Chapter 2

2.1

Suppose,

$X_1 = 1, X_2 = 3, X_3 = 0, X_4 = -2, X_5 = 4, X_6 = -1, X_7 = 5, X_8 = 2, X_9 = 10.$

```
x <- c(1, 3, 0, -2, 4, -1, 5, 2, 10)
```

Find:

a.) $\sum X_i$

```
sum(x)
```

```
[1] 22
```

b.) $\sum_{i=3}^{5} X_i$

```
sum( x[3:5] )
```

```
[1] 2
```

c.) $\sum_{i=1}^{4} X_i^3$

```
sum(x[1:4]^3)
```

```
[1] 20
```

d.) $(\sum X_i)^2$

```
(sum(x))^2
```

```
[1] 484
```

e.) $\sum 3$

```
3 * length(x)
```

```
[1] 27
```

f.) $\sum (X_i - 7)$

```r
sum(x - 7)
```

```
[1] -41
```

g.) $3\sum_{i=1}^{5} X_i - \sum_{i=6}^{9} X_i$

```r
3 * sum( x[1:5] ) - sum( x[6:9] )
```

```
[1] 2
```

h.) $\sum 10X$

```r
sum( 10 * x)
```

```
[1] 220
```

i.) $\sum_{i=2}^{6} iX_i$

```r
i <- 2:6
sum( i * x[i]  )
```

```
[1] 12
```

j.) $\sum 6$

```r
6 * length(x)
```

```
[1] 54
```

2.2

Express the following in summation notation.

a.) $X_1 + \frac{X_2}{2} + \frac{X_3}{3} + \frac{X_4}{4}$

$\ldots = \frac{X_1}{1} + \frac{X_2}{2} + \frac{X_3}{3} + \frac{X_4}{4}$

$\ldots = \sum{}^4{}_{i=1} \frac{X_i}{i}$

b.) $U_1 + U_2^2 + U_3^3 + U_4^4$

$\ldots = U_1^1 + U_2^2 + U_3^3 + U_4^4$

$\ldots = \sum_{i=1}^{4} U_i^i$

c.) $(Y_1 + Y_2 + Y_3)^4$

$(\sum_{i=1}^{3} Y_i)^4$

2.3

Show by numerical example that $\sum X_i^2$ is not necessarily equal to $(\sum X_i{}^2)$.

```
x <- 1:10

sum(x^2)
```

```
[1] 385
```

```
(sum(x))^2
```

```
[1] 3025
```

2.4

Find the mean and median of the following sets of numbers.

a.) -1, 0, 3, 0, 2, -5.

```
x <- c(-1, 0, 3, 0, 2, -5)

mean(x)
```

```
[1] -0.1666667
```

```
median(x)
```

```
[1] 0
```

b.) 2, 2, 3, 10, 100, 1,000

```
x <- c(2, 2, 3, 10, 100, 1000)

mean(x)
```

```
[1] 186.1667
```

```
median(x)
```

```
[1] 6.5
```

2.5

The final exam scores for 15 students are: 73, 74, 92, 98, 100, 72, 74, 85, 76, 94, 89, 73, 76, 99.

Compute the mean, 20% trimmed mean, and median using R.

```
scores <- c(73, 74, 92, 98, 100, 72, 74, 85, 76, 94, 89, 73, 76, 99)

mean(scores)
```

```
[1] 83.92857
```

```
mean(scores, trim = .2)
```

```
[1] 83.1
```

```
median(scores)
```

```
[1] 80.5
```

2.6

The average of 23 numbers is 14.7. What is the sum of these numbers?

```
23 * 14.7
```

```
[1] 338.1
```

2.7

Consider the 10 values: 3, 6, 8, 12, 23, 26, 37, 42, 49, 63.

The mean is $\bar{X} = 26.9$

```
x <- c(3, 6, 8, 12, 23, 26, 37, 42, 49, 63)
mean(x)
```

```
[1] 26.9
```

a.) What is the value of the mean if the largest value, 63, is increased to 100?

```r
x <- c(3, 6, 8, 12, 23, 26, 37, 42, 49, 100)
mean(x)
```

```
[1] 30.6
```

b.) What is the mean if 633 is increased to 1,000?

```r
x <- c(3, 6, 8, 12, 23, 26, 37, 42, 49, 1000)
mean(x)
```

```
[1] 120.6
```

c.) What do these results illustrate about the mean?

