normal distribution.

b. The distribution of the sampled population is unknown.

Must know this proof for the exam

$$s^2 = \frac{1}{(n - 1)}\left[\sum_{i=1}^{n} y_i^2 - \frac{(\sum y_i)^2}{n}\right] = \frac{1}{n - 1}\left[\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2\right]$$

From Theorem 4.4, $\sigma^2 = E(Y^2) - \mu^2$. Consequently, $E(Y^2) = \sigma^2 + \mu^2$ for a random variable Y. Since each Y value, $y_1, y_2, \ldots, y_n$, was randomly selected from a population with mean μ and variance σ^2, it follows that

$$E(y_i^2) = \sigma^2 + \mu^2(i = 1, 2, \ldots, n)$$

and

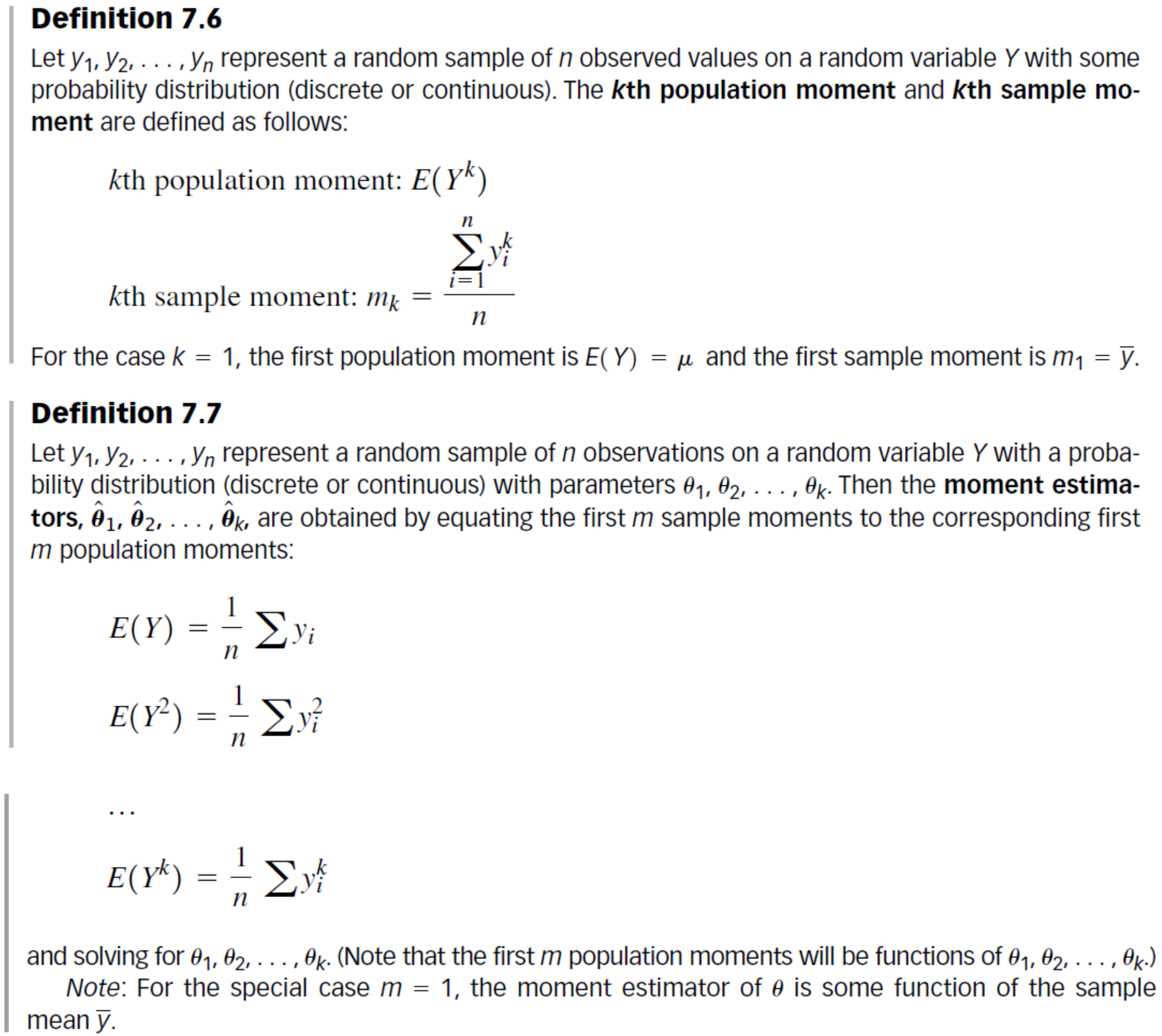
$$E(\bar{y}_i^2) = \sigma_{\bar{y}}^2 + (\mu_{\bar{y}})^2 = \sigma^2/n + \mu^2$$

Taking the expected value of s^2 and substituting these expressions, we obtain

$$\begin{aligned}
E(s^2) &= E\left\{\frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2\right]\right\} \\
&= \frac{1}{n-1}\left\{E\left[\sum_{i=1}^{n} y_i^2\right] - E[n(\bar{y})^2]\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E[y_i^2] - nE[(\bar{y})^2]\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\} \\
&= \frac{1}{n-1}[(n\sigma^2 + n\mu^2) - \sigma^2 - n\mu^2] \\
&= \frac{1}{n-1}[n\sigma^2 - \sigma^2] \\
&= \left(\frac{n-1}{n-1}\right)\sigma^2 = \sigma^2
\end{aligned}$$

This shows that, regardless of the nature of the sampled population, s^2 is an unbiased estimator of σ^2.

Method of moments to create point estimates



Definition 7.6

Let $y_1, y_2, \dots, y_n$ represent a random sample of n observed values on a random variable Y with some probability distribution (discrete or continuous). The ***k*th population moment** and ***k*th sample moment** are defined as follows:

$$k\text{th population moment: } E(Y^k)$$

$$k\text{th sample moment: } m_k = \frac{\sum_{i=1}^{n} y_i^k}{n}$$

For the case $k = 1$, the first population moment is $E(Y) = \mu$ and the first sample moment is $m_1 = \bar{y}$.

Definition 7.7

Let $y_1, y_2, \dots, y_n$ represent a random sample of n observations on a random variable Y with a probability distribution (discrete or continuous) with parameters $\theta_1, \theta_2, \dots, \theta_k$. Then the **moment estimators, $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$,** are obtained by equating the first m sample moments to the corresponding first m population moments:

$$E(Y) = \frac{1}{n}\sum y_i$$

$$E(Y^2) = \frac{1}{n}\sum y_i^2$$

$$\dots$$

$$E(Y^k) = \frac{1}{n}\sum y_i^k$$

and solving for $\theta_1, \theta_2, \dots, \theta_k$. (Note that the first m population moments will be functions of $\theta_1, \theta_2, \dots, \theta_k$.)

Note: For the special case $m = 1$, the moment estimator of θ is some function of the sample mean $\bar{y}$.

Example 7.2

Point Estimate of a Mean: Auditory Nerve Response Rate

Research in the *Journal of the Acoustic Society of America* found that the response rate Y of auditory nerve fibers in cats has an approximate Poisson distribution with unknown mean λ. Suppose the auditory nerve fiber response rate (recorded as number of spikes per 200 milliseconds of noise burst) was measured in each of a random sample of 10 cats. The data follow:

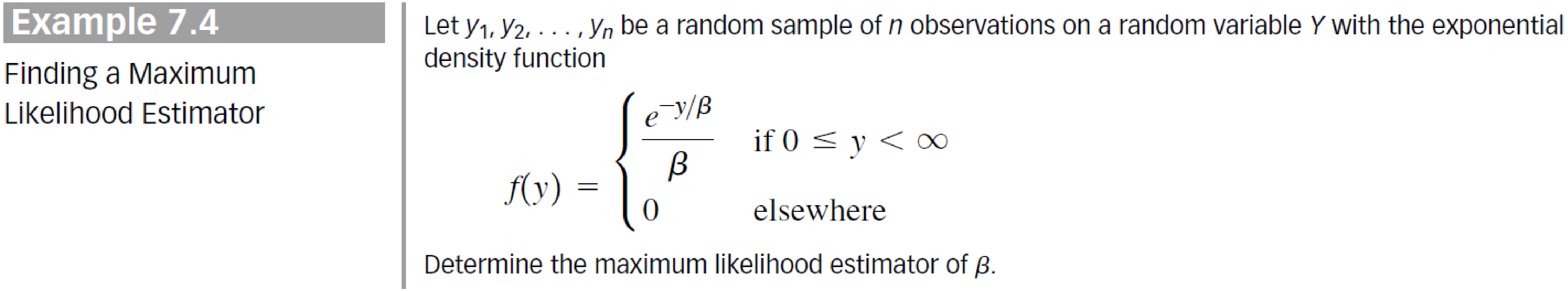
15.1	14.6	12.0	19.2	16.1	15.5	11.3	18.7	17.1	17.2

Calculate a point estimate for the mean response rate λ using the method of moments.

The following will be an example of MLE

Learn by doing!!

Go through all examples



Example 7.4

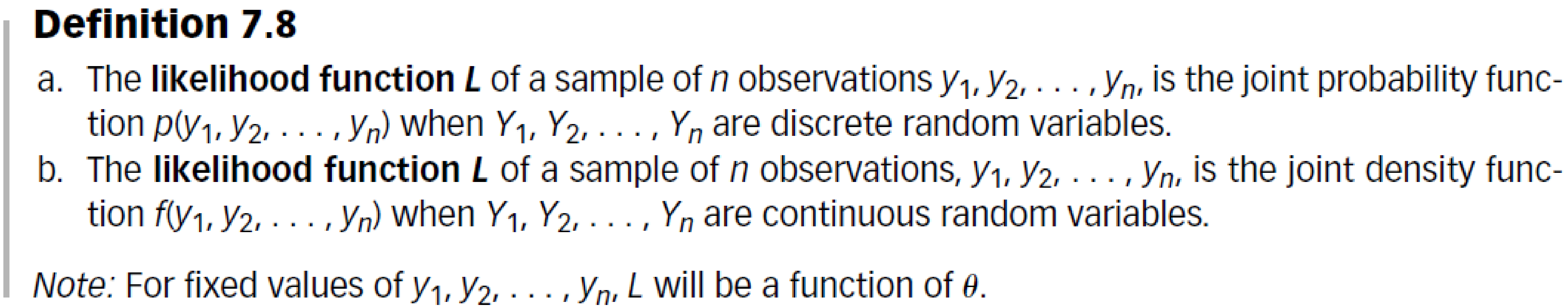
Finding a Maximum Likelihood Estimator

Let $y_1, y_2, \ldots, y_n$ be a random sample of n observations on a random variable Y with the exponential density function

$$f(y) = \begin{cases} \dfrac{e^{-y/\beta}}{\beta} & \text{if } 0 \leq y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

Determine the maximum likelihood estimator of β.

Maximum likelihood method



Definition 7.8

a. The **likelihood function** ***L*** of a sample of n observations $y_1, y_2, \ldots, y_n$, is the joint probability function $p(y_1, y_2, \ldots, y_n)$ when $Y_1, Y_2, \ldots, Y_n$ are discrete random variables.

b. The **likelihood function** ***L*** of a sample of n observations, $y_1, y_2, \ldots, y_n$, is the joint density function $f(y_1, y_2, \ldots, y_n)$ when $Y_1, Y_2, \ldots, Y_n$ are continuous random variables.

Note: For fixed values of $y_1, y_2, \ldots, y_n$, L will be a function of θ.

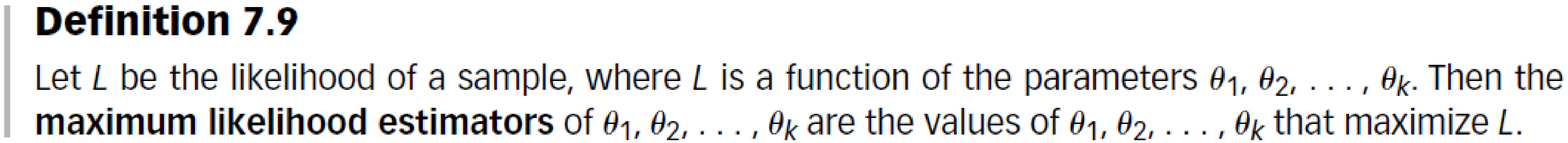
Important!!

THEOREM 7.1

a. Let $y_1, y_2, \ldots, y_n$ represent a random sample of n independent observations on a random variable Y. Then $L = p(y_1)p(y_2) \cdots p(y_n)$ when Y is a discrete random variable with probability distribution $p(y)$.

b. Let $y_1, y_2, \ldots, y_n$ represent a random sample of n independent observations on a random variable Y. Then $L = f(y_1)f(y_2) \cdots f(y_n)$ when Y is a continuous random variable with density function $f(y)$.

Maximum likelihood estimators



Definition 7.9

Let L be the likelihood of a sample, where L is a function of the parameters $\theta_1, \theta_2, \ldots, \theta_k$. Then the **maximum likelihood estimators** of $\theta_1, \theta_2, \ldots, \theta_k$ are the values of $\theta_1, \theta_2, \ldots, \theta_k$ that maximize L.

