1 Introduction

Link to textbook: Richard L. Scheaffer, William Mendenhall III, R. Lyman Ott, and Kenneth G. Gerow (2012), *Elementary Survey Sampling*, 7th edition

Example: Washington Post - ABC Poll: In a poll of 907 adults, June 27-30, 2021, 86% of Democrats have received at least one COVID-19 vaccine shot, and 45% of Republicans have received at least one COVID-19 vaccine shot.

What proportion of Americans were included in that poll?

Is it reasonable to conclude that the difference in vaccination rates among **ALL** Democrats and Republicans in the United States is large, say, at least 20%, during that time period?

What is a purpose of sample surveys in this example?

Does **loyalty** to a political party **cause** people to choose or decline the vaccine?

Do certain **values** which influence political preferences also influence vaccination status?

Is the large discrepancy in vaccination rates between Democrats and Republicans due to ONLY values and loyalty of political affiliation?

Aside, but simple example: Does consumption of ice cream **cause** people to drown?

Do most Americans truly believe that *association does not imply causation?*

□

Remark: With *sample surveys,* sometimes we *like* the results of the survey, in the sense that the survey might confirm our beliefs. In other cases, we might feel *disturbed* by the results of the survey.

Why do we conduct surveys?

What is a **population**?

What is a **sample**?

Quote of the semester: "Sample surveys, when properly performed, often serve as a shield or deterrent against fake political propaganda."

Sample surveys (Math 325) differs from **experimental design** (Math 321).

What are some useful examples of surveys?

Examples of surveys come from social sciences (opinion polls), engineering, agriculture, ecology (study of organisms and their environment), medicine, test marketing/business, natural resource management (timber, wildlife, and recreation).

- What information do the United States Census and American Community Survey (every year for a small percentage of households, www.census.gov) collect?

The *short*-form census counts all residents in the U.S., as well as name, gender, age, date of birth, race, ethnicity, relationships within the household, and housing.

The *long*-form census, previously given to 1/6 households, requests detailed information, including citizenship, education, occupation, income, and housing conditions. This *long*-form census was replaced by the American Community Survey (ACS) in 2010.

ACS determine federal government's allocations of funds to states and cities.

Businesses forecast sales, manage personnel, and establish future site locations.

Urban and regional planners plan land use, transportation networks, and energy consumption.

Social scientists study economic conditions and racial balance.

- Why does the U.S. Bureau of Labor Statistics conduct surveys?

Determine the consumer price index (CPI), a measure of price change for goods and services.

The CPI measures inflation and helps determine wages and pensions for businesses, and affects rents and mortgages.

The Bureau's surveys estimate the labor force, earnings, and future opportunities (10 years in advance) for various occupations.

Example: Quotes from Shere Hite's book *Women and Love: A Cultural Revolution in Progress* (1987):

(a) 84% of women are "not satisfied emotionally with their relationships" (p. 804).

(b) 70% of all women "married five or more years are having sex outside of their marriages" (p. 856).

(c) 95% of women "report forms of emotional and psychological harassment from men with whom they are in love relationships" (p. 810).

(d) 84% of women "report forms of condescension from the men in their love relationships" (p. 809).

Example: News statement in 1993: "Twenty-two percent of Americans doubt that the Holocaust ever occurred."

Question from the Roper organization: "Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?"

Responses were: 22% said "it seemed possible," 12% said they did not know, 65% said it was "impossible it never happened."

Later, question from the Gallup organization: "The term *Holocaust* usually refers to the killing of millions of Jews in Nazi death camps during World War II. In your opinion, did the Holocaust: definitely happen, probably happen, probably did not happen, or definitely not happen?"

Responses were: 83% said Holocaust definitely happened, 13% said Holocaust probably happened, 1% said Holocaust definitely did not happen.

"When statistics are not based on strictly accurate calculations, they mislead instead of guide. The mind easily lets itself be taken in by the false appearance of exactitude which statistics retain in their mistakes, and confidently adopts errors clothed in the form of mathematical truth." – Alexis de Tocqueville, *Democracy in America*

Using *R* or *RStudio* {see https://www.r-project.org or https://rstudio.com}

To download *R*: Click "CRAN", "UCLA" (or some other mirror site), "Download R for Windows", "base", "Download R 4.3.1 for Windows" (or the latest version). Click "Save" and then "Save" again. The saving should take less than five minutes using high-speed internet. Click "Run", "Run", and "OK", and continue to click "Next" until the process is complete.