The mean is very sensitive to outliers.

2.8

Repeat the previous exercise, only compute the median instead.

```r
x <- c(3, 6, 8, 12, 23, 26, 37, 42, 49, 100)
median(x)
```

```
[1] 24.5
```

b.) What is the mean if 633 is increased to 1,000?

```r
x <- c(3, 6, 8, 12, 23, 26, 37, 42, 49, 1000)
median(x)
```

```
[1] 24.5
```

2.9

In general, how many values must be altered to make the sample mean arbitrarily large?

One.

2.10

What is the minimum number of values that must be altered to make the 20% trimmed mean and sample median arbitrarily large?

mean = $g = (.2N), g + 1$

median = ~$.5N$

2.11

For the values 0, 23, -1, 12, -10, -7, 1, -19, -6, 12, 1, -3 compute the lower and upper quartiles using the ideal fourths.

```
# Used in 11 & 12
idealf <- function( x ) {
   n <- length(x)
   sorted <- sort(x)
   i <- (n / 4) + 5/12
   j <- floor(i)
   h <- i - j
   q1 <- (1 - h)*sorted[j] + h*sorted[j + 1]

   k <- n - j + 1
   q2 <- (1 - h)*sorted[k] + h*sorted[k - 1]

   c(x[1], q1, q2, sorted[n])
}
```

```
x <- c(0, 23, -1, 12, -10, -7, 1, -19, -6, 12, 1, -3)

idealf(x)
```

```
[1]  0.000000 -6.583333  7.416667 23.000000
```

2.12

For the values: -1, -10, 2, 2, -7, -2, 3, 3, -6, 12, -1, -12, -6, 8, 6 compute the lower and upper quartiles (the ideal fourths).

```
x <- c(-1, -10, 2, 2, -7, -2, 3, 3, -6, 12, -1, -12, -6, 8, 6)

idealf(x)
```

```
[1] -1 -6  3 12
```

2.13

Approximately how many values must be altered to make *q2*, the estimate of the upper quartile based on the ideal fourths, arbitrarily large?

About $\frac{1}{4}$

2.14

Argue that the smallest observed value, X_1, satisfies the definition of a measure of location.

The smallest (or largest) value in a set defines the boundaries of the set. This is by definition a measure of location.

2.15

The height of 10 plants is measured in inches and found to be 12, 6, 15, 3, 12, 6, 21, 15, 18 and 12.

```
x <- c(12, 6, 15, 3, 12, 6, 21, 15, 18, 12)

xbar <- mean(x)

stopifnot( sum( x - xbar ) == 0 )
```

Verify that $\sum(X_i - \bar{X}) = 0$

2.16

For the data in the previous exercise, compute the range, variance and standard deviation.

```
n <- length(x)

range <- max(x) - min(x)
variance <- (1/(n-1))*sum( ( x - xbar )^2)
stdDev <- sqrt(variance)
```

$\bar{X} = 12$, Range = 18, $Var = 32$, $\sigma = 5.6568542$

2.17

Use the rules of summation notation to show that it is always the case that $\sum(X_i - \bar{X}) = 0$.

$\bar{X} = \frac{1}{n}\sum X_i$

$\sum X_i - \frac{1}{n}\sum X_i$

$\ldots = \frac{1}{n}\sum X_i - \sum X_i$

$\ldots = \frac{1}{n}(\sum X_i - X_i)$

$\ldots = \frac{1}{n}(0)$

$\ldots = 0$

2.18

Seven different thermometers were used to measure the temperature of a substance. The reading in degrees Celsius are -4.10, -4.13, -5.09, -4.08, -4.10, -4.09 and -4.12.

```
x <- c(-4.10, -4.13, -5.09, -4.08, -4.10, -4.09, -4.12)
n <- length(x)
xbar <- 1/n*sum(x)
variance <- (1/(n-1))*sum( (x - xbar)^2 )
stdDev <- sqrt(variance)
```

Find the variance and standard deviation.

