Lab 10

Alexander Hernandez

09/27/2022

```
library(faraway)
library(TeachingDemos)
library(BSDA)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':
##
##     melanoma

##
## Attaching package: 'BSDA'

## The following object is masked from 'package:TeachingDemos':
##
##     z.test

## The following object is masked from 'package:datasets':
##
##     Orange
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

1) Faraway package. Test whether the participants are younger than 65 years

```
# Null:        u >= 65
# Alternative:  u <  65
t.test(prostate$age,
       alternative="less",
       mu=65)
```

```
##
##  One Sample t-test
##
## data:  prostate$age
## t = -1.5002, df = 96, p-value = 0.06843
## alternative hypothesis: true mean is less than 65
## 95 percent confidence interval:
##     -Inf 65.1215
## sample estimates:
## mean of x
##  63.86598
```

```
# With a p-value of 0.068, which is greater than 0.05
# we fail to reject the null hypothesis.
```

2) Naga Valley Marathon Times by Age and Gender, 2015

a) Import the data in R

```
NAPA = read.csv('C:\\repos\\STAT 50001\\Lab 10\\Napa.csv')
```

b) How many runners are older than 50 years old?

```
nrow(NAPA[NAPA$Age > 50,])
```

```
## [1] 383
```

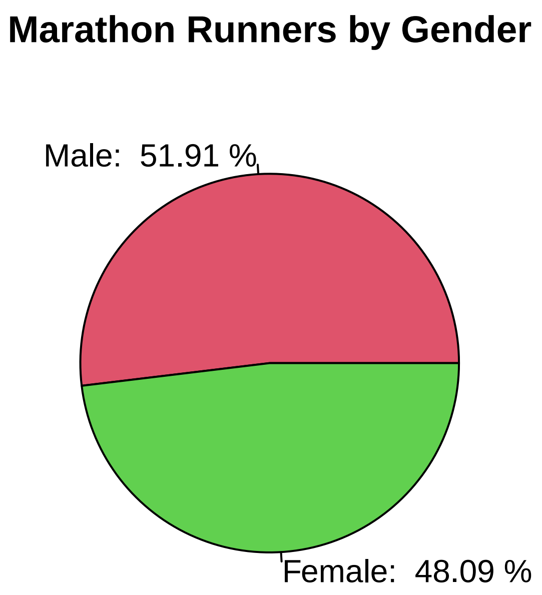
c) Display the age distributions of the runners by gender

```
male = length(which(NAPA$Gender == "M"))
female = length(which(NAPA$Gender == "F"))
total = male + female

male_percent = round((male / total), digits=4) * 100
female_percent = round((female / total), digits=4) * 100

pie_labels = c(paste("Male: ", male_percent, "%"),
               paste("Female: ", female_percent, "%"))

pie(rbind(male, female),
    col=c(2,3),
    labels = pie_labels,
    main = "Marathon Runners by Gender")
```



d) Are men older than women?

```
# Null:         u(m) - u(w) <= 0
# Alternative:  u(m) - u(w) > 0
t.test(NAPA$Age[NAPA$Gender == "M"],
       NAPA$Age[NAPA$Gender == "F"])
```

```
##
##  Welch Two Sample t-test
##
## data:  NAPA$Age[NAPA$Gender == "M"] and NAPA$Age[NAPA$Gender == "F"]
## t = 9.0319, df = 1878.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.444321 5.355064
## sample estimates:
## mean of x mean of y
##  43.44831  39.04862
```

```
# With a p-value of 2.2e-16, we reject the null hypothesis.
```

e) Average completion time is 4.361 hours. Test whether completion time for men is lower than 4.361 hours

```r
# Null:         u >= 4.361
# Alternative:  u < 4.361
t.test(NAPA$Hours[NAPA$Gender == "M"],
       alternative="less",
       mu=4.361)
```

```
##
##  One Sample t-test
##
## data:  NAPA$Hours[NAPA$Gender == "M"]
## t = -6.548, df = 976, p-value = 4.707e-11
## alternative hypothesis: true mean is less than 4.361
## 95 percent confidence interval:
##      -Inf 4.239651
## sample estimates:
## mean of x
##   4.19889
```

```r
# With a p-value of 4.707e-11, we reject the null hypothesis.
```

f) Average age is 41.33. Test whether women are younger than 41.33 years

```r
# Null:         u >= 41.33
# Alternative:  u < 41.33
t.test(NAPA$Age[NAPA$Gender == "F"],
       alternative="less",
       mu=41.33)
```

```
##
##  One Sample t-test
##
## data:  NAPA$Age[NAPA$Gender == "F"]
## t = -6.645, df = 904, p-value = 2.617e-11
## alternative hypothesis: true mean is less than 41.33
## 95 percent confidence interval:
##      -Inf 39.61391
## sample estimates:
## mean of x
##  39.04862
```