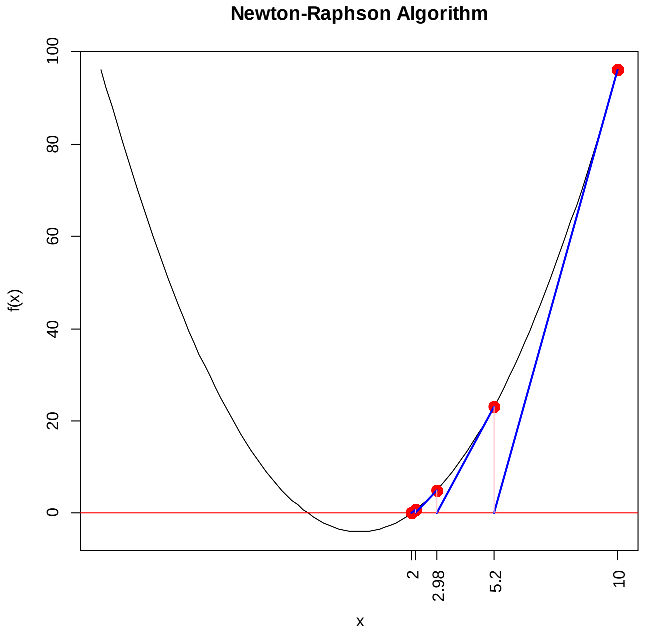
3 ways we will learn to make MLE’s

3 ways to perform MLE

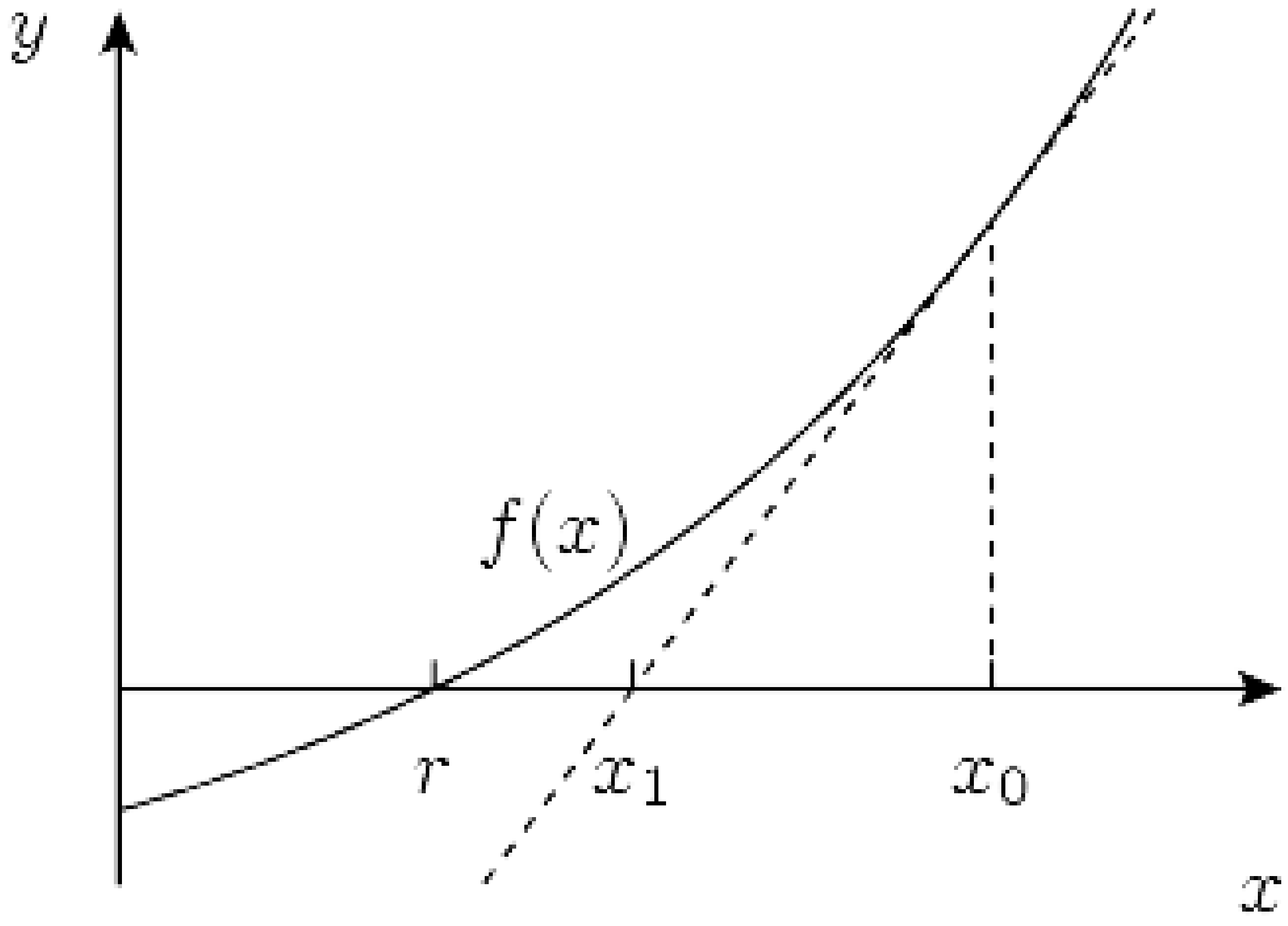
Grid approximation

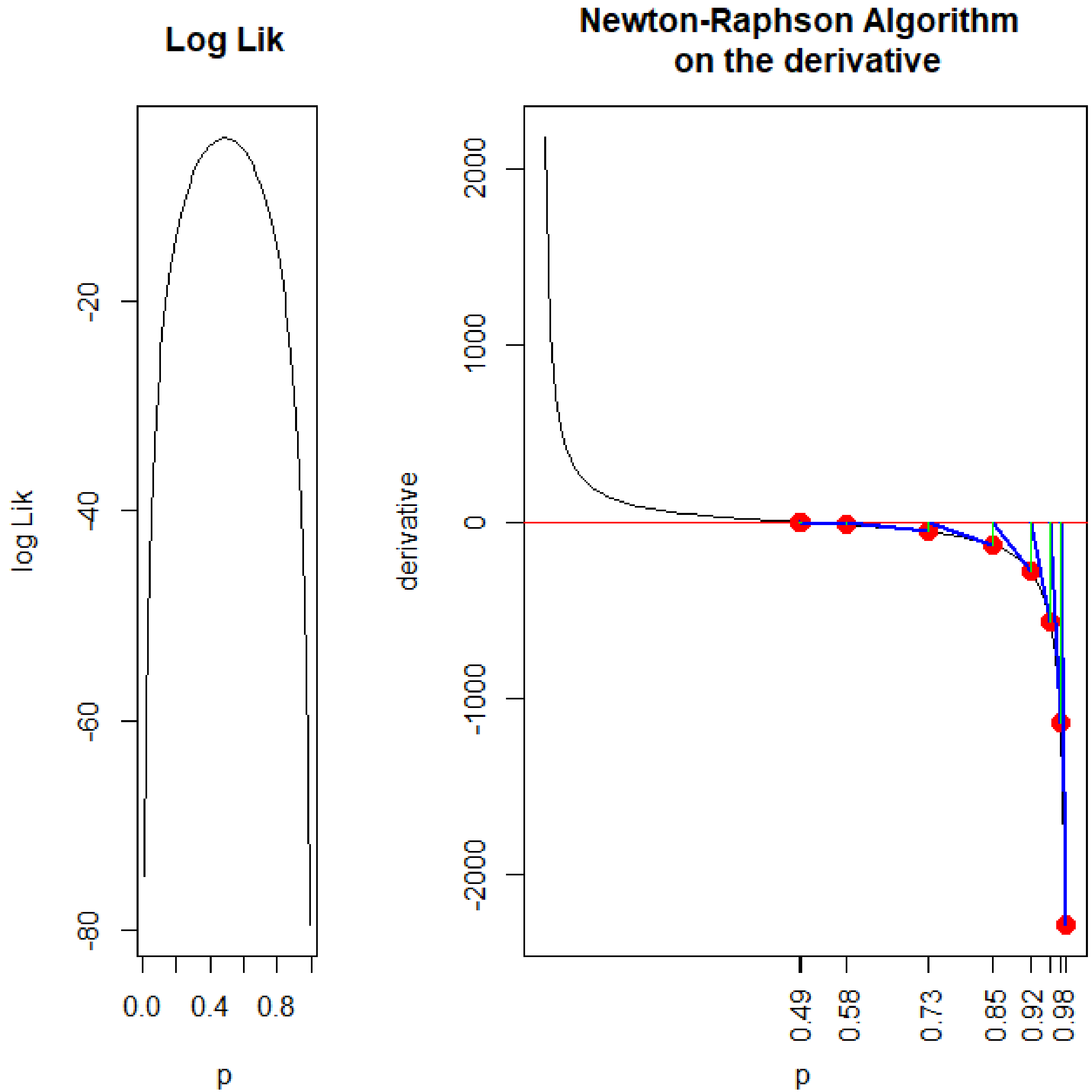
NR approximation

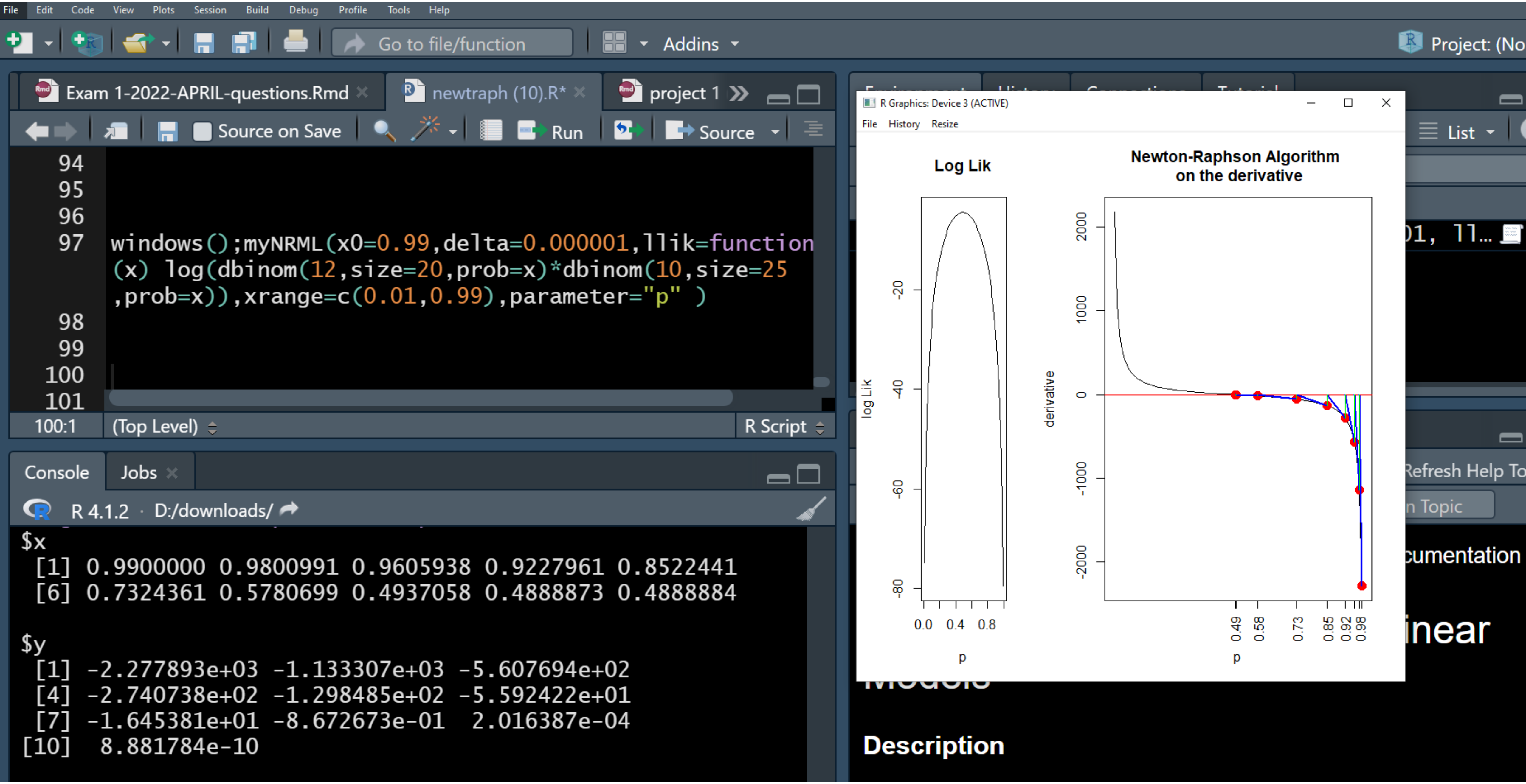
Precise analytical



Newton Raphson derivation



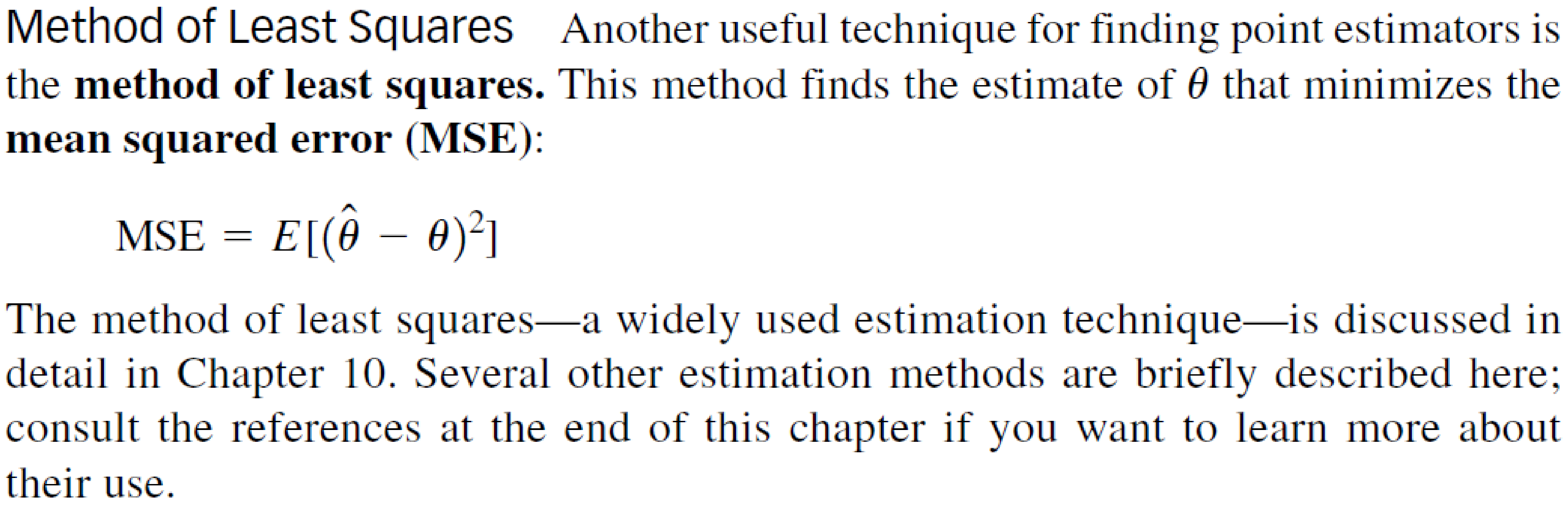

Log Lik
log Lik
-20
-40
-60
-80
0.0
0.4
0.8
p
Newton-Raphson Algorithm
on the derivative
derivative
2000
1000
0
-1000
-2000
0.49
0.58
0.73
0.85
0.92
0.98
p



```
windows();myNRML(x0=0.99,delta=0.000001,llik=function
(x) log(dbinom(12,size=20,prob=x)*dbinom(10,size=25
,prob=x)),xrange=c(0.01,0.99),parameter="p" )
```

```
$x
 [1] 0.9900000 0.9800991 0.9605938 0.9227961 0.8522441
 [6] 0.7324361 0.5780699 0.4937058 0.4888873 0.4888884

$y
 [1] -2.277893e+03 -1.133307e+03 -5.607694e+02
 [4] -2.740738e+02 -1.298485e+02 -5.592422e+01
 [7] -1.645381e+01 -8.672673e-01  2.016387e-04
[10]  8.881784e-10
```

Other methods



Method of Least Squares Another useful technique for finding point estimators is the **method of least squares.** This method finds the estimate of θ that minimizes the **mean squared error (MSE)**:

$$\text{MSE} = E[(\hat{\theta} - \theta)^2]$$

The method of least squares—a widely used estimation technique—is discussed in detail in Chapter 10. Several other estimation methods are briefly described here; consult the references at the end of this chapter if you want to learn more about their use.

Other point estimators not discussed in detail within this course

JACKNIFE

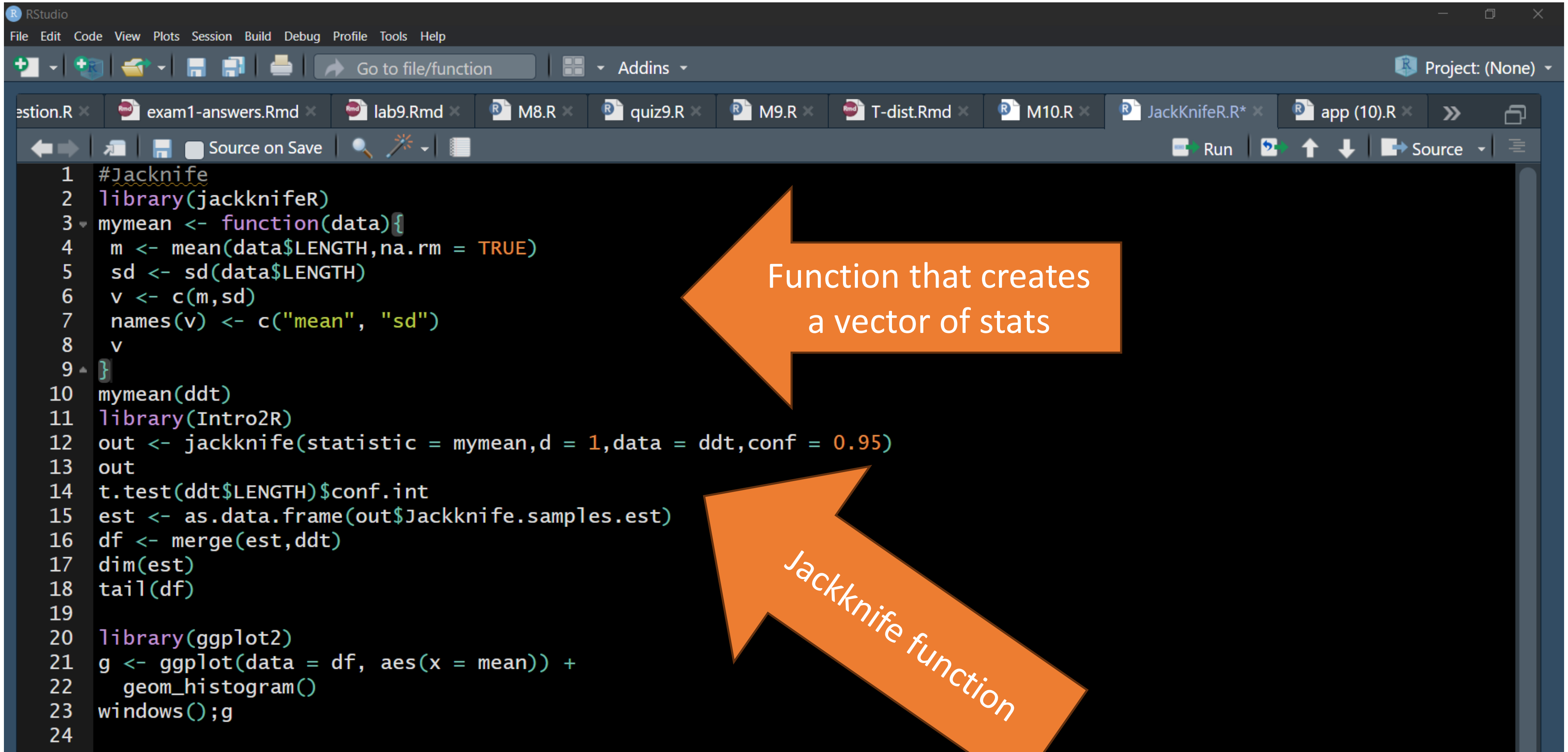
ROBUST

JACK-KNIFE method

JackknifeR

R package

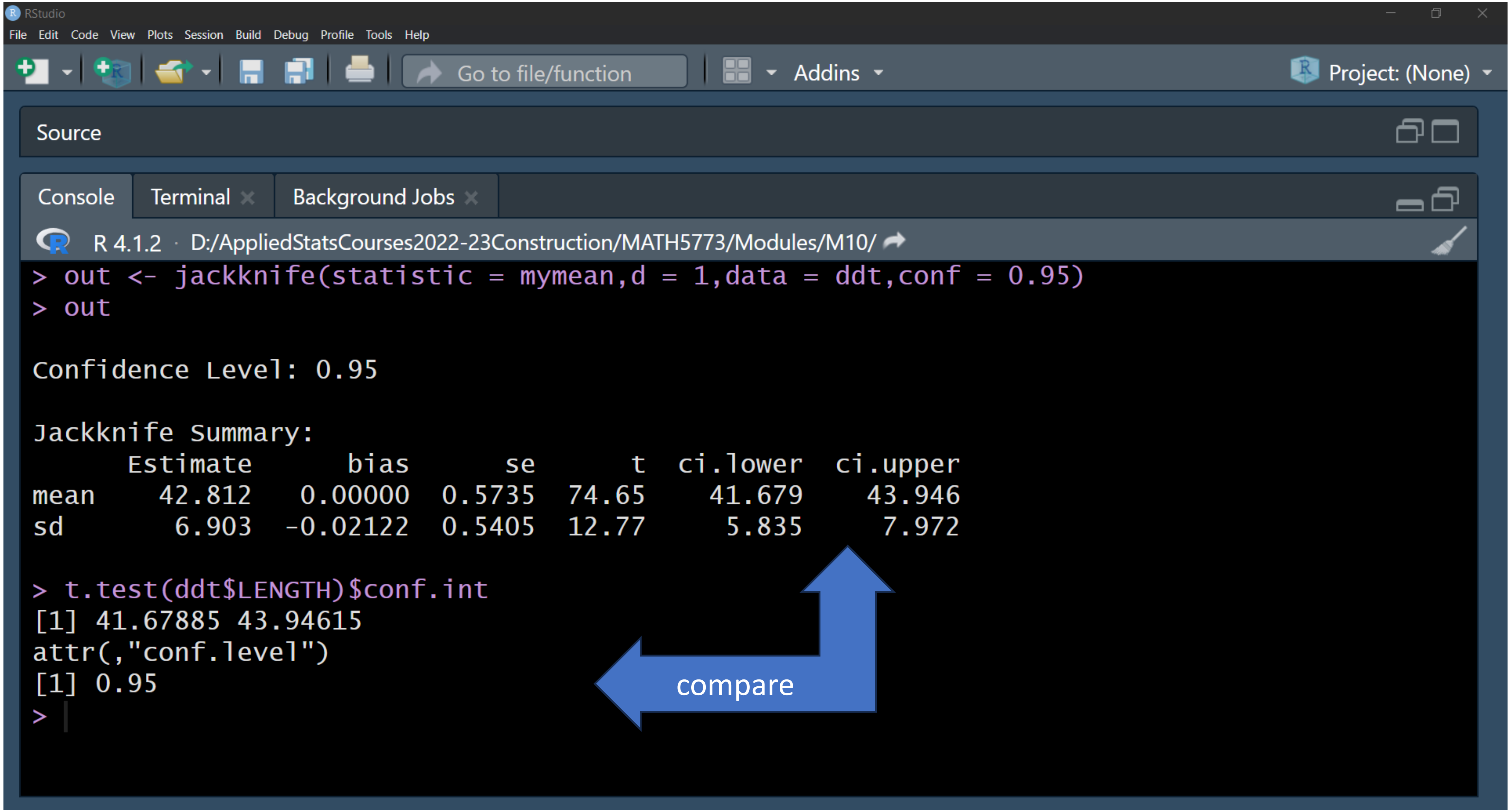
- Leave one(or more) datum out then estimate



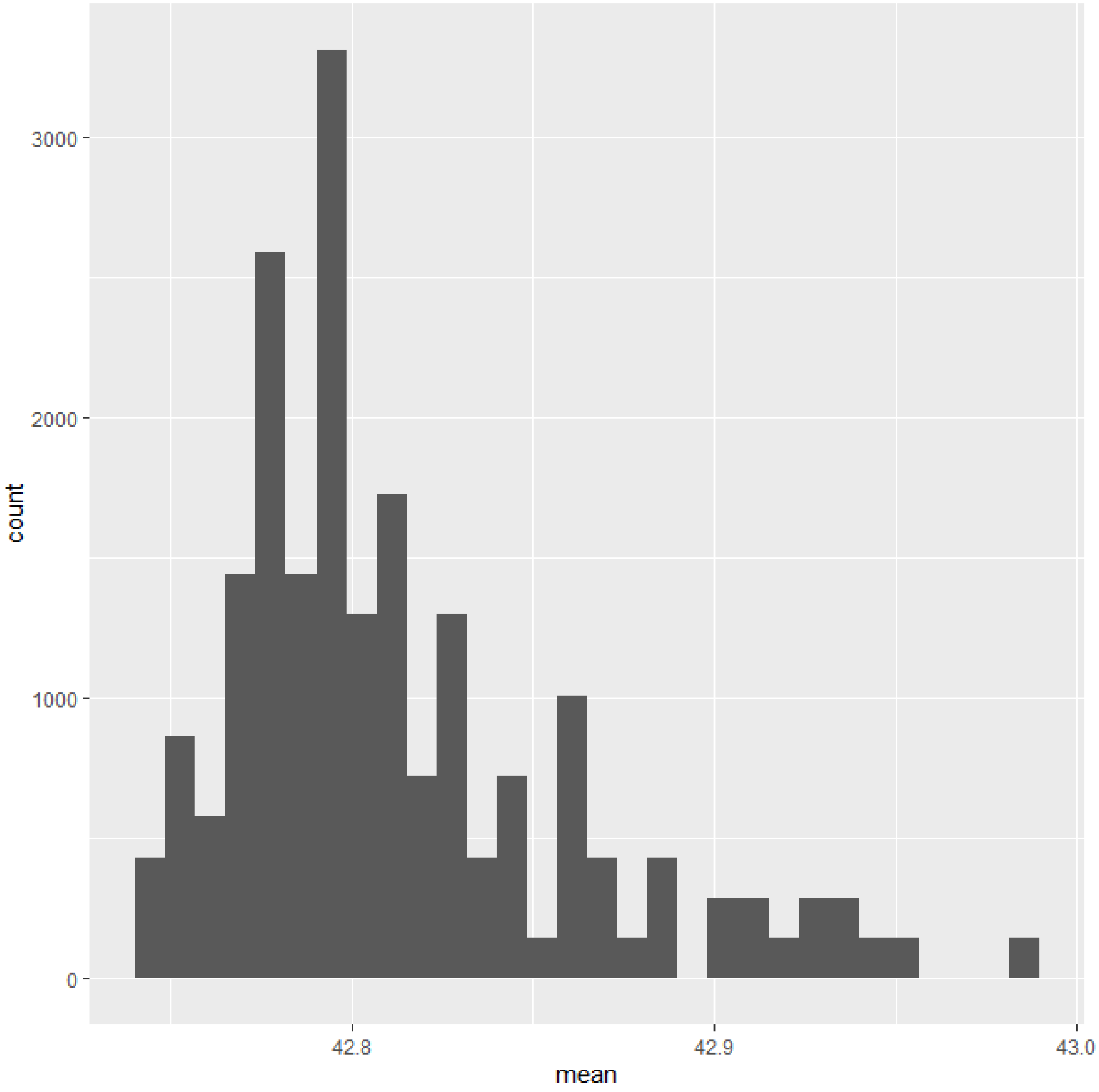
```r
#Jacknife
library(jackknifeR)
mymean <- function(data){
 m <- mean(data$LENGTH,na.rm = TRUE)
 sd <- sd(data$LENGTH)
 v <- c(m,sd)
 names(v) <- c("mean", "sd")
 v
}
mymean(ddt)
library(Intro2R)
out <- jackknife(statistic = mymean,d = 1,data = ddt,conf = 0.95)
out
t.test(ddt$LENGTH)$conf.int
est <- as.data.frame(out$Jackknife.samples.est)
df <- merge(est,ddt)
dim(est)
tail(df)

library(ggplot2)
g <- ggplot(data = df, aes(x = mean)) +
  geom_histogram()
windows();g