Alternatively, use *Rweb* by going to https://rweb.webapps.cla.umn.edu/Rweb/

```
> 11:20 # Generate numbers from 11 to 20
> c(11:20) # Also, generate numbers from 11 to 20.  "c" means combine.

> x = c(11:20) # Save numbers from 11 to 20 as the variable x.


> x # List contents of x.

> x[4]

> x[4:7]

> y = 3 * x ^ 2 + 5

> y

> plot( x, y )

> c(5:-3, 10:15, 20, 6)

> ls( ) # Lists all variables.
```

Example: Consider data in exercise 4.40 on p. 109, EXER4_40.txt. The columns are “item number”, “audited amount (dollars)”, “recorded amount (dollars)” and “overstatement (difference in dollars)”.

```
> y = read.table(
   "http://educ.jmu.edu/~garrenst/math325.dir/datasets/EXER4_40.txt") #
   Read the data from the website.

> help( read.table )

> ?read.table
```

Tables may be saved using `write.table`.

```
> source( "http://educ.jmu.edu/~garrenst/math325.dir/Rfunctions" ) #
   Read in the functions.
```

Look at `ls` and `read.table2`.

```
> y[5:10, 2] # To view the audited amount for items 5 to 10.
> # Now, view only the odd numbered items.
> odd = (0:7) * 2 + 1



> odd
> y[odd, ]
> # Functions use ( ), and variables use [ ].
> colnames(y)
> colnames(y) = c( "item number", "audited amount", "recorded amount",
 "overstatement" )
> y

> # Compute the mean recorded amount.

> x2 = y[ , 3]; mean(x2) # Computes the mean recorded amount.


> # Suppose a penalty equal to double the overstatement is applied.
> # Append this new variable.
> penalty = 2 * y[ , "overstatement" ]
> penalty
> x2 = cbind(y, penalty)
> x2
> apply( x2, 2, mean ) # Computes the mean for all 5 categories.
> # Using only one or two commands in R, in the ‘y’ table, change the
   9th recorded amount from 82 to 62, and then fix the overstatement as
   well.
```

Example: Read in the following data set regarding income (in thousands of dollars): {48, 75, 93, NA, 81, 53}. Compute the mean and standard deviation of this data set, while discarding all values of NA.

Example: Read in the data from exercise 6.21 on p. 209, EXER6_21.txt.

```
> # As a table.

> x = read.table(
   "http://educ.jmu.edu/~garrenst/math325.dir/datasets/EXER6_21.txt" )

> # As a single vector.
```

Example: Consider data from CARS93.txt in Appendix C, pp. 416-419. Compute the average city mpg and the mean number of cylinders among all 57 cars.

```
> z = read.table(
   "http://educ.jmu.edu/~garrenst/math325.dir/datasets/CARS93.txt" )
```

```
> # Take a simple random sample of five values of city mpg withOUT
  replacement.
```

Definition: A **data frame** is an easy-to-access structured collection of variable(s) of the same length, and may contain different types of variables (e.g., numeric, string, logical - TRUE/FALSE).

Example: Enter the data from Table 2.4 on p. 15 into a data frame, and label the rows and columns.

```
> history( ) # View the history.

> q( ) # Quits R.
```

Homework **C1.1*** **:** Clearly label and highlight parts (a), (b), (c), etc., in your source code and output. Using the CLASSSUR.txt data set from Appendix C, pp. 414-416, and using *R*, list your source code and your output.

(a) Read all of the data into a table.

(b) Take a simple random sample of 20 GPAs withOUT replacement.

(c) Take a simple random sample of 100 GPAs WITH replacement.

(d) Compute the mean and standard deviation of each column in the table, while discarding all values of NA. (Note that, say, the standard deviation of gender, is not all that meaningful, but don't worry about that!)

(e) Attach your answers from part (d) to the bottom of the table. (*Hint:* use `rbind`.)

(f) Print your final table.

(g) Display the height and weight of students #11 through #20 ONLY, in a table.

(h) Add 2 inches to the height of student #12 only, using only one command in *R*.