```r
# With a p-value of 2.617e-11, we reject the null hypothesis.
```

3) birthwt data

a) Import Data

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

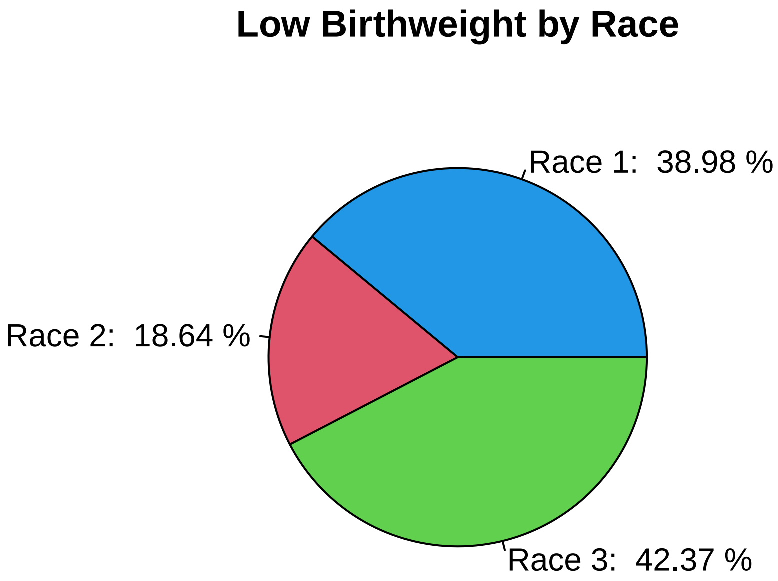
b) Identify proportion of low birthweights based on Race

```
race_birthwt = birthwt$race[birthwt$low == 1]

one_percent = round(100 * length(race_birthwt[race_birthwt == 1]) / length(race_birthwt),
                    digits = 2)
two_percent = round(100 * length(race_birthwt[race_birthwt == 2]) / length(race_birthwt),
                    digits = 2)
three_percent = round(100 * length(race_birthwt[race_birthwt == 3]) / length(race_birthwt),
                      digits = 2)

race_birthwt_labels = c(paste("Race 1: ", one_percent, "%"),
                        paste("Race 2: ", two_percent, "%"),
                        paste("Race 3: ", three_percent, "%"))

pie(table(race_birthwt),
    main="Low Birthweight by Race",
    col=c(4,2,3),
    labels = race_birthwt_labels)
```



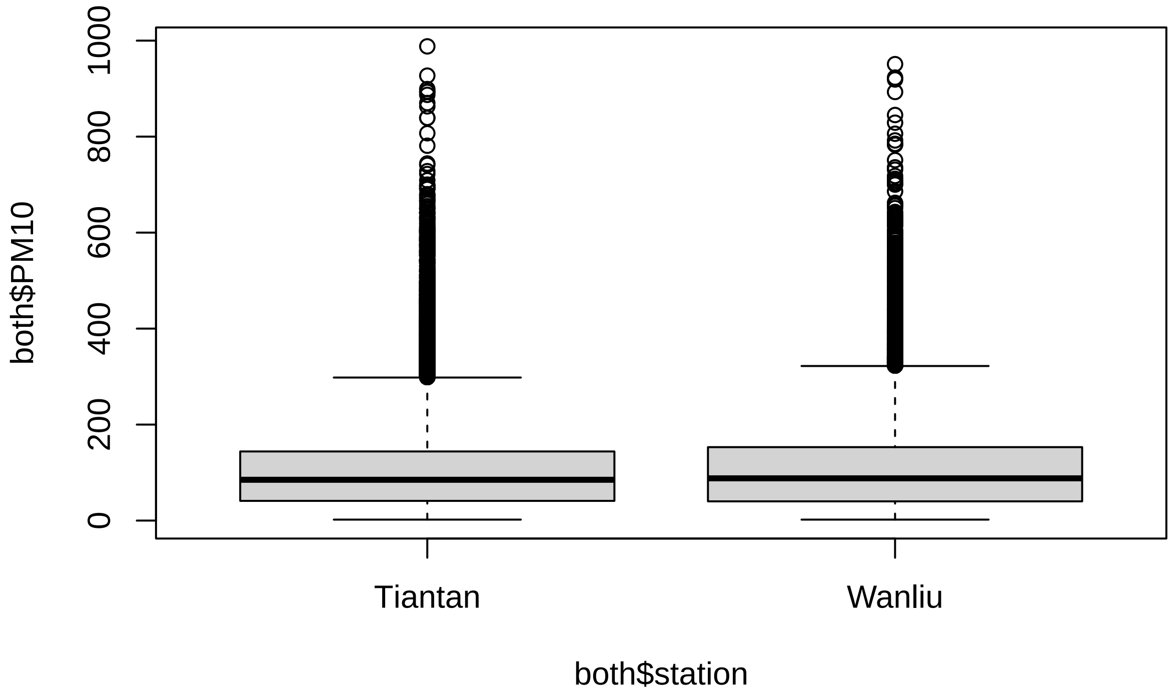
4) Beijing Multi-Site Air-Quality Data

a) Import data from Wanliu and Tiantan into a single dataframe

```
Wanliu = read.csv('C:\\repos\\STAT 50001\\Lab 10\\PRSA_Data_Wanliu.csv')
Tiantan = read.csv('C:\\repos\\STAT 50001\\Lab 10\\PRSA_Data_Tiantan.csv')
both = rbind(Wanliu, Tiantan)
```

b) Display the PM10 values by creating a side-by-side box plot for both stations

```
boxplot(both$PM10 ~ both$station)
```



c) Test for significance difference in PM10 values in Wanlio and Tiantan

```
# Two sample test
# Null:         u(w) - u(t) = 0
# Alternative:  u(w) - u(t) != 0
t.test(both$PM10 ~ both$station)
```

```
##
##  Welch Two Sample t-test
##
## data:  both$PM10 by both$station
## t = -5.9129, df = 69202, p-value = 3.377e-09
## alternative hypothesis: true difference in means between group Tiantan and group Wanliu is not equal
## 95 percent confidence interval:
##  -5.460323 -2.741567
## sample estimates:
## mean in group Tiantan  mean in group Wanliu
##              106.3637              110.4646
```

```
# Given that the p-value is 3.377e^-09, we can conclude that
# we have enough evidence to reject the null hypothesis.
```