```

Function that creates a vector of stats

Jackknife function

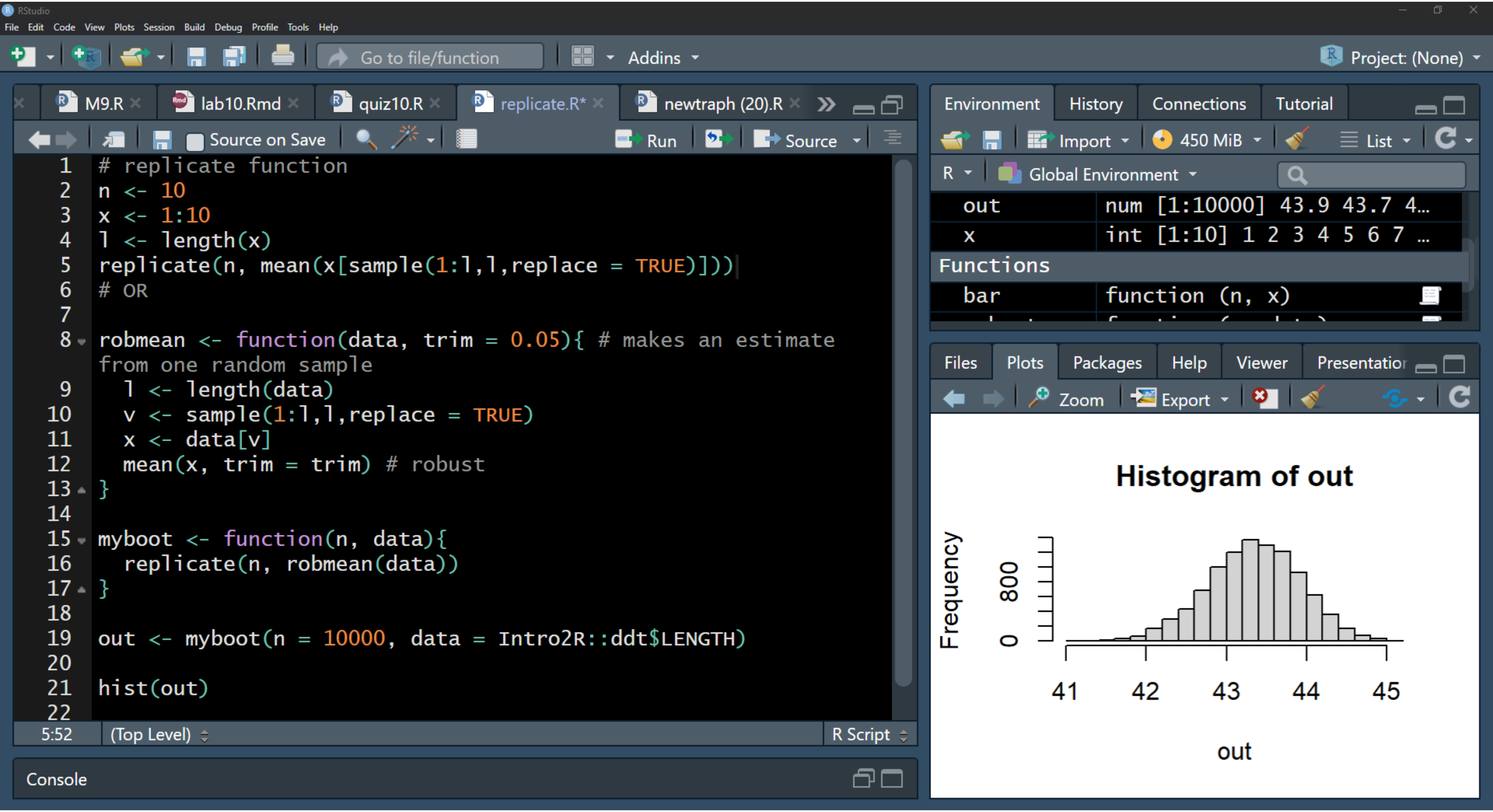


```
> out <- jackknife(statistic = mymean,d = 1,data = ddt,conf = 0.95)
> out

Confidence Level: 0.95

Jackknife Summary:
      Estimate     bias     se     t  ci.lower  ci.upper
mean    42.812  0.00000 0.5735 74.65    41.679    43.946
sd       6.903 -0.02122 0.5405 12.77     5.835     7.972

> t.test(ddt$LENGTH)$conf.int
[1] 41.67885 43.94615
attr(,"conf.level")
[1] 0.95
> 
```



3000
2000
1000
0
count
42.8
42.9
43.0
mean

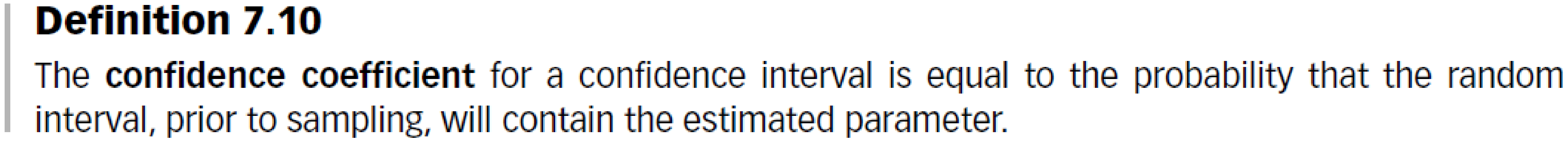
BOOTSTRAP



```r
# replicate function
n <- 10
x <- 1:10
l <- length(x)
replicate(n, mean(x[sample(1:l,l,replace = TRUE)]))
# OR

robmean <- function(data, trim = 0.05){ # makes an estimate from one random sample
  l <- length(data)
  v <- sample(1:l,l,replace = TRUE)
  x <- data[v]
  mean(x, trim = trim) # robust
}

myboot <- function(n, data){
  replicate(n, robmean(data))
}

out <- myboot(n = 10000, data = Intro2R::ddt$LENGTH)

hist(out)

```

```
out      num [1:10000] 43.9 43.7 4…
x        int [1:10] 1 2 3 4 5 6 7 …
Functions
bar      function (n, x)
```

Interval estimates



Definition 7.10

The **confidence coefficient** for a confidence interval is equal to the probability that the random interval, prior to sampling, will contain the estimated parameter.

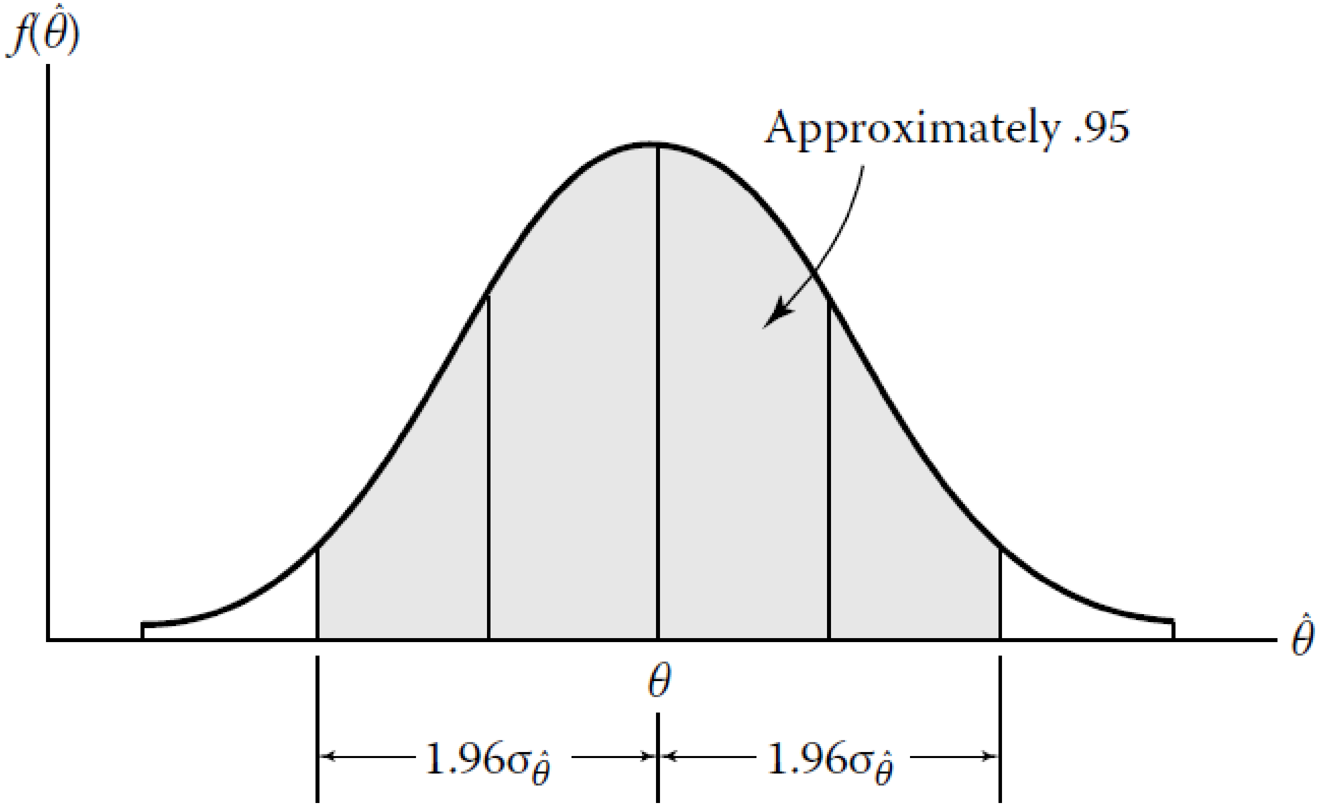
THEOREM 7.2

Let $\hat{\theta}$ be normally distributed for large samples with $E(\hat{\theta}) = \theta$ and standard error $\sigma_{\hat{\theta}}$. Then a $(1 - \alpha)100\%$ confidence interval for θ is

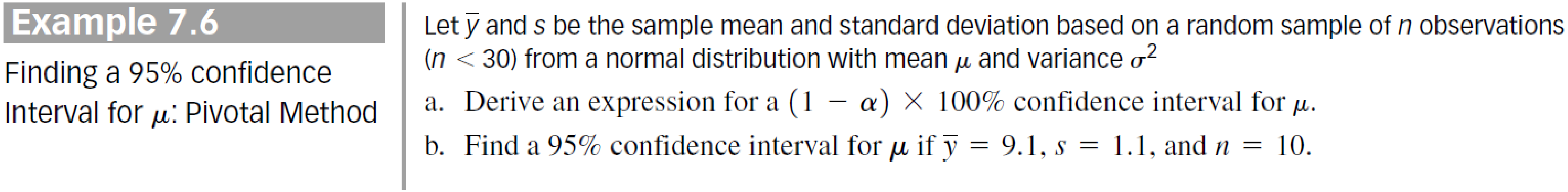
$$\hat{\theta} \pm (z_{\alpha/2})\sigma_{\hat{\theta}}$$

FIGURE 7.6

The sampling distribution of $\hat{\theta}$ for large samples



Pivotal method



Example 7.6

Finding a 95% confidence Interval for μ: Pivotal Method

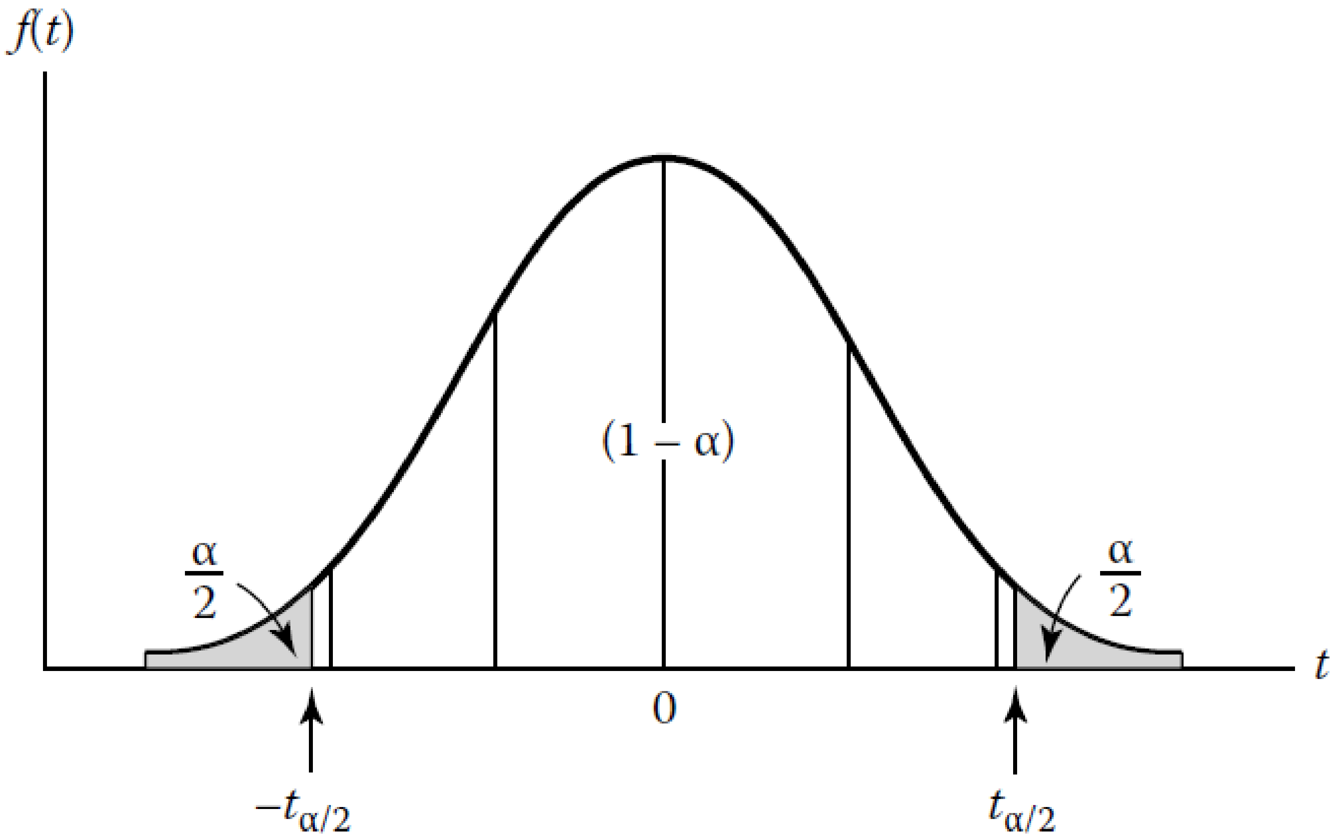
Let $\overline{y}$ and s be the sample mean and standard deviation based on a random sample of n observations ($n < 30$) from a normal distribution with mean μ and variance σ^2

a. Derive an expression for a $(1 - \alpha) \times 100\%$ confidence interval for μ.

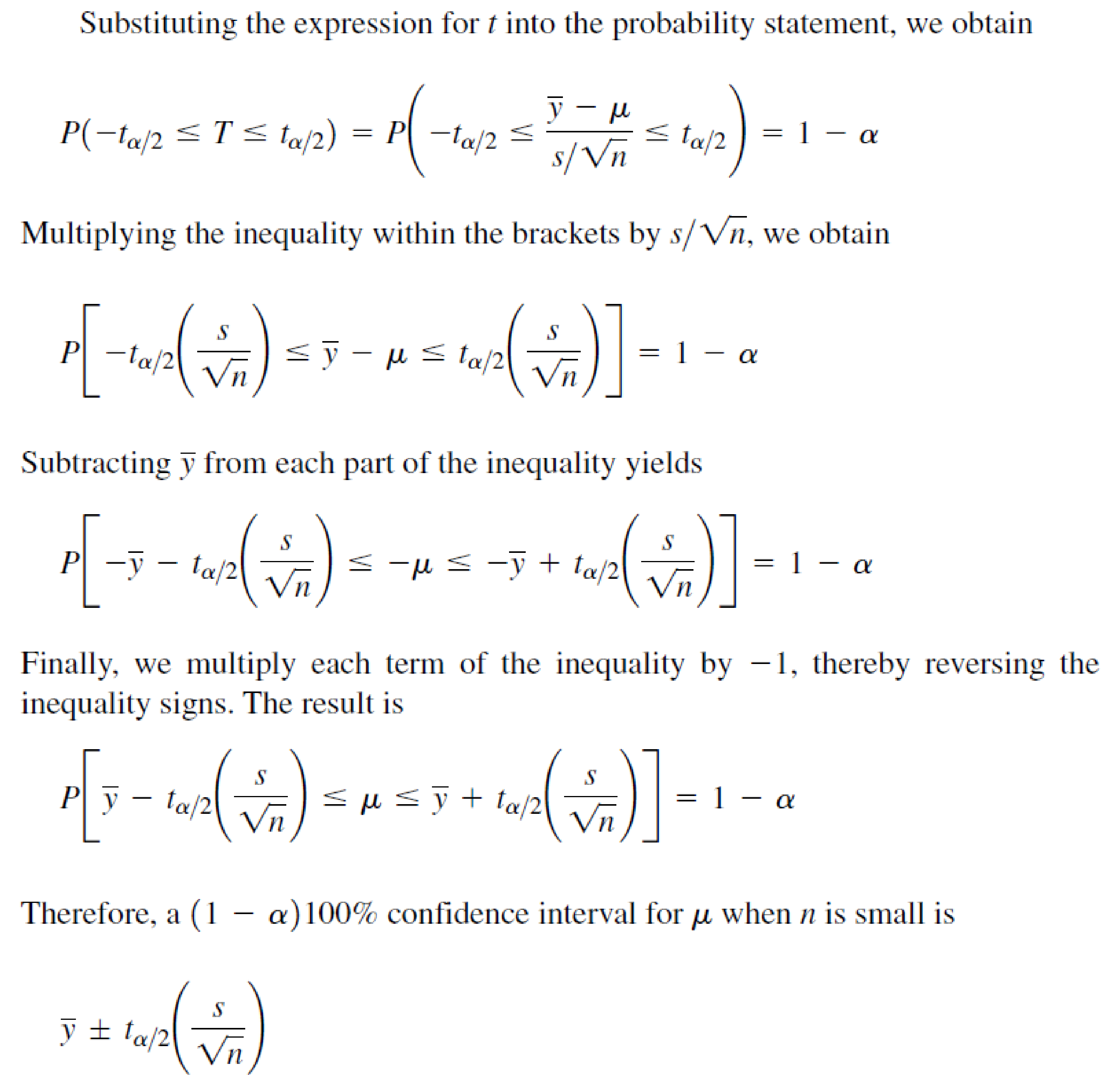
b. Find a 95% confidence interval for μ if $\overline{y} = 9.1$, $s = 1.1$, and $n = 10$.