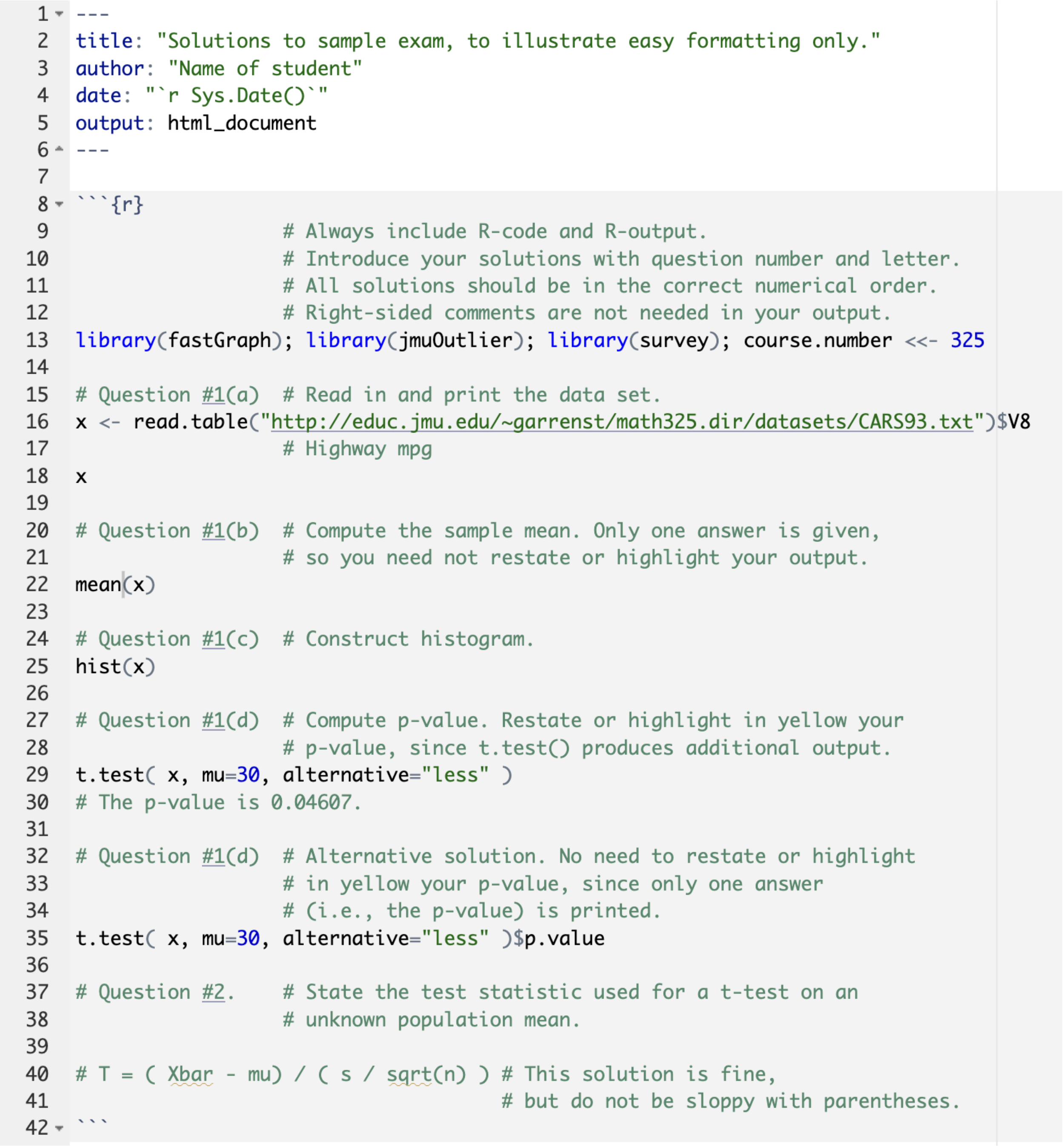
(i) List ‘height’ in centimeters, rather than inches, and display the height and weight of students #11 through #20 only, in a table.

End of Homework C1.1*. □

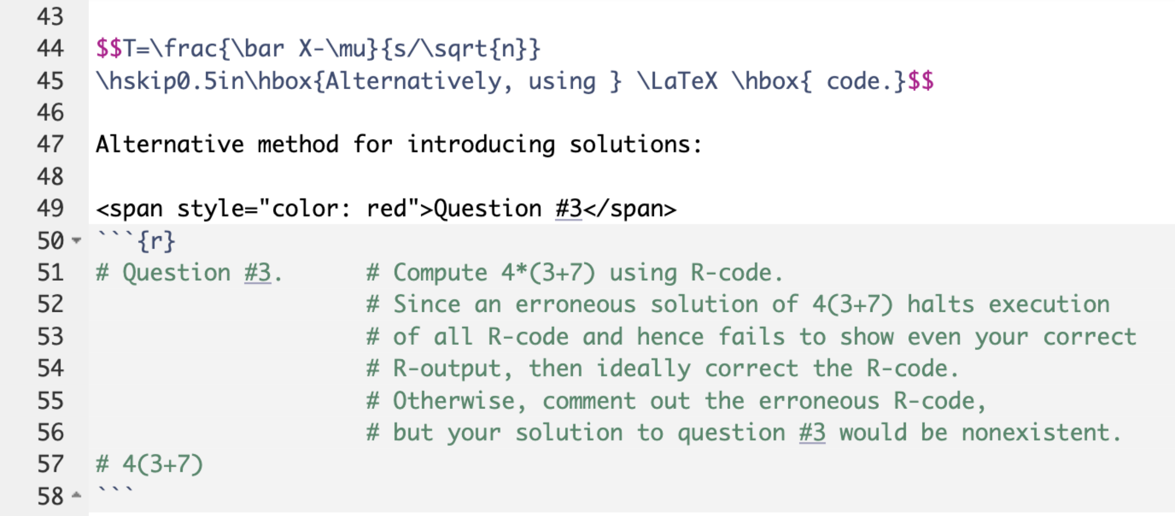
Homework C1.2*: Using the table in exercise 5.29 on p. 158, save the data as a data frame in *R*, provide appropriate row and column names in *R*, and print the data frame with row and column names in *R*.

End of Homework C1.2*. □

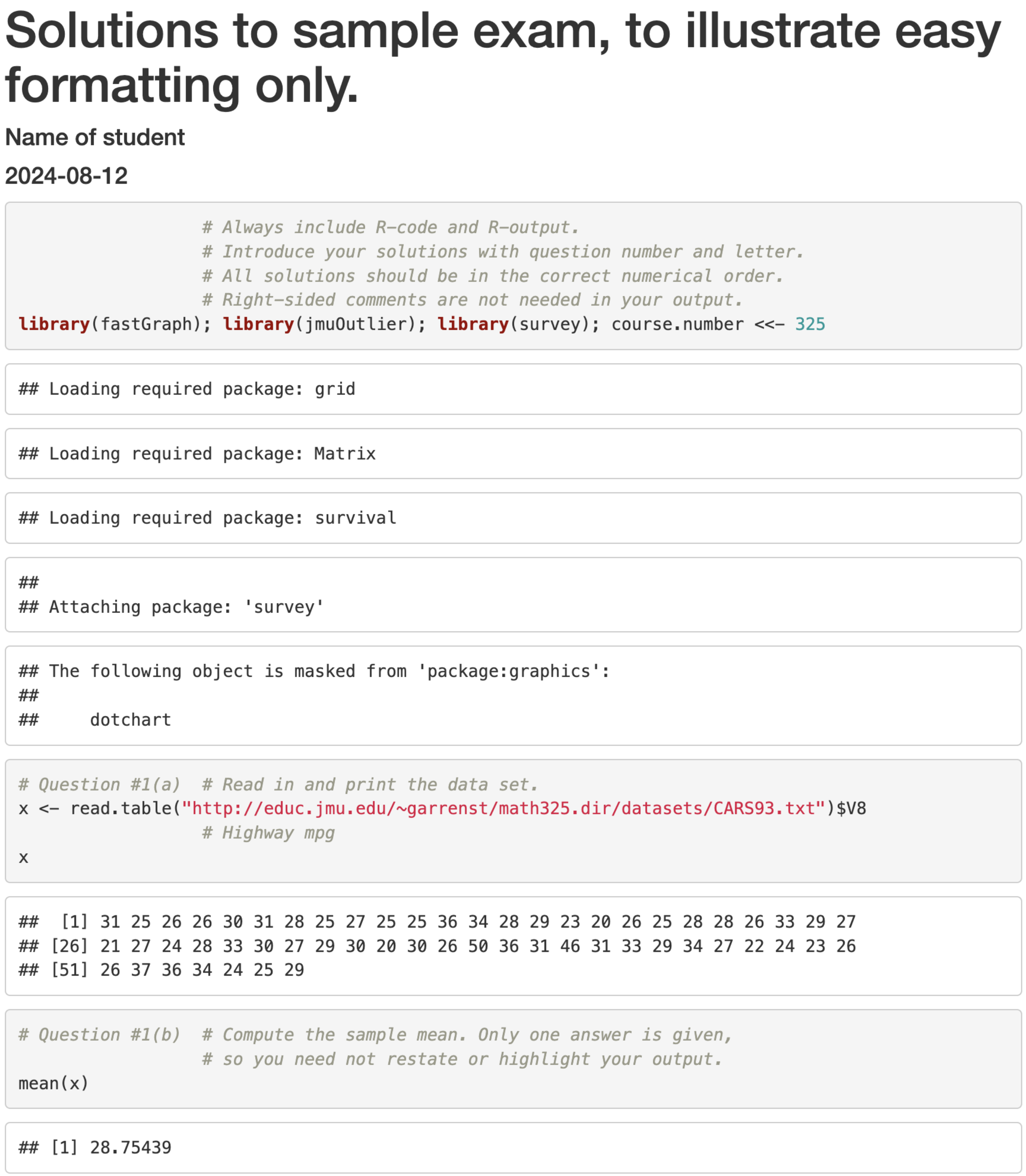
Below are solutions to a sample exam, to illustrate easy formatting only.



~~~
---
title: "Solutions to sample exam, to illustrate easy formatting only."
author: "Name of student"
date: "`r Sys.Date()`"
output: html_document
---