FIGURE 7.7

The location of $t_{\alpha/2}$ and $-t_{\alpha/2}$ for a Student's T distribution



- Follow the proof



Substituting the expression for t into the probability statement, we obtain

$$P(-t_{\alpha/2} \le T \le t_{\alpha/2}) = P\left(-t_{\alpha/2} \le \frac{\bar{y} - \mu}{s/\sqrt{n}} \le t_{\alpha/2}\right) = 1 - \alpha$$

Multiplying the inequality within the brackets by $s/\sqrt{n}$, we obtain

$$P\left[-t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \le \bar{y} - \mu \le t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

Subtracting $\bar{y}$ from each part of the inequality yields

$$P\left[-\bar{y} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \le -\mu \le -\bar{y} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

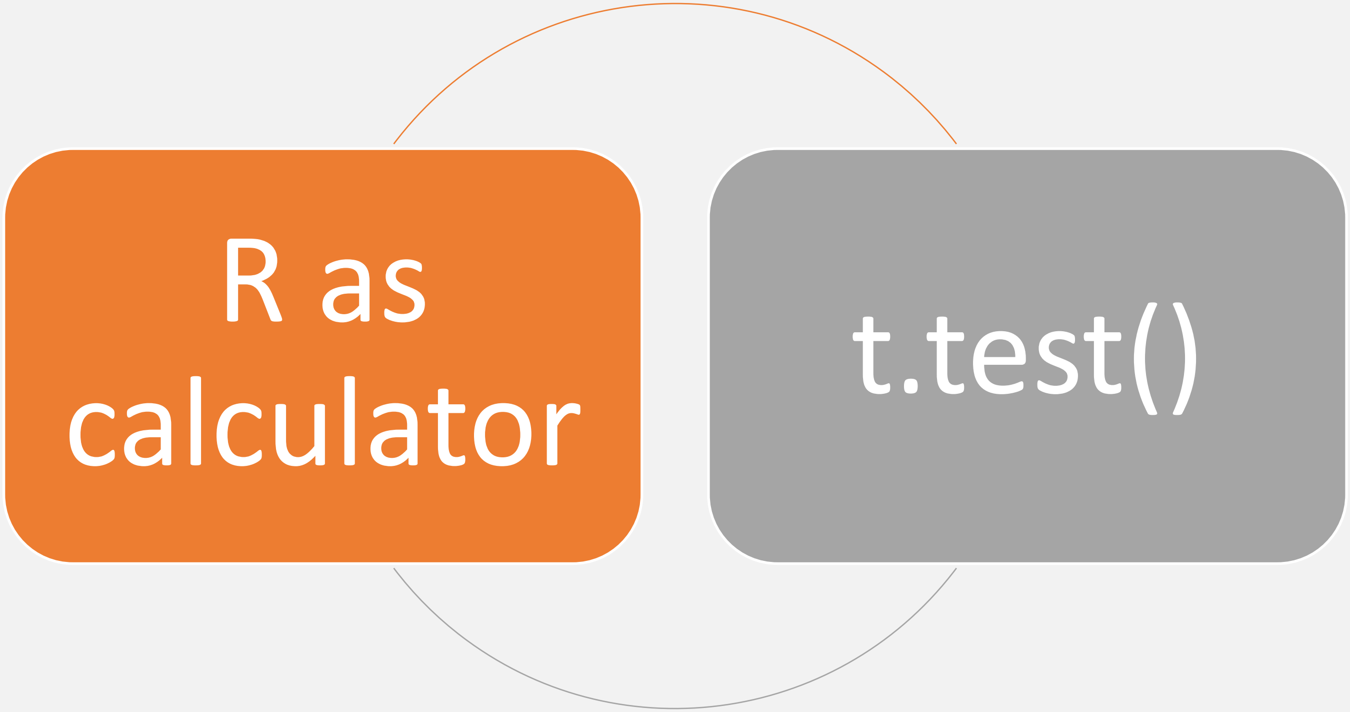
Finally, we multiply each term of the inequality by -1, thereby reversing the inequality signs. The result is

$$P\left[\bar{y} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{y} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

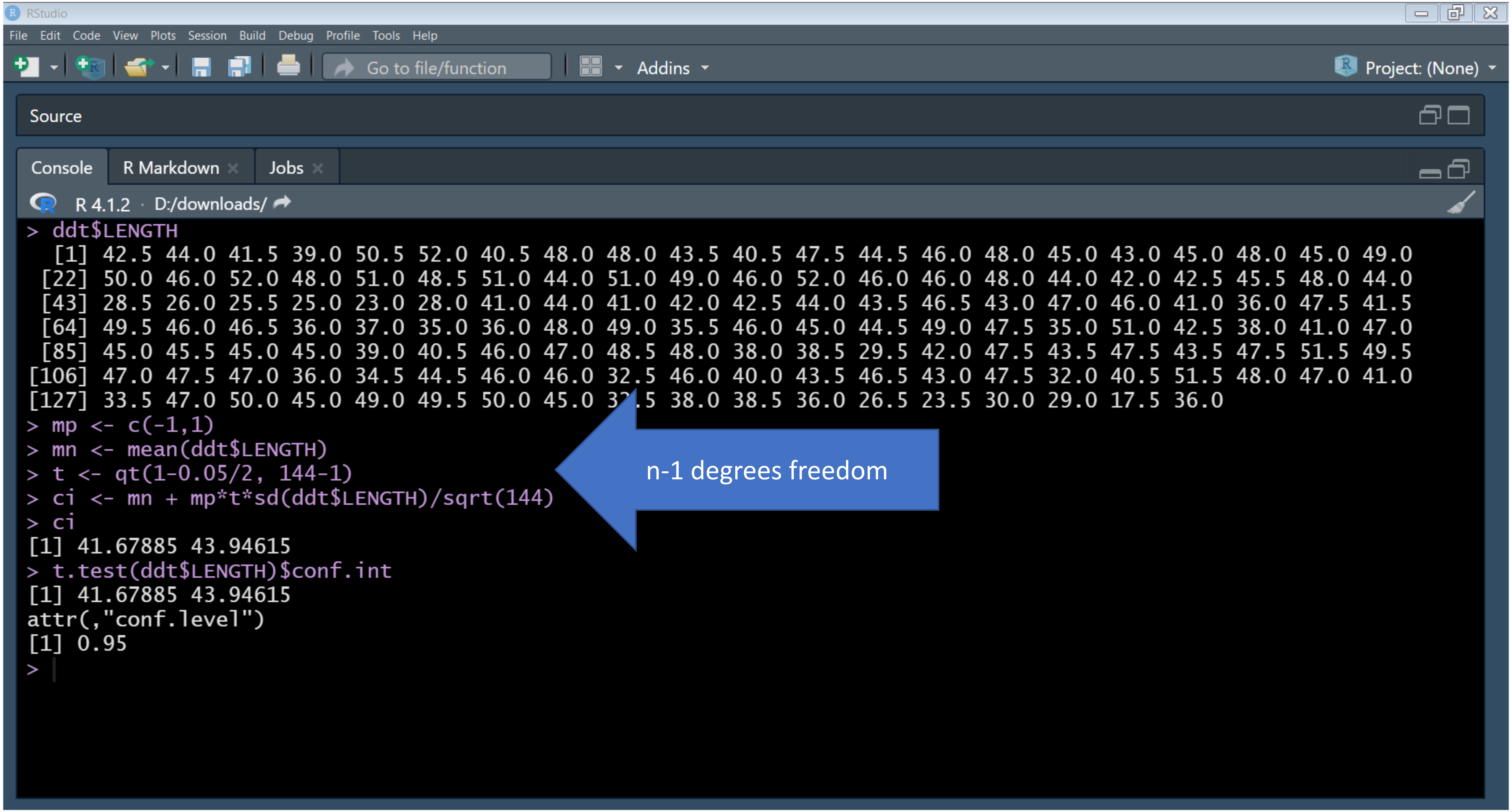
Therefore, a $(1 - \alpha)100\%$ confidence interval for μ when n is small is

$$\bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Two ways of obtaining ci’s


R as calculator

t.test()

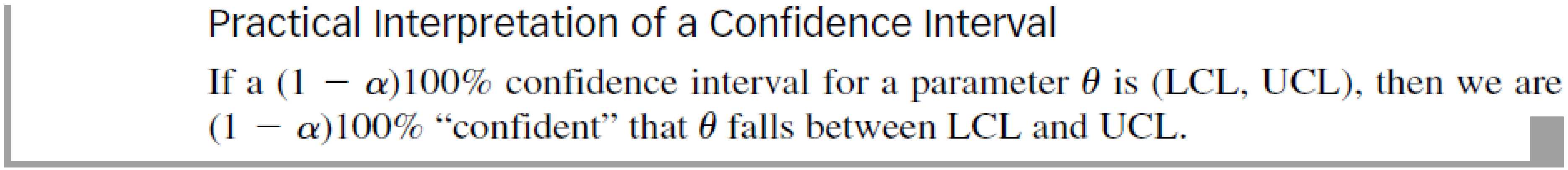


```
> ddt$LENGTH
  [1] 42.5 44.0 41.5 39.0 50.5 52.0 40.5 48.0 48.0 43.5 40.5 47.5 44.5 46.0 48.0 45.0 43.0 45.0 48.0 45.0 49.0
 [22] 50.0 46.0 52.0 48.0 51.0 48.5 51.0 44.0 51.0 49.0 46.0 52.0 46.0 46.0 48.0 44.0 42.0 42.5 45.5 48.0 44.0
 [43] 28.5 26.0 25.5 25.0 23.0 28.0 41.0 44.0 41.0 42.0 42.5 44.0 43.5 46.5 43.0 47.0 46.0 41.0 36.0 47.5 41.5
 [64] 49.5 46.0 46.5 36.0 37.0 35.0 36.0 48.0 49.0 35.5 46.0 45.0 44.5 49.0 47.5 35.0 51.0 42.5 38.0 41.0 47.0
 [85] 45.0 45.5 45.0 45.0 39.0 40.5 46.0 47.0 48.5 48.0 38.0 38.5 29.5 42.0 47.5 43.5 47.5 43.5 47.5 51.5 49.5
[106] 47.0 47.5 47.0 36.0 34.5 44.5 46.0 46.0 32.5 46.0 40.0 43.5 46.5 43.0 47.5 32.0 40.5 51.5 48.0 47.0 41.0
[127] 33.5 47.0 50.0 45.0 49.0 49.5 50.0 45.0 32.5 38.0 38.5 36.0 26.5 23.5 30.0 29.0 17.5 36.0
> mp <- c(-1,1)
> mn <- mean(ddt$LENGTH)
> t <- qt(1-0.05/2, 144-1)
> ci <- mn + mp*t*sd(ddt$LENGTH)/sqrt(144)
> ci
[1] 41.67885 43.94615
> t.test(ddt$LENGTH)$conf.int
[1] 41.67885 43.94615
attr(,"conf.level")
[1] 0.95
> 
```

n-1 degrees freedom

Two interpretations of ci’s

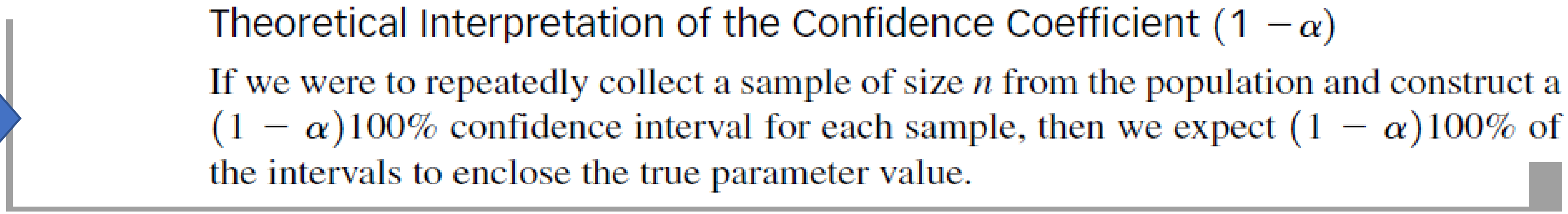
1



Practical Interpretation of a Confidence Interval

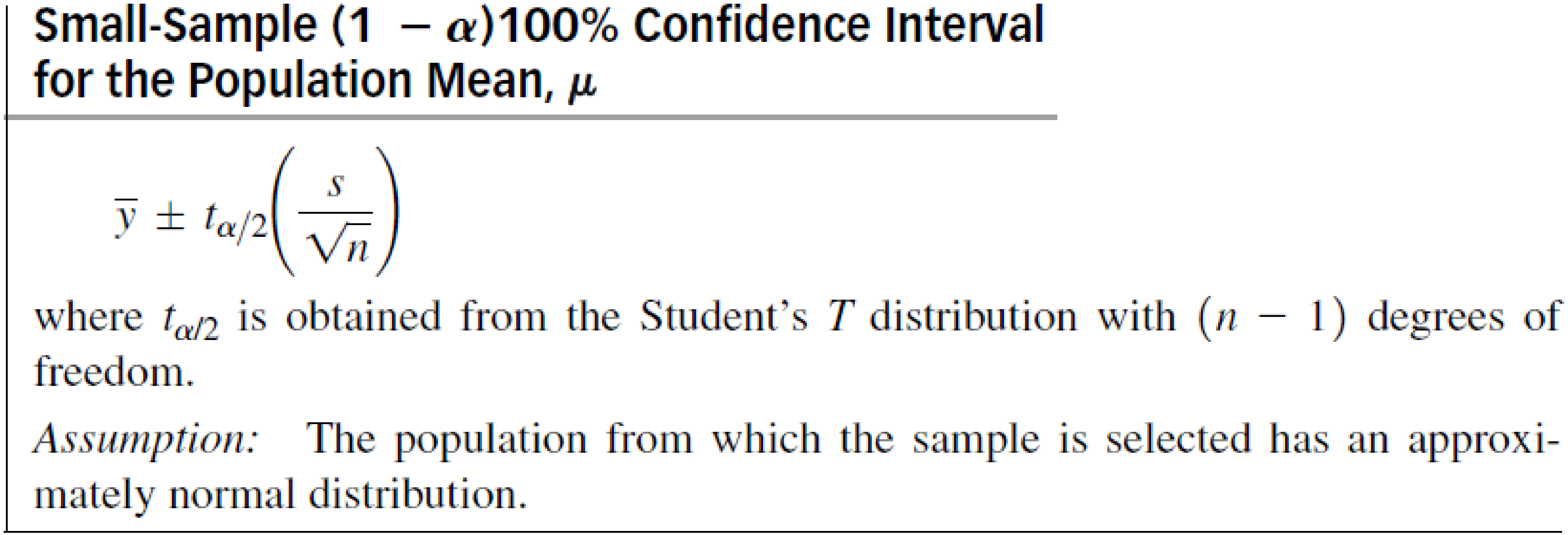
If a $(1 - \alpha)100\%$ confidence interval for a parameter θ is (LCL, UCL), then we are $(1 - \alpha)100\%$ “confident” that θ falls between LCL and UCL.

2



Theoretical Interpretation of the Confidence Coefficient $(1 - \alpha)$

If we were to repeatedly collect a sample of size n from the population and construct a $(1 - \alpha)100\%$ confidence interval for each sample, then we expect $(1 - \alpha)100\%$ of the intervals to enclose the true parameter value.



Small-Sample $(1 - \alpha)100\%$ Confidence Interval for the Population Mean, μ

$$\bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

where $t_{\alpha/2}$ is obtained from the Student's T distribution with $(n - 1)$ degrees of freedom.

Assumption: The population from which the sample is selected has an approximately normal distribution.

Small-Sample $(1-\alpha)100\%$ Confidence Interval for $(\mu_1 - \mu_2)$: Independent Samples and $\sigma_1^2 = \sigma_2^2$

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and the value of $t_{\alpha/2}$ is based on $(n_1 + n_2 - 2)$ degrees of freedom.

Assumptions:
1. Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.
2. The variances σ_1^2 and σ_2^2 of the two populations are equal.
3. The random samples are selected in an independent manner from the two populations.

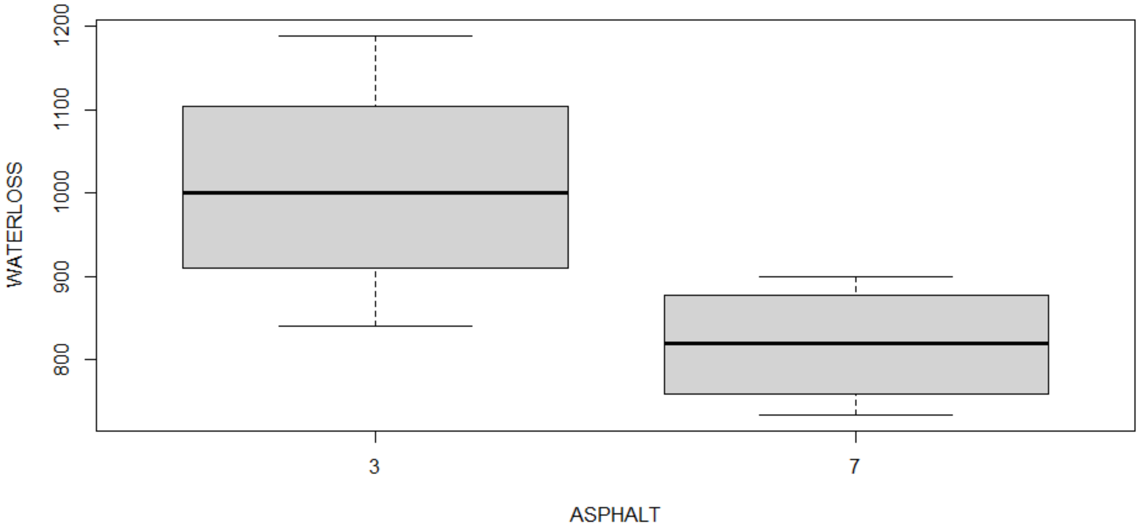
CONCRETE

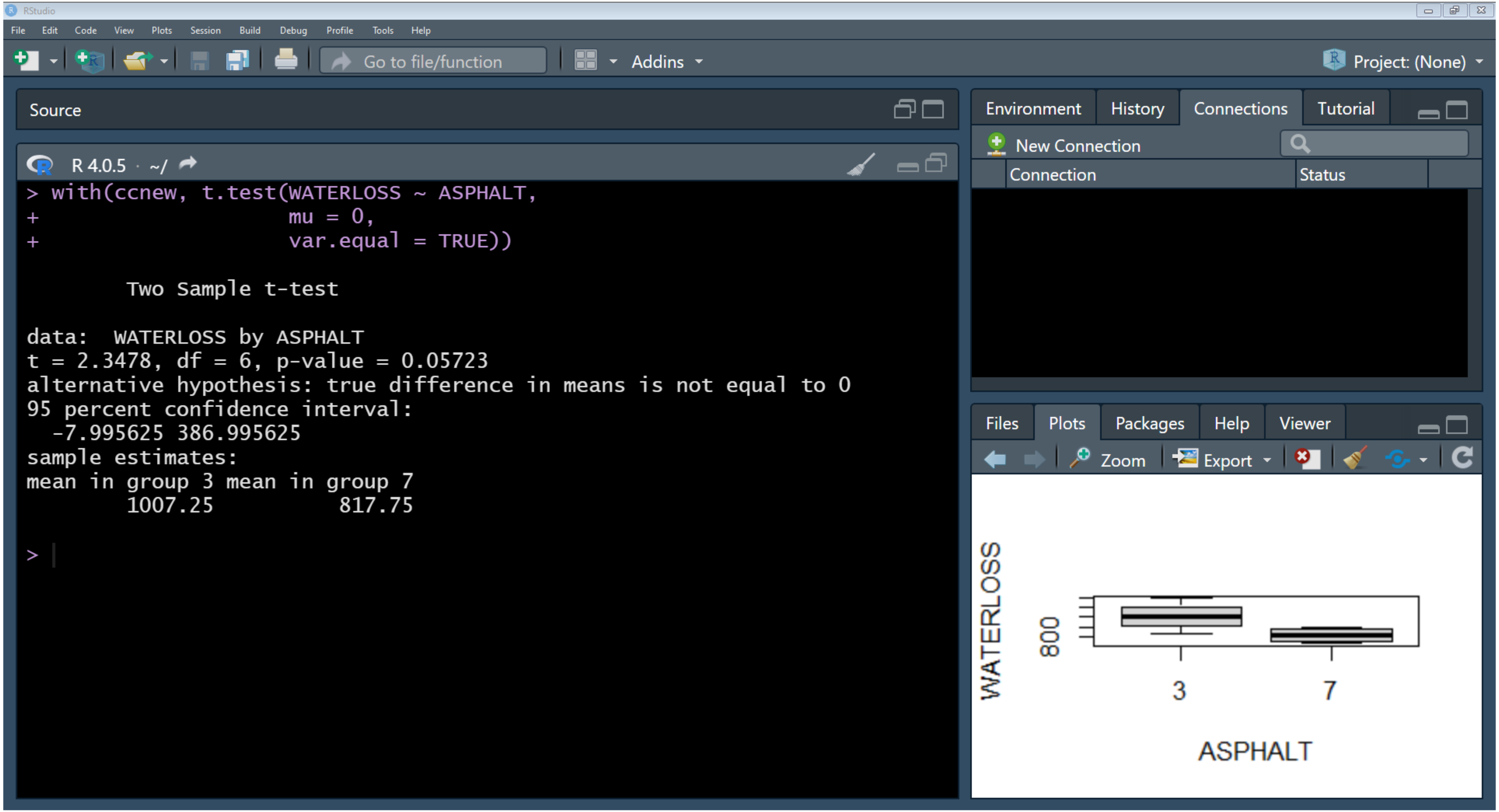
TABLE 7.4 Permeability Measurements for 3% and 7% Asphalt Concrete, Example 7.10

Asphalt Content	3%	1,189	840	1,020	980
	7%	853	900	733	785

Source: Woelfl, G., Wei, I., Faulstich, C., and Litwack, H. “Laboratory testing of asphalt concrete for porous pavements.” *Journal of Testing and Evaluation*, Vol. 9, No. 4, July 1981, pp. 175–181. Copyright American Society for Testing and Materials.

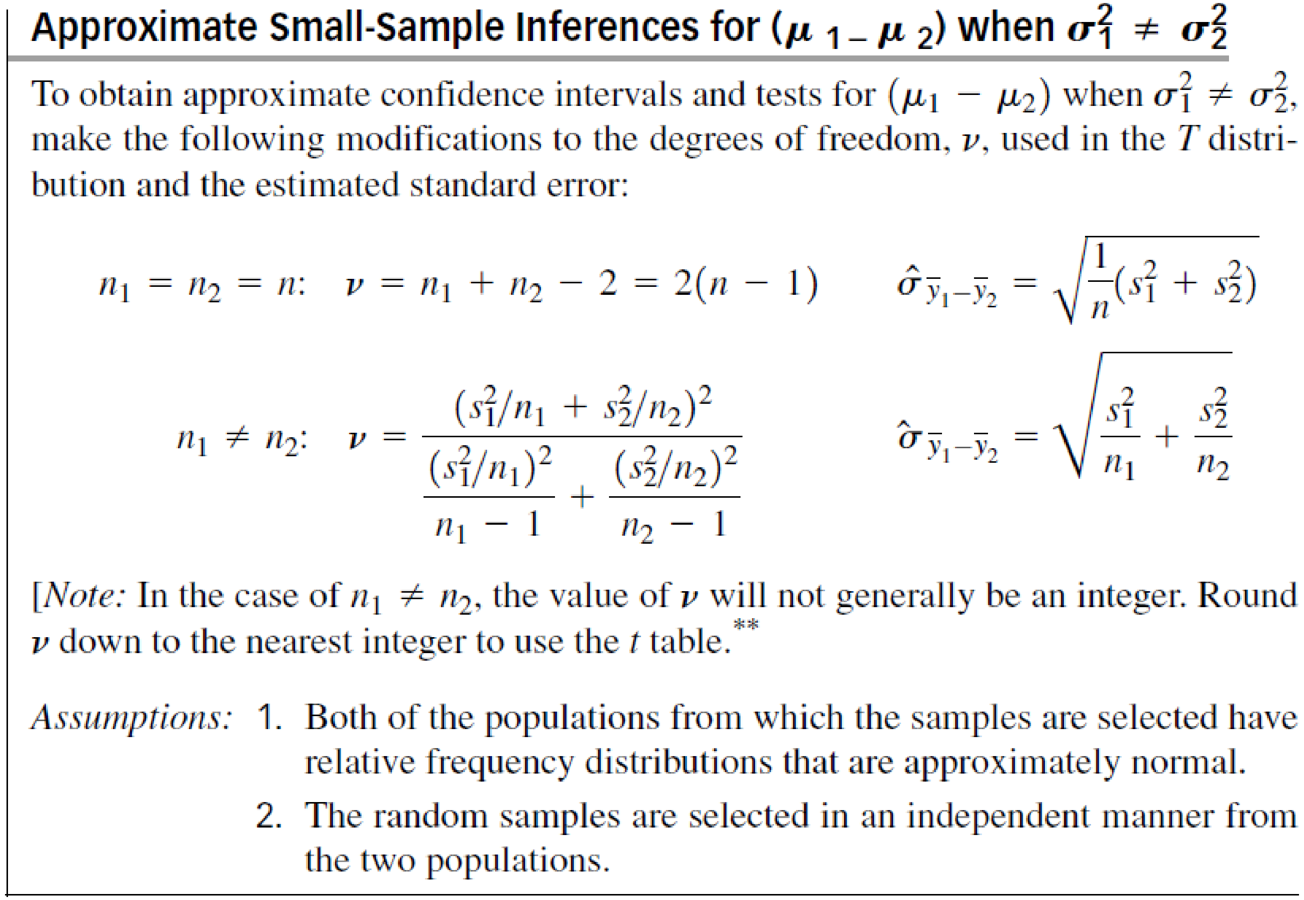
```
names(cc)
head(cc)
ccnew<-within(cc, ASPHALT <- factor(ASPHALT))
with(ccnew, boxplot(WATERLOSS ~ ASPHALT))
with(ccnew,var.test(WATERLOSS~ASPHALT))
with(ccnew, t.test(WATERLOSS ~ ASPHALT,
                   mu = 0,
                   var.equal = TRUE))
```





```
> with(ccnew, t.test(WATERLOSS ~ ASPHALT,
+                    mu = 0,
+                    var.equal = TRUE))

        Two Sample t-test

data:  WATERLOSS by ASPHALT
t = 2.3478, df = 6, p-value = 0.05723
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -7.995625 386.995625
sample estimates:
mean in group 3 mean in group 7
        1007.25          817.75

> 
```

Welch 2 sample t-test



Approximate Small-Sample Inferences for $(\mu_1 - \mu_2)$ when $\sigma_1^2 \neq \sigma_2^2$

To obtain approximate confidence intervals and tests for $(\mu_1 - \mu_2)$ when $\sigma_1^2 \neq \sigma_2^2$, make the following modifications to the degrees of freedom, ν, used in the T distribution and the estimated standard error:

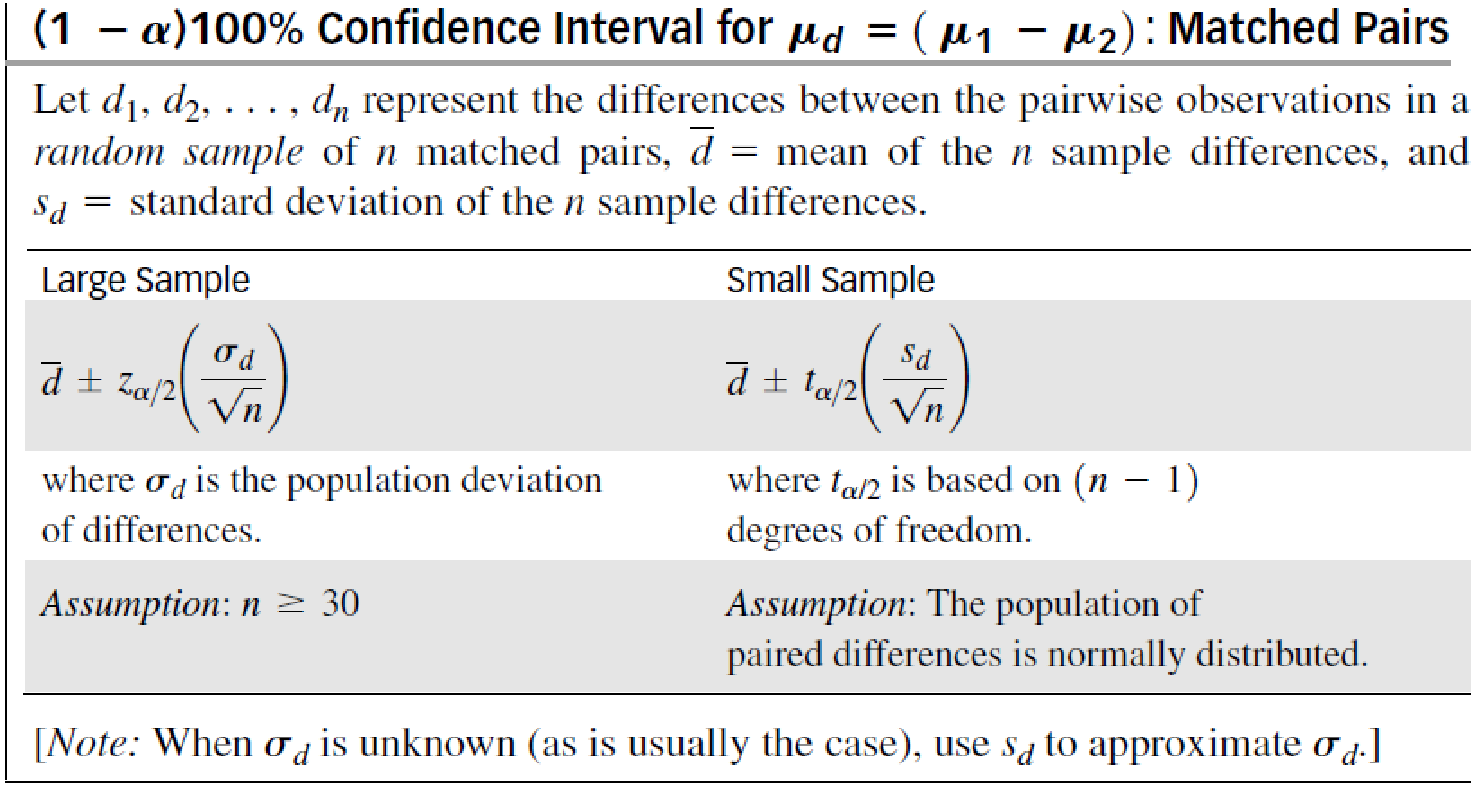
$$n_1 = n_2 = n: \quad \nu = n_1 + n_2 - 2 = 2(n-1) \qquad \hat{\sigma}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{1}{n}(s_1^2 + s_2^2)}$$

$$n_1 \neq n_2: \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}} \qquad \hat{\sigma}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

[*Note:* In the case of $n_1 \neq n_2$, the value of ν will not generally be an integer. Round ν down to the nearest integer to use the t table.**

Assumptions:
1. Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.
2. The random samples are selected in an independent manner from the two populations.

paired samples



$(1 - \alpha)100\%$ Confidence Interval for $\mu_d = (\mu_1 - \mu_2)$: Matched Pairs

Let $d_1, d_2, \ldots, d_n$ represent the differences between the pairwise observations in a *random sample* of n matched pairs, $\bar{d}$ = mean of the n sample differences, and s_d = standard deviation of the n sample differences.

Large Sample	Small Sample
$\bar{d} \pm z_{\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right)$	$\bar{d} \pm t_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right)$
where σ_d is the population deviation of differences.	where $t_{\alpha/2}$ is based on $(n - 1)$ degrees of freedom.
Assumption: $n \geq 30$	*Assumption*: The population of paired differences is normally distributed.

[*Note:* When σ_d is unknown (as is usually the case), use s_d to approximate σ_d.]

Example 7.11

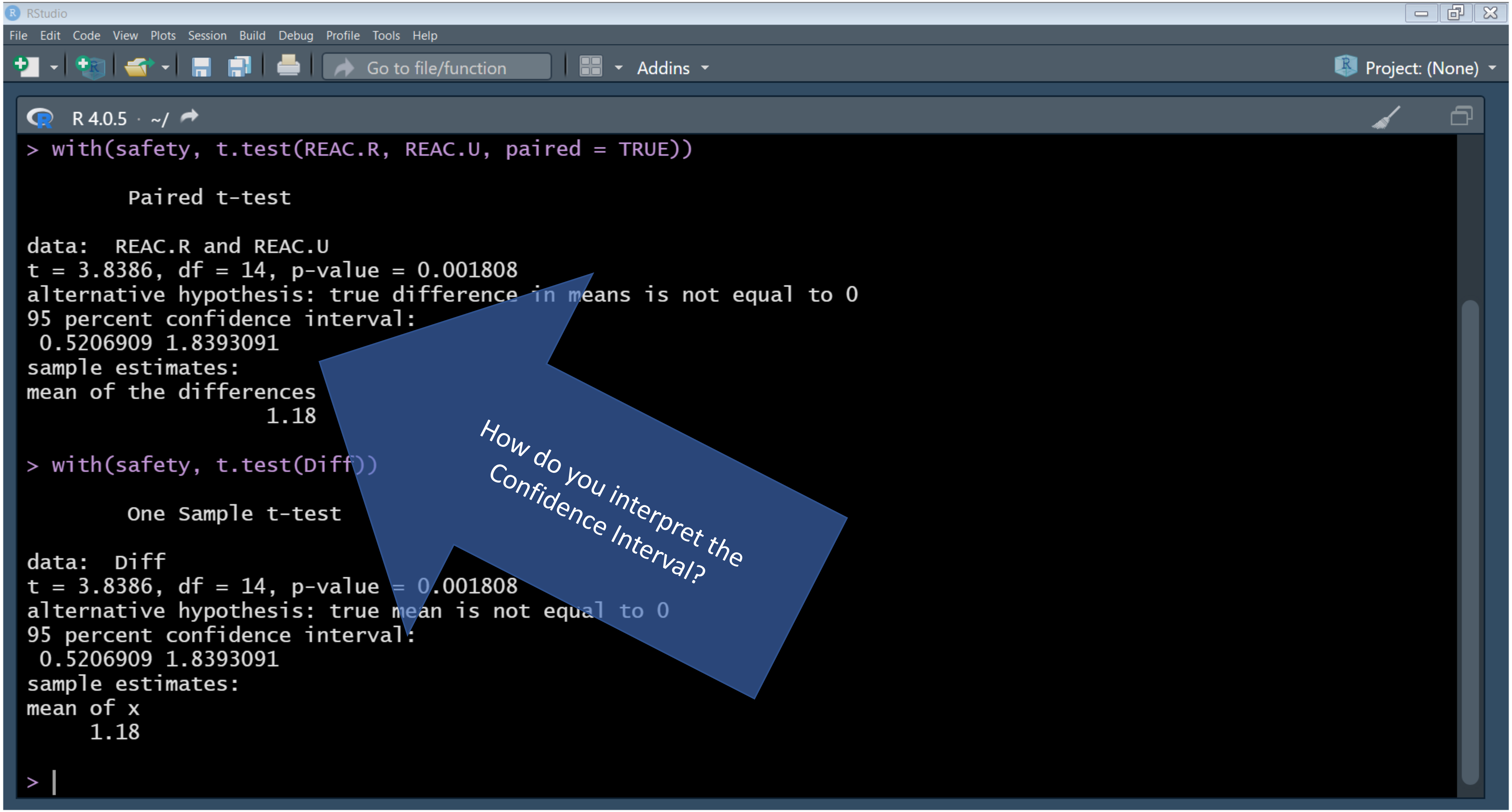
Matched Pairs Confidence Interval — Driver Reaction Time

A federal traffic safety engineer wants to ascertain the effect of wearing safety devices (shoulder harnesses, seat belts) on reaction times to peripheral stimuli. A study was designed as follows: A random sample of 15 student drivers was selected from students enrolled in a driver-education program. Each driver performed a simulated driving task that allowed reaction times to be recorded under two conditions, wearing a safety device (restrained condition) and wearing no safety device (unrestrained condition). Thus, each student driver received two reaction-time scores, one for the restrained condition and one for the unrestrained condition. The data (in hundredths of a second) are shown in Table 7.7 and saved in the **SAFETY** file. Find and interpret a

SAFETY

TABLE 7.7 Reaction time data for Example 7.11

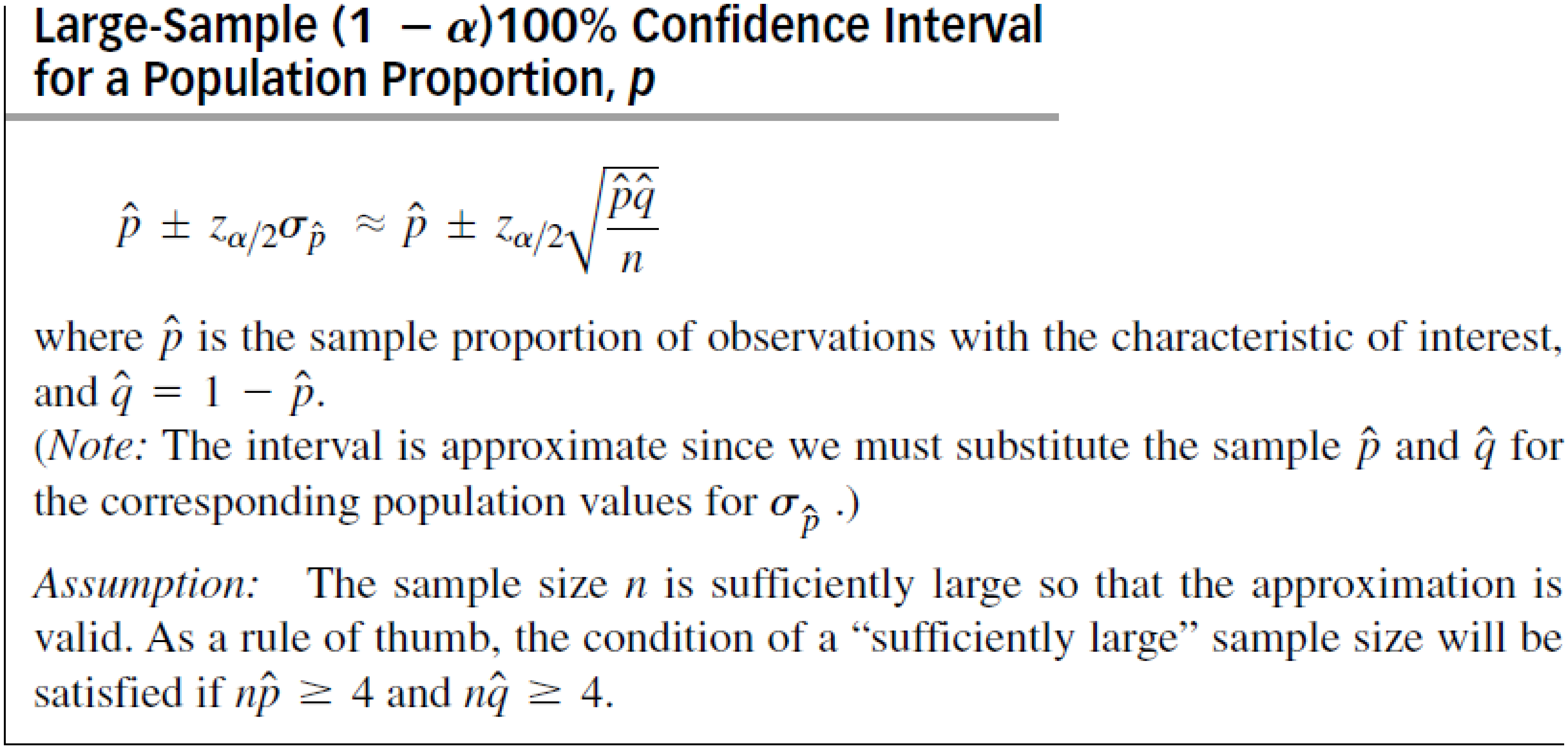
	Condition		
Driver	Restrained	Unrestrained	Difference
1	36.7	36.1	0.6
2	37.5	35.8	1.7
3	39.3	38.4	0.9
4	44.0	41.7	2.3
5	38.4	38.3	0.1
6	43.1	42.6	0.5



```
> with(safety, t.test(REAC.R, REAC.U, paired = TRUE))

        Paired t-test

data:  REAC.R and REAC.U
t = 3.8386, df = 14, p-value = 0.001808
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5206909 1.8393091
sample estimates:
mean of the differences
                   1.18

> with(safety, t.test(Diff))

        One Sample t-test

data:  Diff
t = 3.8386, df = 14, p-value = 0.001808
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.5206909 1.8393091
sample estimates:
mean of x
     1.18

> |
```

How do you interpret the Confidence Interval?