```{r}
                  # Always include R-code and R-output.
                  # Introduce your solutions with question number and letter.
                  # All solutions should be in the correct numerical order.
                  # Right-sided comments are not needed in your output.
library(fastGraph); library(jmuOutlier); library(survey); course.number <<- 325

# Question #1(a)  # Read in and print the data set.
x <- read.table("http://educ.jmu.edu/~garrenst/math325.dir/datasets/CARS93.txt")$V8
                  # Highway mpg
x

# Question #1(b)  # Compute the sample mean. Only one answer is given,
                  # so you need not restate or highlight your output.
mean(x)

# Question #1(c)  # Construct histogram.
hist(x)

# Question #1(d)  # Compute p-value. Restate or highlight in yellow your
                  # p-value, since t.test() produces additional output.
t.test( x, mu=30, alternative="less" )
# The p-value is 0.04607.

# Question #1(d)  # Alternative solution. No need to restate or highlight
                  # in yellow your p-value, since only one answer
                  # (i.e., the p-value) is printed.
t.test( x, mu=30, alternative="less" )$p.value

# Question #2.    # State the test statistic used for a t-test on an
                  # unknown population mean.

# T = ( Xbar - mu) / ( s / sqrt(n) ) # This solution is fine,
                                     # but do not be sloppy with parentheses.
```
~~~



~~~

$$T=\frac{\bar X-\mu}{s/\sqrt{n}}
\hskip0.5in\hbox{Alternatively, using } \LaTeX \hbox{ code.}$$

Alternative method for introducing solutions:

<span style="color: red">Question #3</span>
```{r}
# Question #3.      # Compute 4*(3+7) using R-code.
                    # Since an erroneous solution of 4(3+7) halts execution
                    # of all R-code and hence fails to show even your correct
                    # R-output, then ideally correct the R-code.
                    # Otherwise, comment out the erroneous R-code,
                    # but your solution to question #3 would be nonexistent.
# 4(3+7)
```
~~~



Solutions to sample exam, to illustrate easy formatting only.