Large sample ci for a proportion



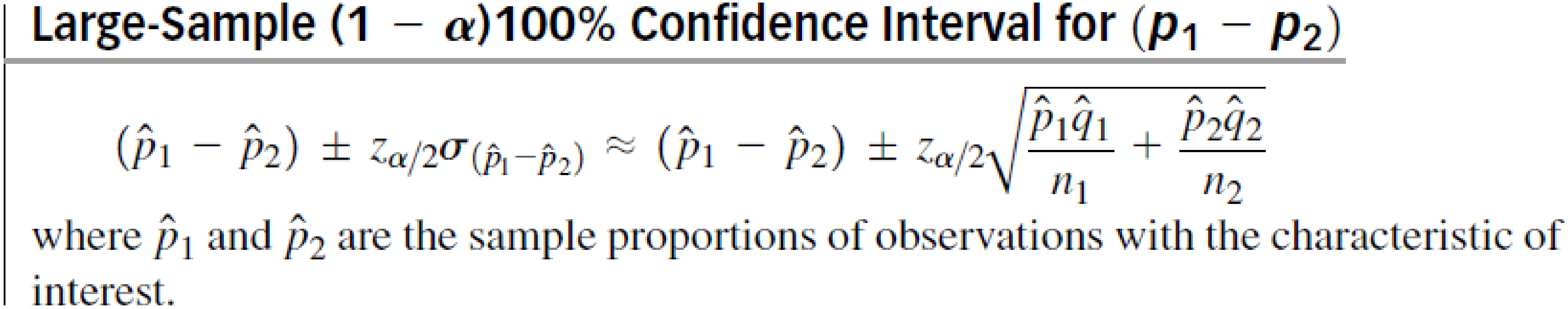
Large-Sample $(1 - \alpha)100\%$ Confidence Interval for a Population Proportion, *p*

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}} \approx \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p}$ is the sample proportion of observations with the characteristic of interest, and $\hat{q} = 1 - \hat{p}$.
(*Note:* The interval is approximate since we must substitute the sample $\hat{p}$ and $\hat{q}$ for the corresponding population values for $\sigma_{\hat{p}}$.)

Assumption: The sample size n is sufficiently large so that the approximation is valid. As a rule of thumb, the condition of a "sufficiently large" sample size will be satisfied if $n\hat{p} \geq 4$ and $n\hat{q} \geq 4$.

ci for difference in proportions

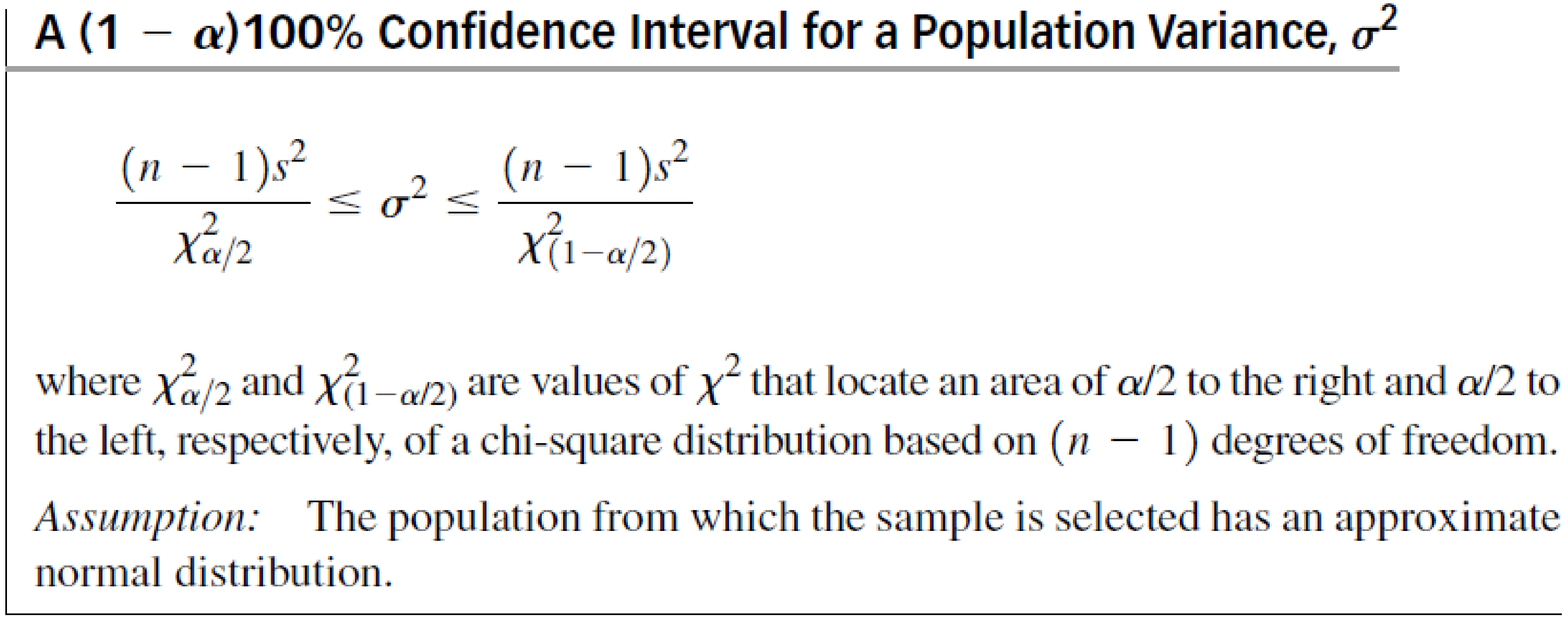


Large-Sample $(1 - \alpha)100\%$ Confidence Interval for $(p_1 - p_2)$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of observations with the characteristic of interest.

ci for a population variance



A (1 − α)100% Confidence Interval for a Population Variance, σ^2

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{(1-\alpha/2)}$ are values of χ^2 that locate an area of $\alpha/2$ to the right and $\alpha/2$ to the left, respectively, of a chi-square distribution based on $(n - 1)$ degrees of freedom.

Assumption: The population from which the sample is selected has an approximate normal distribution.

The F pivotal statistic

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} = \frac{\dfrac{(n_1-1)s_1^2}{\sigma_1^2}\Big/(n_1-1)}{\dfrac{(n_2-1)s_2^2}{\sigma_2^2}\Big/(n_2-1)}$$

$$= \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

$$= \left(\frac{s_1^2}{s_2^2}\right)\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$$

Properties of the f statistic

$$F_{\frac{\alpha}{2}}(\nu_1, \nu_2) = \frac{1}{F_{1-\frac{\alpha}{2}}(\nu_2, \nu_1)}$$

see proof

F quantile equivalents

Dr Wayne Stewart

4/30/2021

Introduction

You may be slightly confused by the varied representations of confidence intervals for sigma because the quantile F multipliers are different to what is expected. This is often because there is an equality that can be expressed:

$$F_\alpha(\nu_1,\nu_2) = \frac{1}{F_{1-\alpha}(\nu_2,\nu_1)}$$

Proof

We will use F_{12} to represent the random variable

$$F_{12} = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

Therefore,

$$F_{21} = \frac{1}{F_{12}}$$

Hence,

$$\begin{aligned} P(F_{12} \leq F_\alpha(\nu_1,\nu_2)) &= 1-\alpha \\ P(1/F_{12} \geq 1/F_\alpha(\nu_1,\nu_2)) &= 1-\alpha \\ P(F_{21} \geq 1/F_\alpha(\nu_1,\nu_2)) &= 1-\alpha \\ P(F_{21} \geq F_{1-\alpha}(\nu_2,\nu_1)) &= 1-\alpha \end{aligned}$$

Therefore the result follows from

$$F_{1-\alpha}(\nu_2,\nu_1) = 1/F_\alpha(\nu_1,\nu_2)$$

after taking reciprocals of both sides.

R confirmation

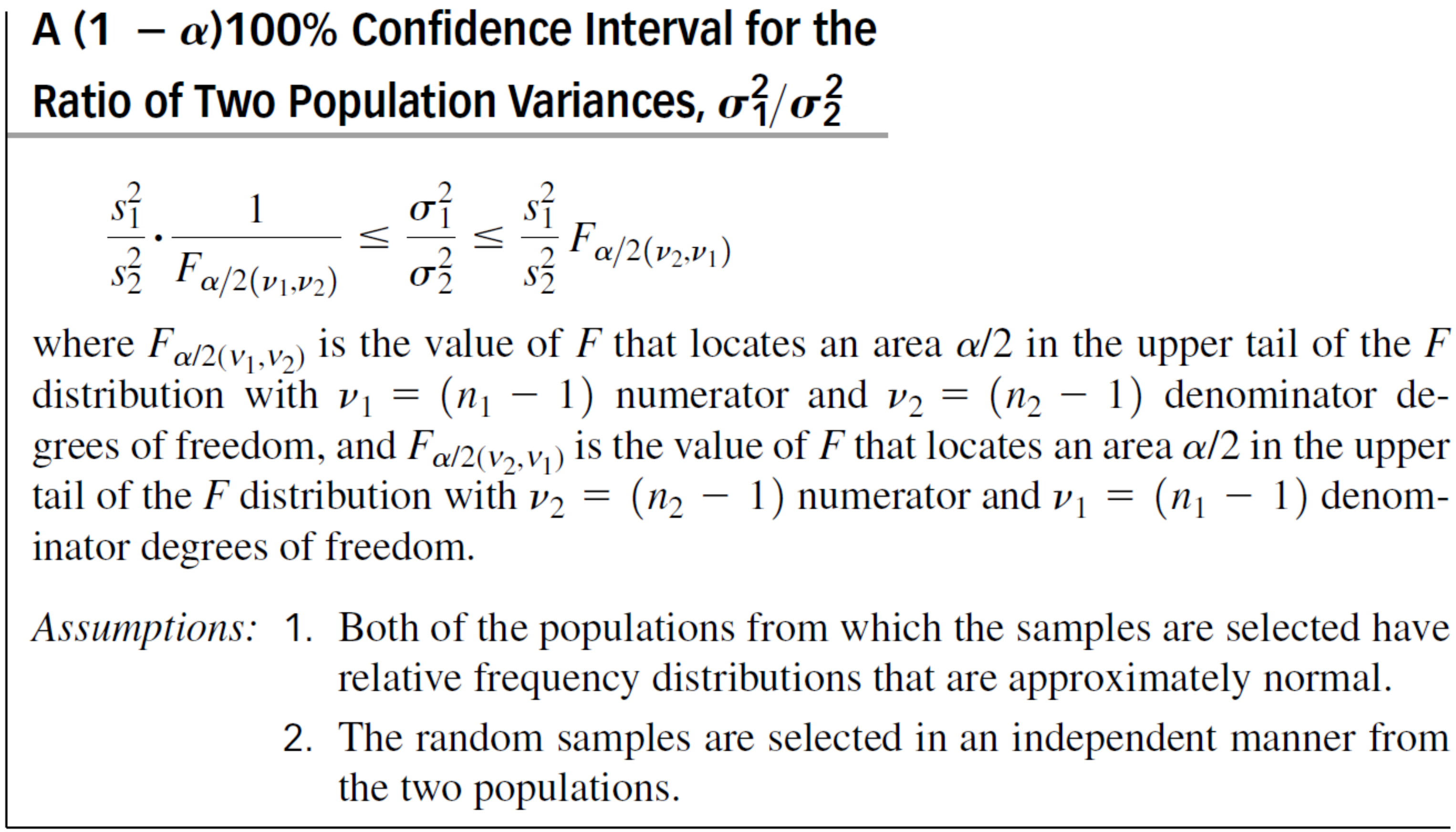
```
alpha = 0.05
nu1 = 30
nu2 = 40
qf(alpha, nu1, nu2)
```

```
## [1] 0.5581011
```

```
1/qf(1-alpha, nu2,nu1)
```

```
## [1] 0.5581011
```

ci for ratio of population variances



A $(1-\alpha)100\%$ Confidence Interval for the Ratio of Two Population Variances, σ_1^2/σ_2^2

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2(\nu_1,\nu_2)}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}F_{\alpha/2(\nu_2,\nu_1)}$$

where $F_{\alpha/2(\nu_1,\nu_2)}$ is the value of F that locates an area $\alpha/2$ in the upper tail of the F distribution with $\nu_1 = (n_1 - 1)$ numerator and $\nu_2 = (n_2 - 1)$ denominator degrees of freedom, and $F_{\alpha/2(\nu_2,\nu_1)}$ is the value of F that locates an area $\alpha/2$ in the upper tail of the F distribution with $\nu_2 = (n_2 - 1)$ numerator and $\nu_1 = (n_1 - 1)$ denominator degrees of freedom.

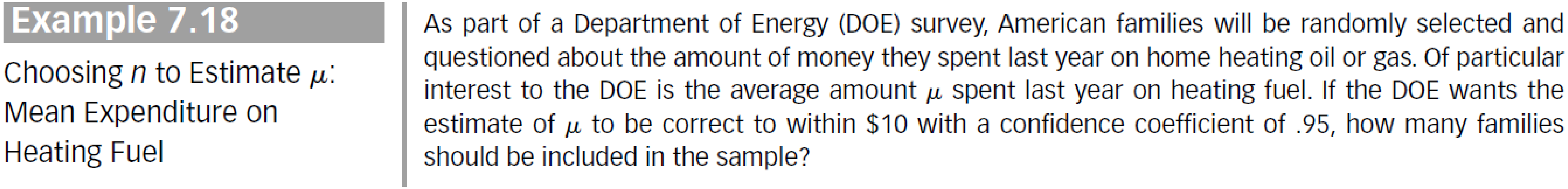
Assumptions:

1. Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.
2. The random samples are selected in an independent manner from the two populations.

Calculate n for a half width

$$H = t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

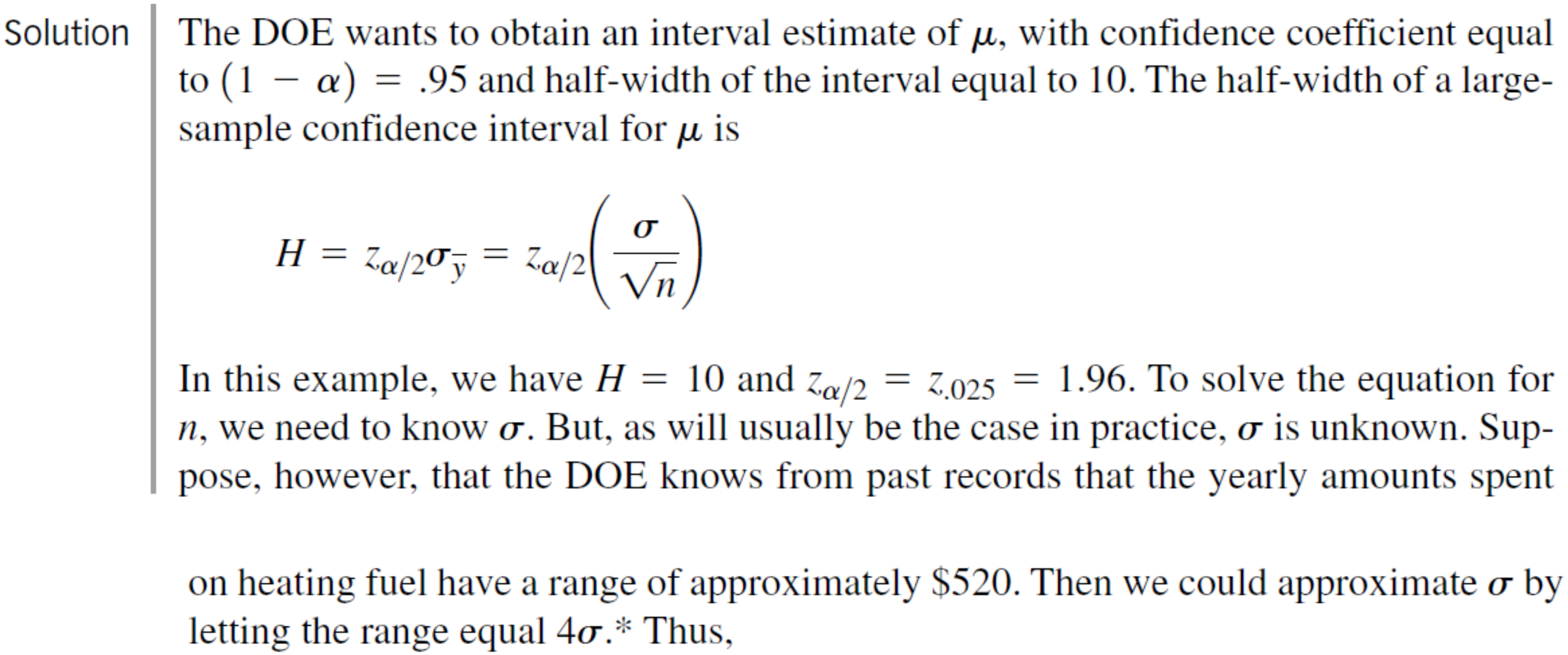
Check out all examples



Example 7.18

Choosing n to Estimate μ: Mean Expenditure on Heating Fuel

As part of a Department of Energy (DOE) survey, American families will be randomly selected and questioned about the amount of money they spent last year on home heating oil or gas. Of particular interest to the DOE is the average amount μ spent last year on heating fuel. If the DOE wants the estimate of μ to be correct to within \$10 with a confidence coefficient of .95, how many families should be included in the sample?



Solution

The DOE wants to obtain an interval estimate of μ, with confidence coefficient equal to $(1 - \alpha) = .95$ and half-width of the interval equal to 10. The half-width of a large-sample confidence interval for μ is

$$H = z_{\alpha/2}\sigma_{\bar{y}} = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

In this example, we have $H = 10$ and $z_{\alpha/2} = z_{.025} = 1.96$. To solve the equation for n, we need to know σ. But, as will usually be the case in practice, σ is unknown. Suppose, however, that the DOE knows from past records that the yearly amounts spent on heating fuel have a range of approximately \$520. Then we could approximate σ by letting the range equal 4σ.* Thus,

Solution

$$4\sigma \approx 520 \quad \text{or} \quad \sigma \approx 130$$

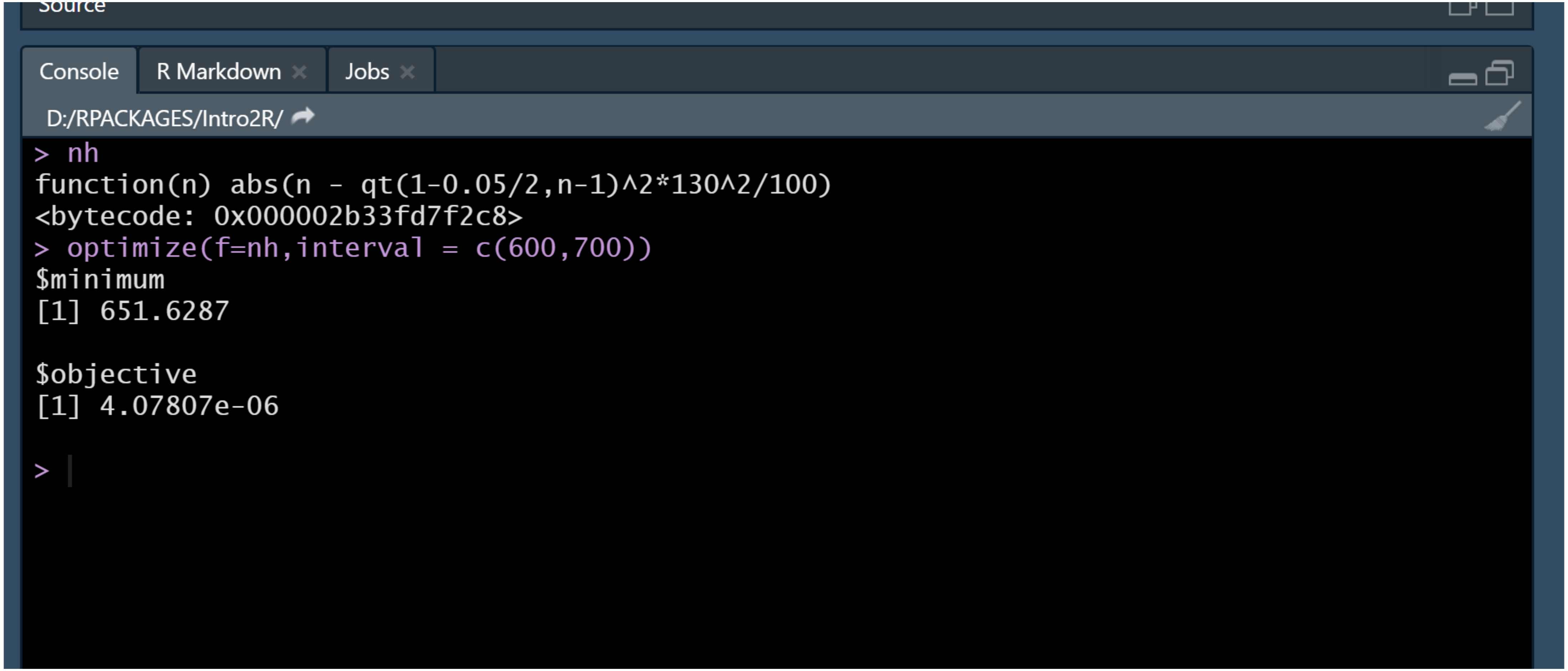
Solving for n, we have

$$H = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{or} \quad 10 = 1.96\left(\frac{130}{\sqrt{n}}\right)$$

or

$$n = \frac{(1.96)^2(130)^2}{(10)^2} \approx 650$$

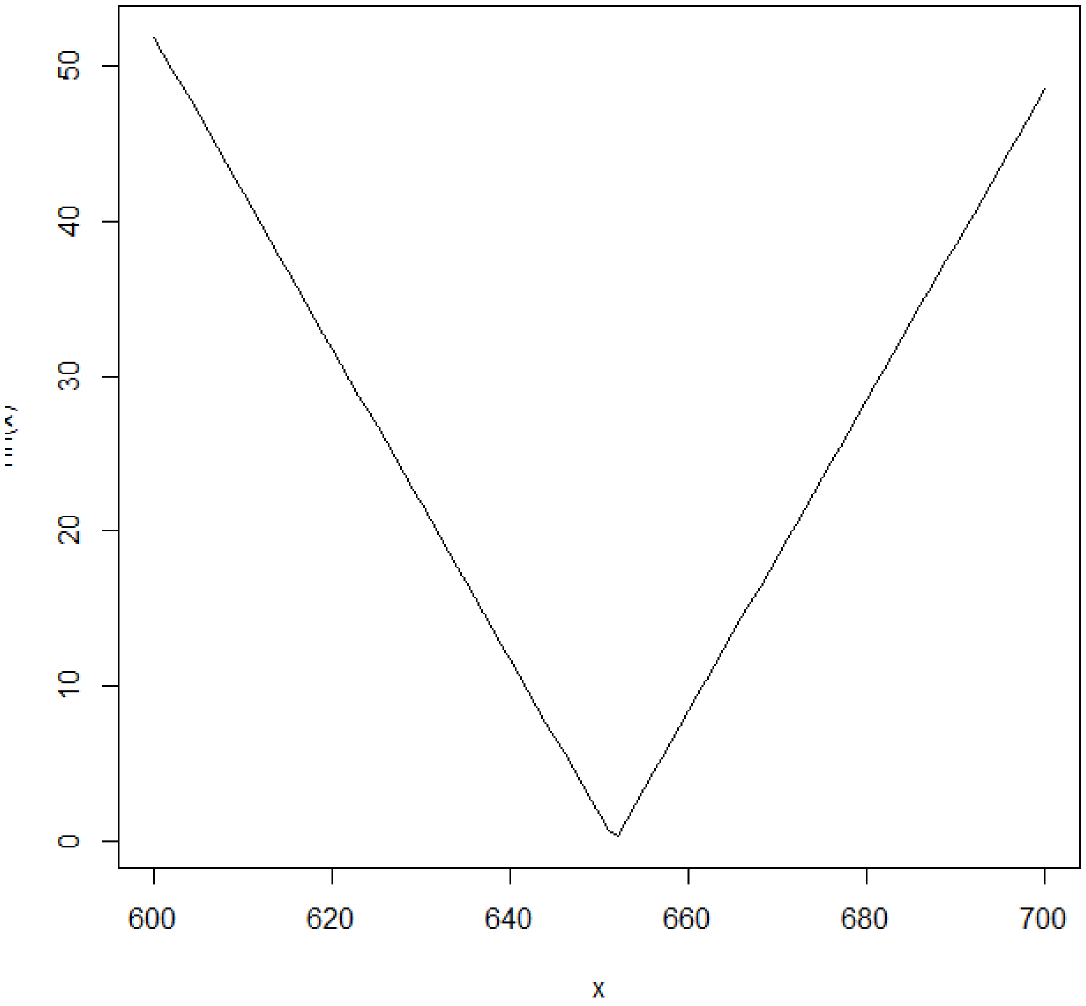
We could use $t_{\frac{\alpha}{2}}$



```
> nh
function(n) abs(n - qt(1-0.05/2,n-1)^2*130^2/100)
<bytecode: 0x000002b33fd7f2c8>
> optimize(f=nh,interval = c(600,700))
$minimum
[1] 651.6287

$objective
[1] 4.07807e-06

> 
```

The use of $Z_{\frac{\alpha}{2}}$ instead of $t_{\frac{\alpha}{2}}$ is to simplify the computation

Not much difference



Choosing the Sample Size for Estimating a Population Mean μ to Within H Units with Probability $(1-\alpha)$

$$n = \left(\frac{z_{\alpha/2}\sigma}{H}\right)^2$$

(*Note:* The population standard deviation σ will usually have to be approximated.)

Choosing the Sample Sizes for Estimating the Difference $(\mu_1 - \mu_2)$ Between a Pair of Population Means Correct to Within H Units with Probability $(1-\alpha)$

$$n_1 = n_2 = \left(\frac{z_{\alpha/2}}{H}\right)^2(\sigma_1^2 + \sigma_2^2)$$

where n_1 and n_2 are the numbers of observations sampled from each of the two populations, and σ_1^2 and σ_2^2 are the variances of the two populations.

Choosing the Sample Size for Estimating a Population Proportion p to Within H Units with Probability $(1-\alpha)$

$$n = \left(\frac{z_{\alpha/2}}{H}\right)^2 pq$$

where p is the value of the population proportion that you are attempting to estimate, and $q = 1 - p$.
(*Note:* This technique requires previous estimates of p and q. If none are available, use $p = q = .5$ for a conservative choice of n.)

Choosing the Sample Sizes for Estimating the Difference $(p_1 - p_2)$ Between Two Population Proportions to Within H Units with Probability $(1-\alpha)$

$$n_1 = n_2 = \left(\frac{z_{\alpha/2}}{H}\right)^2(p_1q_1 + p_2q_2)$$

where p_1 and p_2 are the proportions for populations 1 and 2, respectively, and n_1 and n_2 are the numbers of observations to be sampled from each population.

STATISTICS IN ACTION

Bursting Strength of PET Beverage Bottles

Polyethylene terephthalate (PET) bottles are used for carbonated beverages. At a certain facility, PET bottles are produced by inserting injection molded pre-forms into a stretch blow machine with 24 cavities. Each machine can produce 440 bottles per minute. A critical property of PET bottles is their bursting strength — the pressure at which bottles filled with water burst when pressurized.

In the *Journal of Data Science* (May 2003), researchers measured and analyzed the bursting strength of PET bottles made from two different mold designs – an old design and a new design. The new mold design reduces the time to change molds in the blow machine, thus reducing the downtime of the machine. This advantage, however, will be negated if the new design has problems with low bursting strength. Consequently, an analysis was performed to compare bursting strengths of PET bottles produced from the two mold designs.

The data for the analysis was obtained by testing one bottle per cavity per day produced over a period of 32 days for each design. Since the machine has 24 cavities, there were a total of 32 × 24 = 768 PET bottles tested for each design. Each bottle was filled with water and pressurized until it burst and the resulting pressure (in pounds per square inch) was recorded. These bursting strengths are saved in the **PETBOTTLE** file, described in Table SIA7.1. The researchers showed that there were no significant trends in bursting strength over time (days) and that there were no "cavity effects" (i.e., no significant bursting strength differences among the 24 cavities within each design). Thus, the data for all cavities and all days were pooled in order to compare the two mold designs.

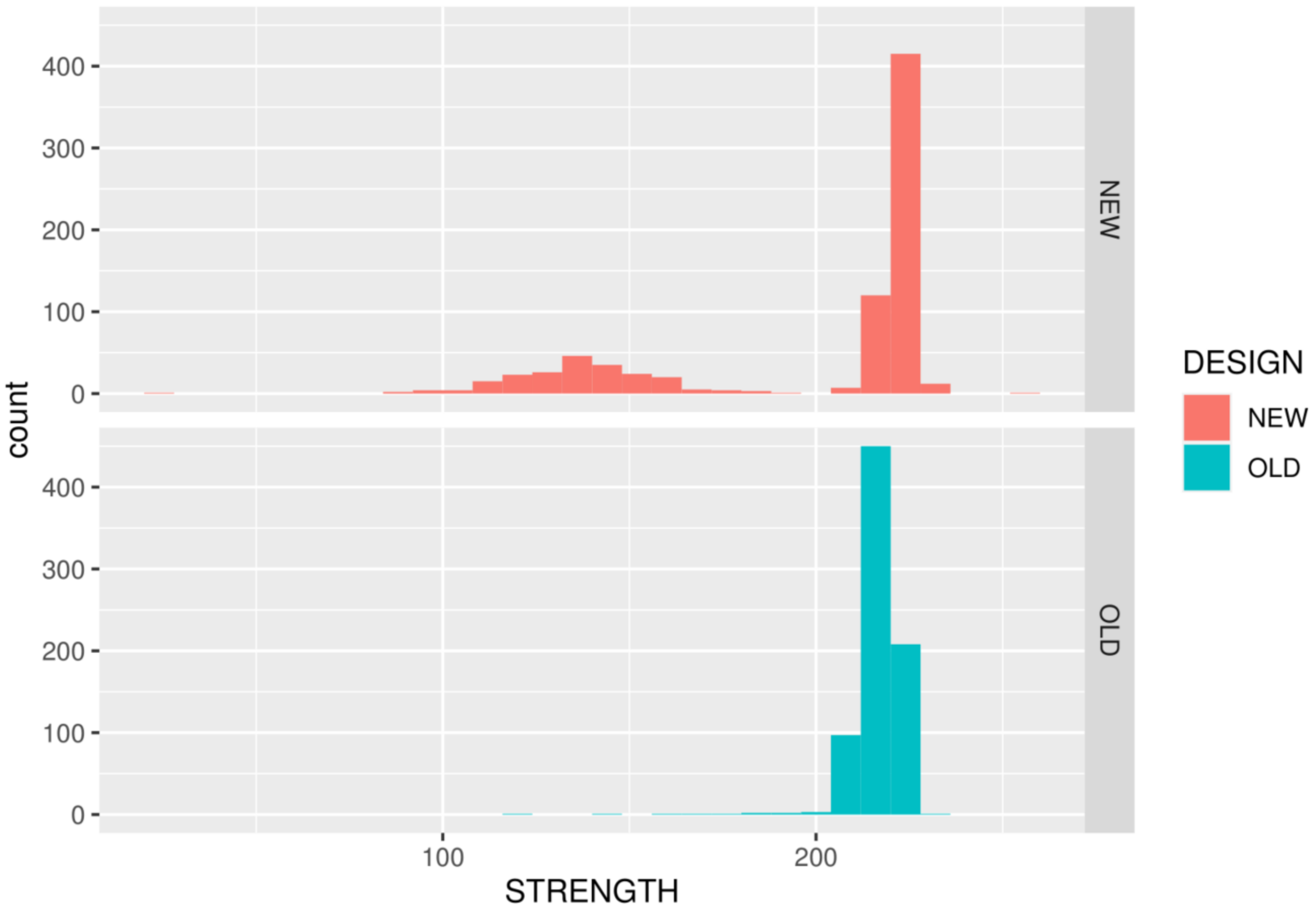
PETBOTTLE

TABLE SIA7.1:

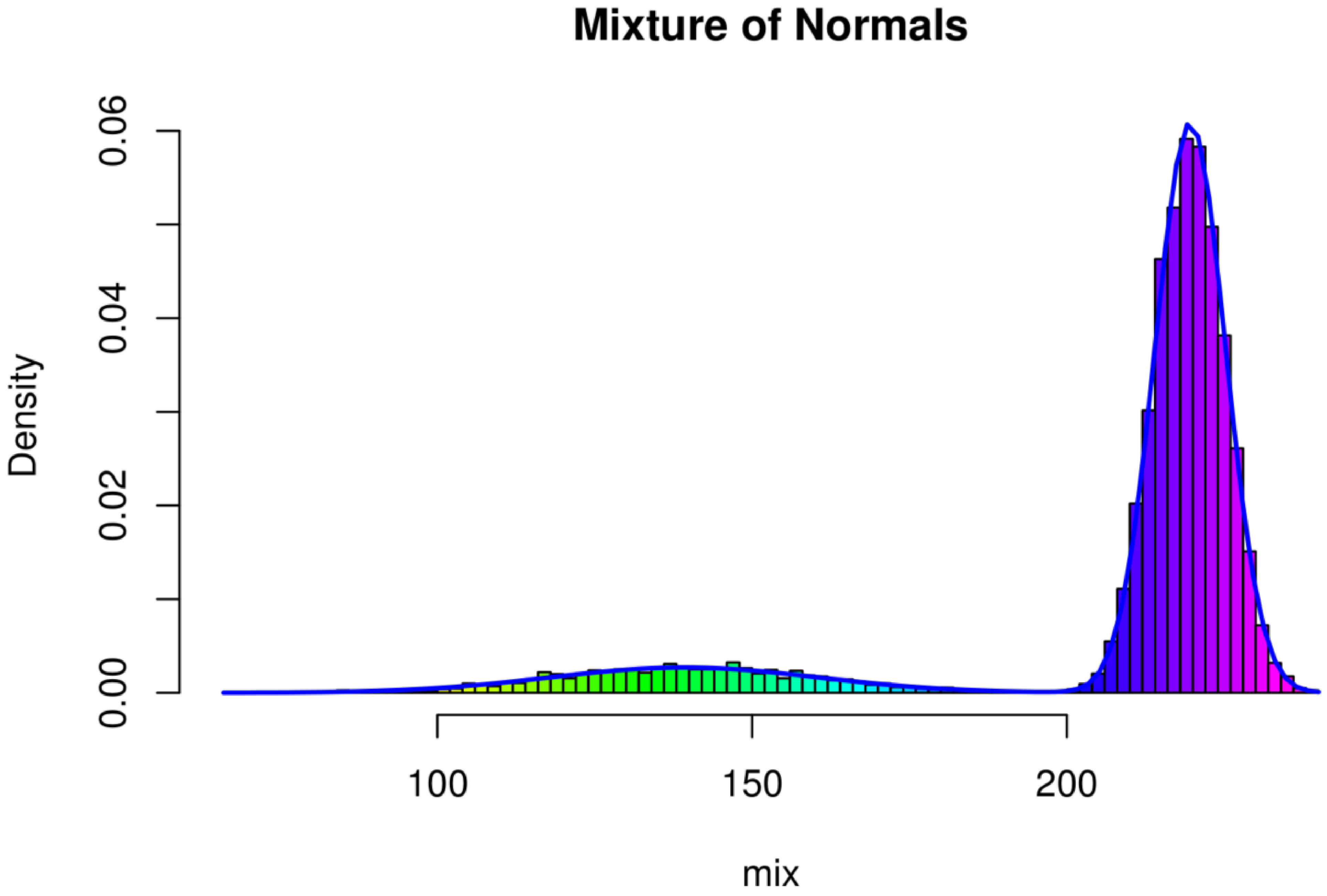
Variable Name	Description	Data Type
DESIGN	Mold design (OLD or NEW)	Qualitative
DAY	Day number	Quantitative
CAVITY	Cavity number	Quantitative
STRENGTH	Bursting strength (psi)	Quantitative

In the Statistics in *Action Revisited* at the end of this chapter, we demonstrate how the methods outlined in this chapter can be used to compare bursting strengths of PET bottles produced from the two mold designs.

PET EXAMPLE



count
400
300
200
100
0
NEW
OLD
100
200
STRENGTH
DESIGN
NEW
OLD


Mixture of Normals
Density
0.00
0.02
0.04
0.06
100
150
200
mix

Problems

CLT

We can see that there is some bimodality for the New design and non Normality also apparent in the Old design. Does this fact in itself make the assumptions of Normality for $\bar{Y}_{NEW} - \bar{Y}_{OLD}$ invalid? The answer is NO!! Why? Because of the CLT, which says that regardles of the distribution of Y the mean $\bar{Y}$ will be approx. Normal with n sufficiently big.

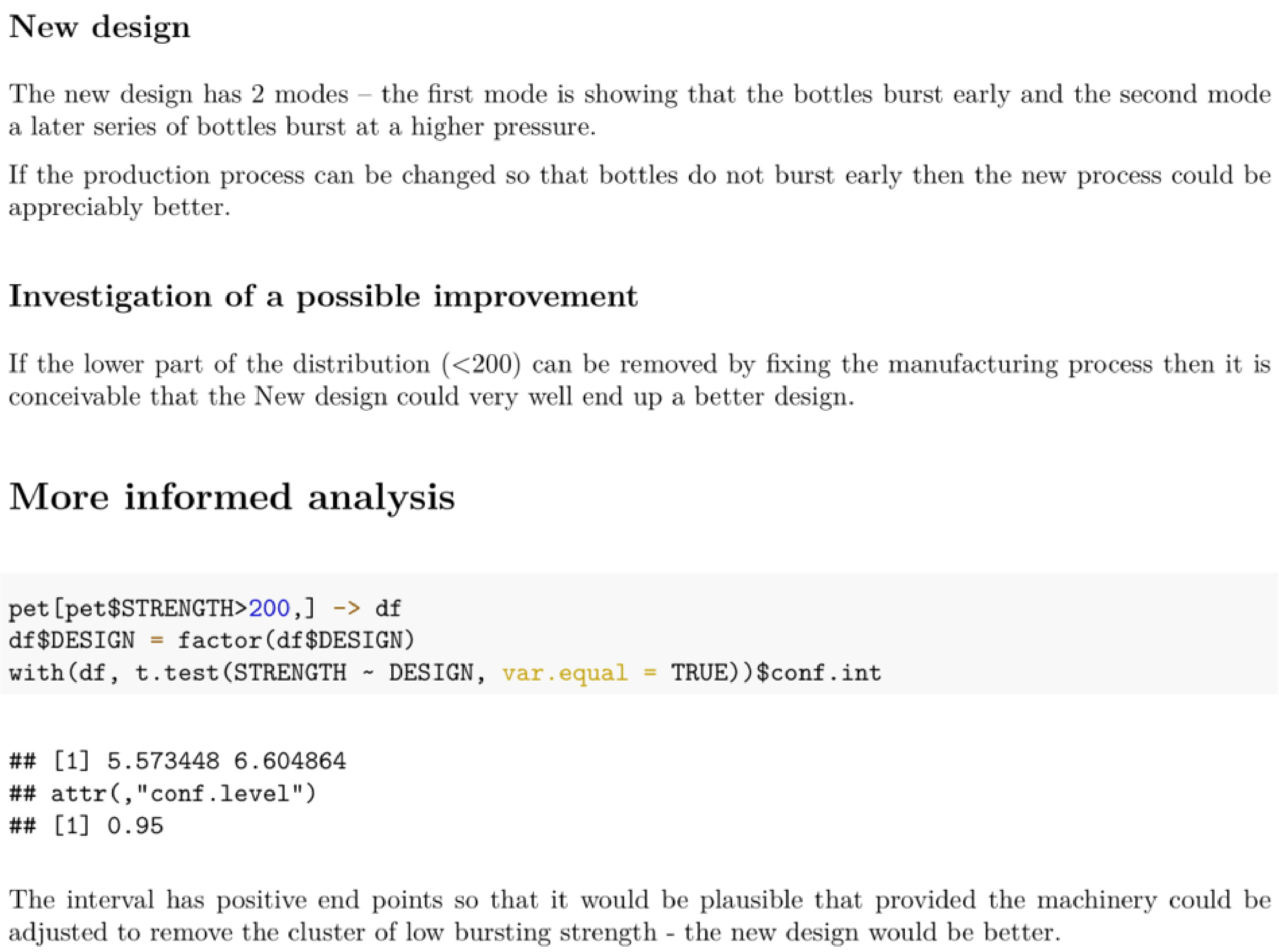
Here is the ci for $\mu_{NEW} - \mu_{OLD}$ – how do we know it is in this order? R will take the alphabetial order by default.

```
tt<-with(pet, t.test(STRENGTH ~ DESIGN, var.equal=TRUE))
unique(pet$DESIGN)
```

```
## [1] "OLD" "NEW"
```

Assumming equality of population variances, appealing to the CLT and creating a ci using the side effect of the t.test yields -19.9100336, -14.1628831.