Name of student

2024-08-12

```
                 # Always include R-code and R-output.
                 # Introduce your solutions with question number and letter.
                 # All solutions should be in the correct numerical order.
                 # Right-sided comments are not needed in your output.
library(fastGraph); library(jmuOutlier); library(survey); course.number <<- 325
```

```
## Loading required package: grid
```

```
## Loading required package: Matrix
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survey'
```

```
## The following object is masked from 'package:graphics':
##
##     dotchart
```

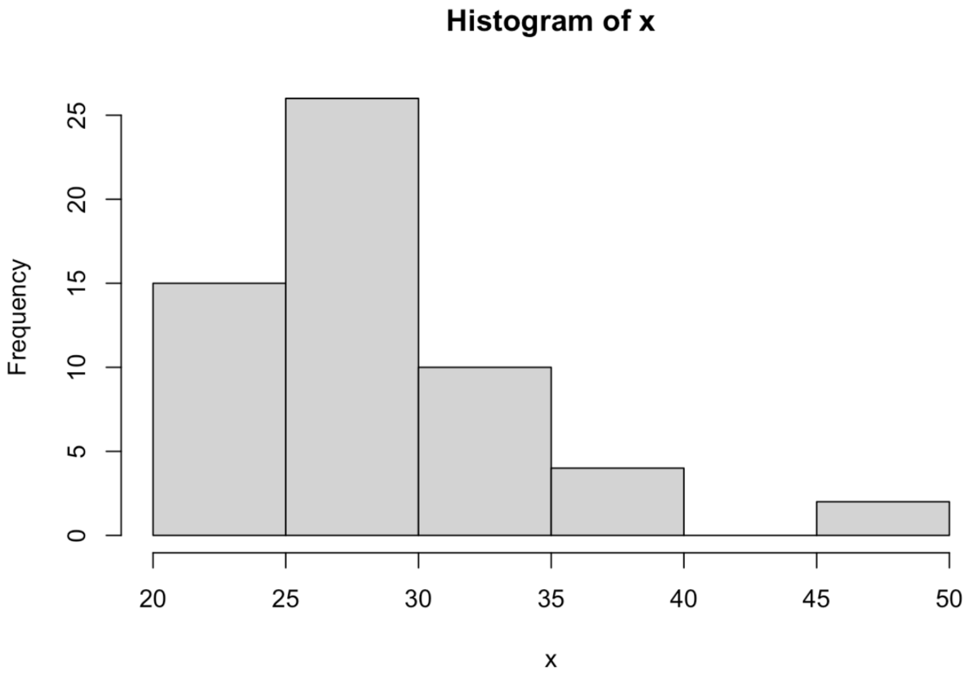
```
# Question #1(a)  # Read in and print the data set.
x <- read.table("http://educ.jmu.edu/~garrenst/math325.dir/datasets/CARS93.txt")$V8
                 # Highway mpg
x
```