This means that it is plausible that the new pop mean is smaller than the old pop mean. (Both sides of the ci are negative)



New design

The new design has 2 modes – the first mode is showing that the bottles burst early and the second mode a later series of bottles burst at a higher pressure.

If the production process can be changed so that bottles do not burst early then the new process could be appreciably better.

Investigation of a possible improvement

If the lower part of the distribution (<200) can be removed by fixing the manufacturing process then it is conceivable that the New design could very well end up a better design.

More informed analysis

R!!

```
pet[pet$STRENGTH>200,] -> df
df$DESIGN = factor(df$DESIGN)
with(df, t.test(STRENGTH ~ DESIGN, var.equal = TRUE))$conf.int
```

```
## [1] 5.573448 6.604864
## attr(,"conf.level")
## [1] 0.95
```

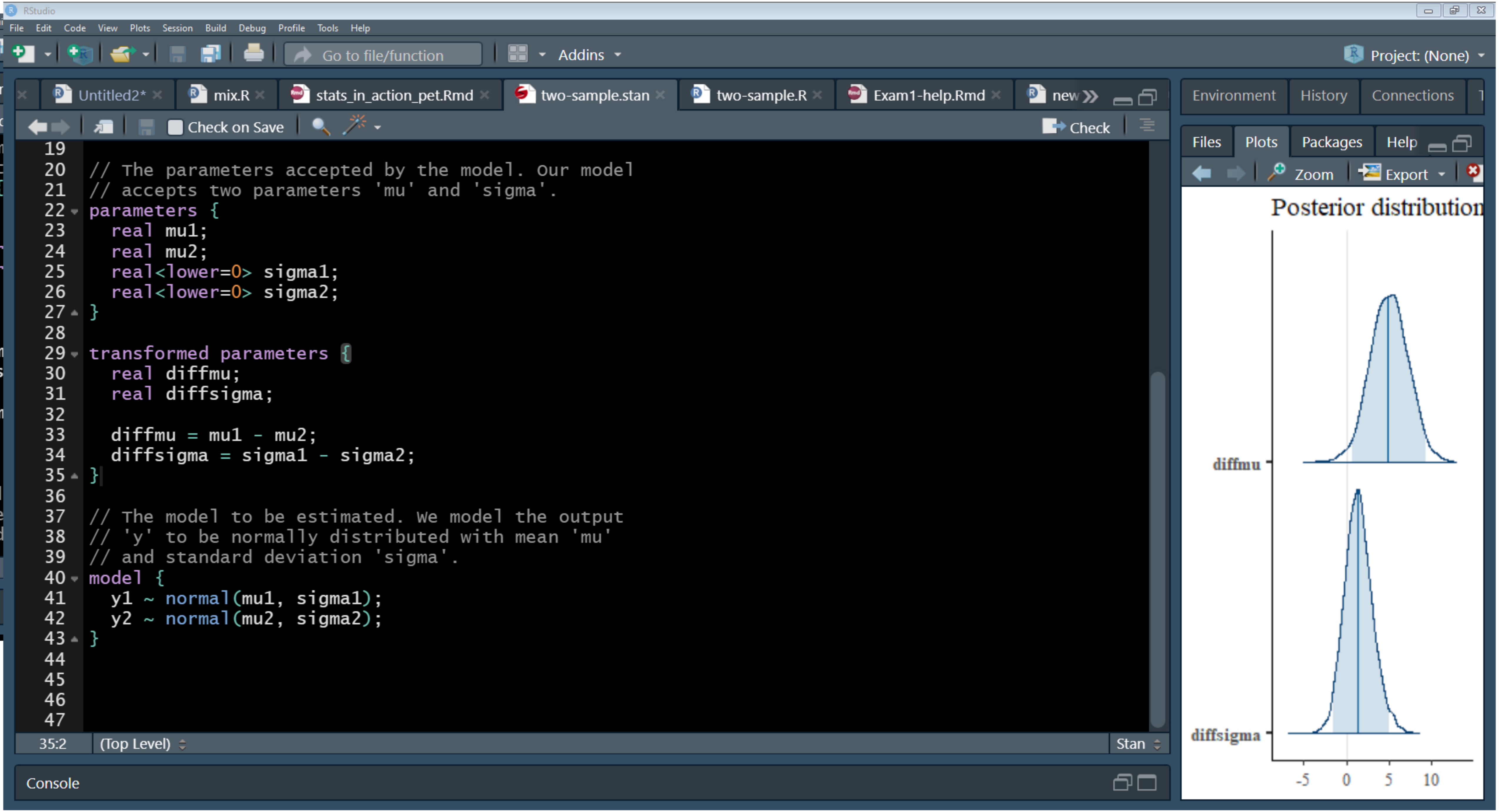
The interval has positive end points so that it would be plausible that provided the machinery could be adjusted to remove the cluster of low bursting strength - the new design would be better.

Using the Bayesian paradigm

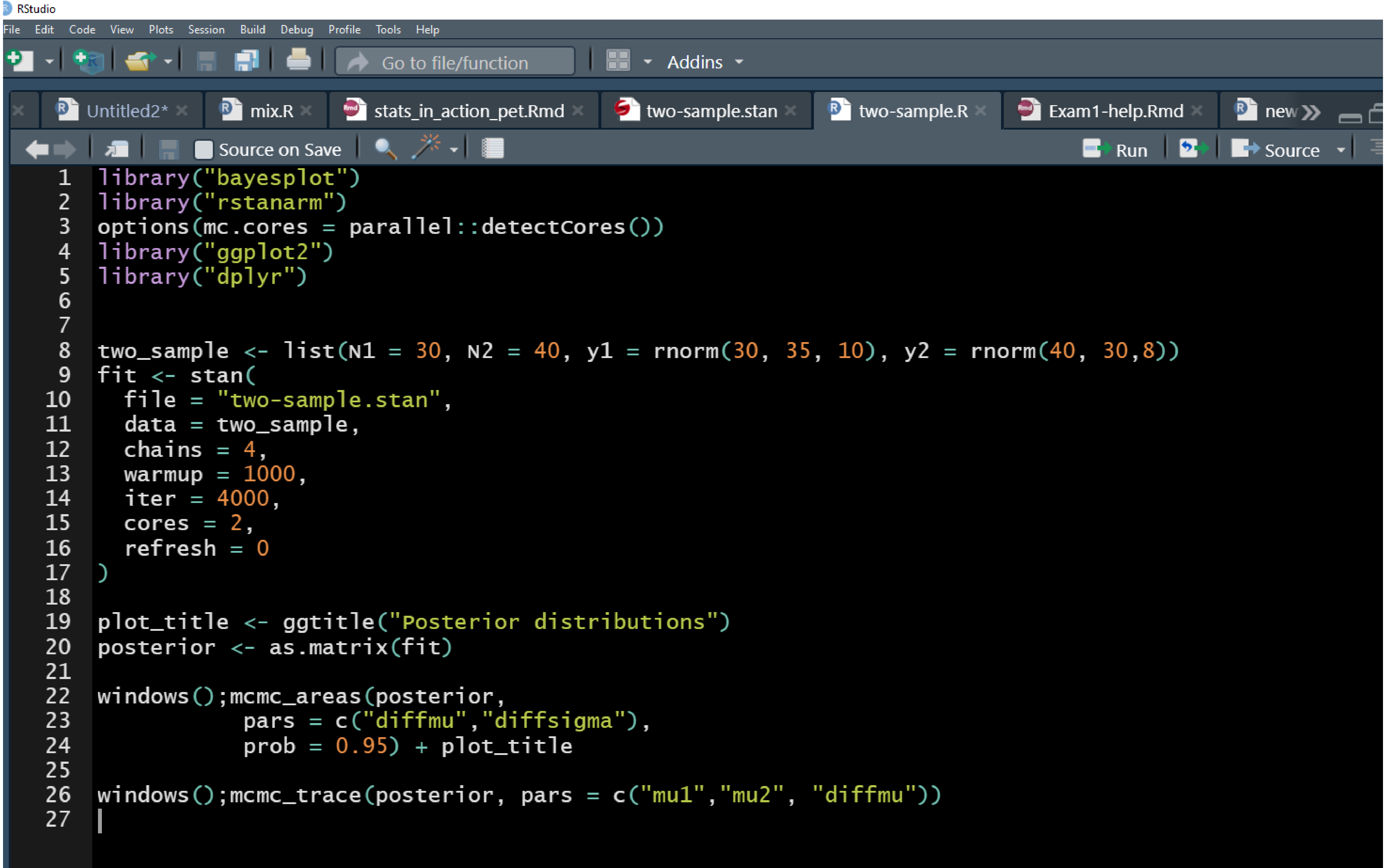
Bayes' Rule

$$\boldsymbol{p(\theta|x) = \frac{p(\theta)f(x|\theta)}{p(x)}}$$

Use

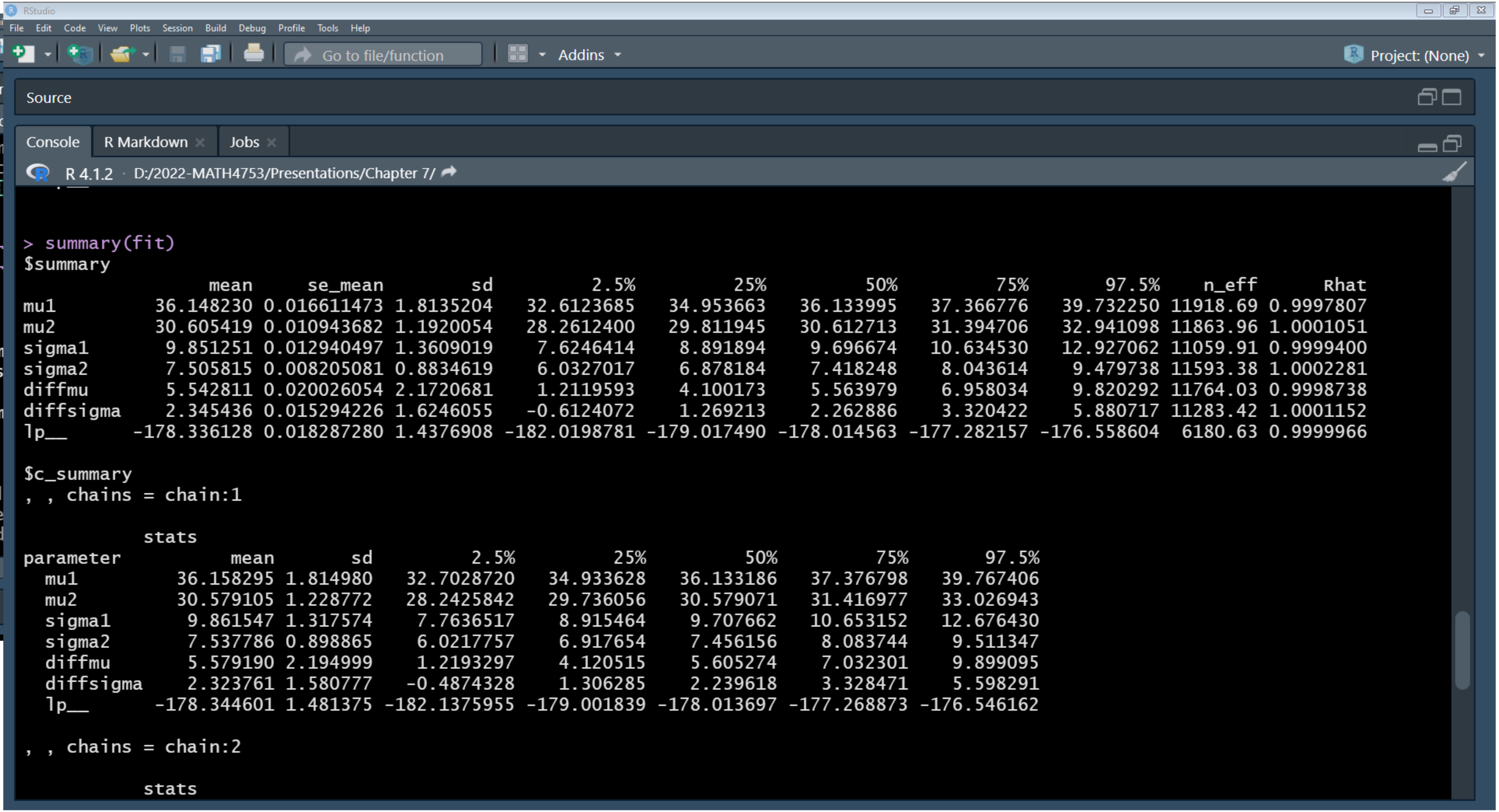


```

// The parameters accepted by the model. Our model
// accepts two parameters 'mu' and 'sigma'.
parameters {
  real mu1;
  real mu2;
  real<lower=0> sigma1;
  real<lower=0> sigma2;
}

transformed parameters {
  real diffmu;
  real diffsigma;

  diffmu = mu1 - mu2;
  diffsigma = sigma1 - sigma2;
}

// The model to be estimated. We model the output
// 'y' to be normally distributed with mean 'mu'
// and standard deviation 'sigma'.
model {
  y1 ~ normal(mu1, sigma1);
  y2 ~ normal(mu2, sigma2);
}
```



```r
library("bayesplot")
library("rstanarm")
options(mc.cores = parallel::detectCores())
library("ggplot2")
library("dplyr")


two_sample <- list(N1 = 30, N2 = 40, y1 = rnorm(30, 35, 10), y2 = rnorm(40, 30,8))
fit <- stan(
  file = "two-sample.stan",
  data = two_sample,
  chains = 4,
  warmup = 1000,
  iter = 4000,
  cores = 2,
  refresh = 0
)

plot_title <- ggtitle("Posterior distributions")
posterior <- as.matrix(fit)

windows();mcmc_areas(posterior,
           pars = c("diffmu","diffsigma"),
           prob = 0.95) + plot_title

windows();mcmc_trace(posterior, pars = c("mu1","mu2", "diffmu"))

```


Plot Zoom
Posterior distributions
diffmu
diffsigma
-5
0
5
10



```
> summary(fit)
$summary
                mean     se_mean        sd        2.5%         25%         50%         75%       97.5%    n_eff      Rhat
mu1        36.148230 0.016611473 1.8135204   32.6123685   34.953663   36.133995   37.366776   39.732250 11918.69 0.9997807
mu2        30.605419 0.010943682 1.1920054   28.2612400   29.811945   30.612713   31.394706   32.941098 11863.96 1.0001051
sigma1      9.851251 0.012940497 1.3609019    7.6246414    8.891894    9.696674   10.634530   12.927062 11059.91 0.9999400
sigma2      7.505815 0.008205081 0.8834619    6.0327017    6.878184    7.418248    8.043614    9.479738 11593.38 1.0002281
diffmu      5.542811 0.020026054 2.1720681    1.2119593    4.100173    5.563979    6.958034    9.820292 11764.03 0.9998738
diffsigma   2.345436 0.015294226 1.6246055   -0.6124072    1.269213    2.262886    3.320422    5.880717 11283.42 1.0001152
lp__     -178.336128 0.018287280 1.4376908 -182.0198781 -179.017490 -178.014563 -177.282157 -176.558604  6180.63 0.9999966

$c_summary
, , chains = chain:1

           stats
parameter          mean       sd        2.5%         25%         50%         75%       97.5%
  mu1         36.158295 1.814980   32.7028720   34.933628   36.133186   37.376798   39.767406
  mu2         30.579105 1.228772   28.2425842   29.736056   30.579071   31.416977   33.026943
  sigma1       9.861547 1.317574    7.7636517    8.915464    9.707662   10.653152   12.676430
  sigma2       7.537786 0.898865    6.0217757    6.917654    7.456156    8.083744    9.511347
  diffmu       5.579190 2.194999    1.2193297    4.120515    5.605274    7.032301    9.899095
  diffsigma    2.323761 1.580777   -0.4874328    1.306285    2.239618    3.328471    5.598291
  lp__      -178.344601 1.481375 -182.1375955 -179.001839 -178.013697 -177.268873 -176.546162

, , chains = chain:2

           stats
```

Quick Review

Key Terms

(*Note: Items marked with an asterisk (*) are from the optional section in this chapter.*)

*Bayes credible (probability) intervals 354
*Bayesian estimation 352
Biased estimator 291
*Bootstrap estimation 350
Confidence coefficient 301
Confidence interval 301
Interval estimator 290
Jackknife estimators 299
Least-squares method 488
Likelihood function 296
Lower confidence limit 302
Matched pairs 322
Maximum likelihood method 297
Mean squared error 291
Method of moments 294
Minimum variance unbiased estimator 291
Paired observations 322
*Percentile bootstrap confidence intervals 352
Pivotal method 301
Pivotal statistic 301
Point estimator 289
Pooled estimate of variance 316
Population moment 294
*Posterior probability distribution 354
*Prior probability distribution 354
*Resampling 350
Robust estimators 300
Sample moment 294
*Squared error loss 353
Unbiased estimator 291
Upper confidence limit 302

Chapter Summary Notes

- A point estimator $\hat{\theta}$ of a population parameter θ is **unbiased** if $E(\hat{\theta}) = \theta$; otherwise, the estimator is **biased**.
- The **minimum variance unbiased estimator (MVUE)** of a population parameter θ has the smallest variance among all unbiased estimators.
- Methods of estimation: **pivotal method** (either the **method of moments** or the **maximum likelihood method**), **jackknife method, robust estimation methods, bootstrapping**, and **Bayes' estimation**.
- **Confidence interval**—an interval that encloses an unknown population parameter with a certain level of confidence
- **Confidence coefficient**—the probability that a randomly selected confidence interval encloses the value of the population parameter
- Interpretation of the phrase "*$(1 - \alpha)100\%$ confident*": In repeated sampling, $(1 - \alpha)100\%$ of all similarly constructed intervals will enclose the true parameter value.
- Key words for identifying $\boldsymbol{\mu}$ as the parameter of interest: *mean, average*.
- Key words/phrases for identifying $\boldsymbol{\mu_1 - \mu_2}$ as the parameter of interest: *difference between means or averages, compare two means using independent samples*.
- Key words/phrases for identifying $\boldsymbol{\mu_d}$ as the parameter of interest: *mean or average of paired differences, compare two means using matched pairs*.
- Key words for identifying $\boldsymbol{p}$ as the parameter of interest: *proportion, percentage, rate*.
- Key words/phrases for identifying $\boldsymbol{p_1 - p_2}$ as the parameter of interest: *difference between proportions or percentages, compare two proportions using independent samples*.
- Key words for identifying $\boldsymbol{\sigma^2}$ as the parameter of interest: *variance, spread, variation*.
- Key words/phrases for identifying $\boldsymbol{\sigma_1^2/\sigma_2^2}$ as the parameter of interest: *difference between variances, compare variation in two populations using independent samples*.