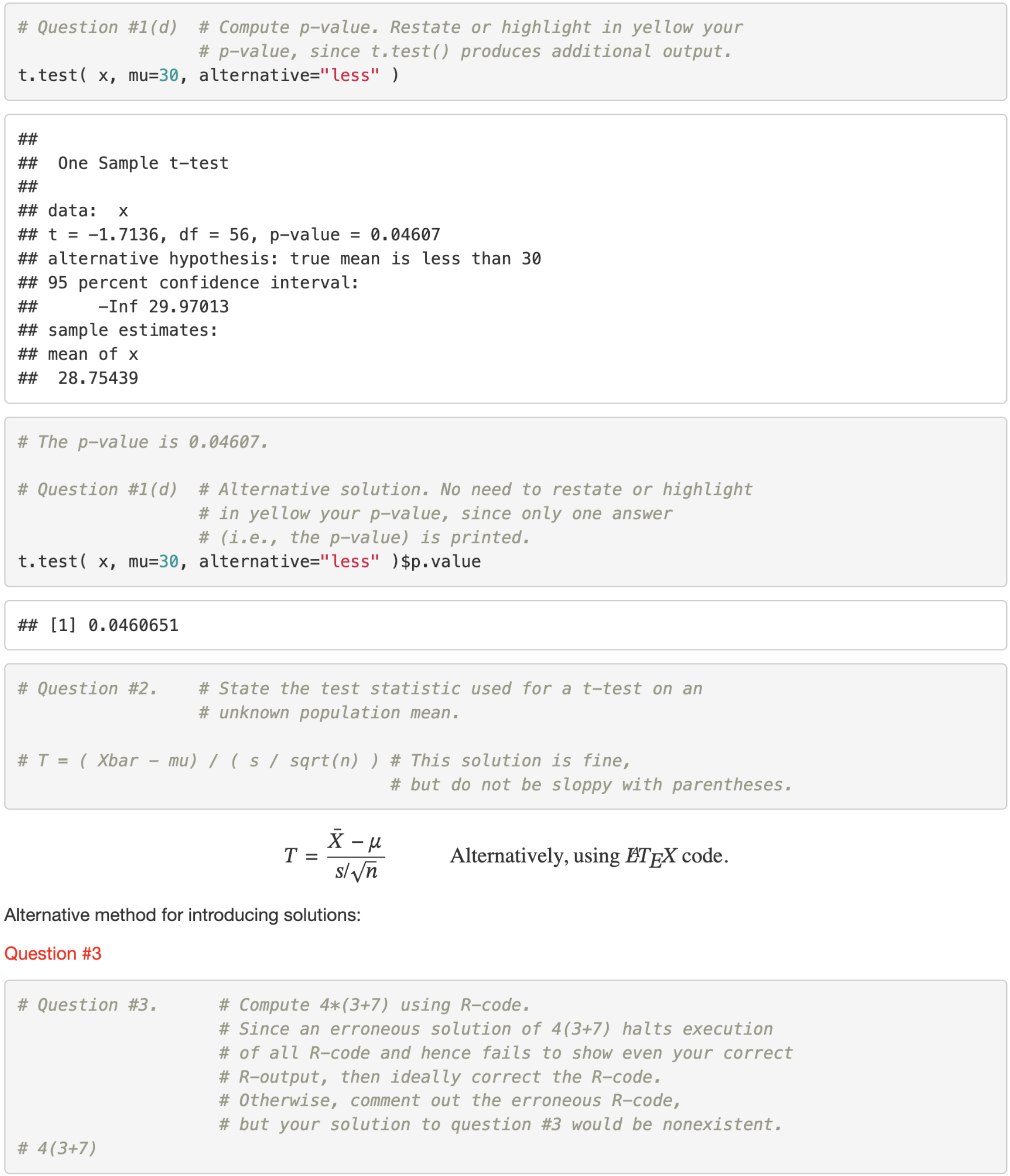
```
##  [1] 31 25 26 26 30 31 28 25 27 25 25 36 34 28 29 23 20 26 25 28 28 26 33 29 27
## [26] 21 27 24 28 33 30 27 29 30 20 30 26 50 36 31 46 31 33 29 34 27 22 24 23 26
## [51] 26 37 36 34 24 25 29
```

```
# Question #1(b)  # Compute the sample mean. Only one answer is given,
                 # so you need not restate or highlight your output.
mean(x)
```

```
## [1] 28.75439
```

```
# Question #1(c)  # Construct histogram.
hist(x)
```





```
# Question #1(d)  # Compute p-value. Restate or highlight in yellow your
                  # p-value, since t.test() produces additional output.
t.test( x, mu=30, alternative="less" )
```

```
##
##  One Sample t-test
##
## data:  x
## t = -1.7136, df = 56, p-value = 0.04607
## alternative hypothesis: true mean is less than 30
## 95 percent confidence interval:
##      -Inf 29.97013
## sample estimates:
## mean of x
##  28.75439
```

```
# The p-value is 0.04607.

# Question #1(d)  # Alternative solution. No need to restate or highlight
                  # in yellow your p-value, since only one answer
                  # (i.e., the p-value) is printed.
t.test( x, mu=30, alternative="less" )$p.value
```

```
## [1] 0.0460651
```

```
# Question #2.    # State the test statistic used for a t-test on an
                  # unknown population mean.

# T = ( Xbar - mu) / ( s / sqrt(n) ) # This solution is fine,
                                     # but do not be sloppy with parentheses.
```

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \qquad \text{Alternatively, using \LaTeX\ code.}$$

Alternative method for introducing solutions:

Question #3

```
# Question #3.      # Compute 4*(3+7) using R-code.
                    # Since an erroneous solution of 4(3+7) halts execution
                    # of all R-code and hence fails to show even your correct
                    # R-output, then ideally correct the R-code.
                    # Otherwise, comment out the erroneous R-code,
                    # but your solution to question #3 would be nonexistent.
# 4(3+7)
```