$Var = 0.1393619, \sigma = 0.3733121$

2.19

A weightlifter's maximum bench press (in pounds) in each of 6 successive weaks was 280, 295, 275, 305, 300, 290.

Find the standard deviation.

```
x <- c(280, 295, 275, 305, 300, 290)
n <- length(x)
xbar <- 1/n*(sum(x))
variance <- (1/(n-1))*sum( (x - xbar)^2)
sqrt(variance)
```

```
[1] 11.58303
```

2.20

For the values,

20, 121, 132, 123, 145, 151, 119, 133, 134, 130, 200

use the classic outlier detection rule to determine whether any outliers exist.

```
x <- c(20, 121, 132, 123, 145, 151, 119, 133, 134, 130, 200)

x[ abs(x - mean(x)) / sd(x) > 2 ]
```

```
[1] 20
```

2.21

Apply the boxplot rule and the MAD-median rule using the values in the preceding exercise. Note that the results differ, compared to using the classic rule.

```
x[ abs( x - median(x) ) / mad(x) > 2.27 ]
```

```
[1]  20 200
```

Explain why this happened.

The classic method masks the value 200 as an outlier.

2.22

Consider the values,

0, 121, 132, 123, 145, 151, 119, 133, 134, 130, 250.

```
x <- c(0, 121, 132, 123, 145, 151, 119, 133, 134, 130, 250)

x[ ( abs(x - mean(x)) /sd(x) ) > 2 ]
```

```
[1]   0 250
```

Are the values 0 and 250 declared outliers using the classic outlier detection rule?

Yes.

2.23

Verify that for the data in the previous exercise, the boxplot rule declares the values 0 and 250 outliers.

```
bounds <- quantile(x)[c(2,4)]

lower <- bounds[1] - 1.5*(bounds[2] - bounds[1])
upper <- bounds[2] + 1.5*(bounds[2] - bounds[1])

x[ x < lower | x > upper]
```

```
[1]   0 250
```

2.24

Consider the values

20, 121, 132, 123, 145, 151, 119, 133, 134, 240, 250

Verify that no outliers are found using the classic outlier detection rule.

```
x <- c(20, 121, 132, 123, 145, 151, 119, 133, 134, 240, 250)

stopifnot( length( x[ ( abs( x - mean(x) ) / sd(x) ) > 2] ) == 0 )
```

2.25

Verify that for the data in the previous exercise, the boxplot rule declares the values 20, 240 and 250 outliers.

```
x <- c(20, 121, 132, 123, 145, 151, 119, 133, 134, 240, 250)

bounds <- quantile(x)[c(2,4)]

lower <- bounds[1] - 1.5*(bounds[2] - bounds[1])
upper <- bounds[2] + 1.5*(bounds[2] - bounds[1])

x[ x < lower | x > upper]
```

```
[1]  20 240 250
```

2.26

What do the last three exercises suggest about the boxplot rule versus the classic rule for detecting outliers?

The classic boxplot rule masks outliers.

2.27

What is the typical pulse rate (beats per minute) among adults? Imagine that you sample 21 adults, measure their pulse rate, and get:

80, 85, 81, 75, 77, 79, 74, 86, 79, 55, 82, 89, 73, 79, 83, 82, 88, 79, 77, 81, 82

```
x <- c(80, 85, 81, 75, 77, 79, 74, 86, 79, 55, 82, 89, 73, 79, 83, 82, 88, 79, 77, 81, 82)
```

compute the 20% trimmed mean.

```
n <- length(x)
o <- sort(x)
g <- floor(.2*n)
t <- o[ (g+1):(n-g) ]
tbar <- 1/(n - 2*g)*sum(t)

stopifnot(tbar == mean(x, trim = .2))
```

2.28

For the observations,

21, 36, 42, 24, 25, 36, 35, 49, 32

```
x <- c(21, 36, 42, 24, 25, 36, 35, 49, 32)
n <- length(x)

xbar <- 1/n * sum(x)

o <- sort(x)
g <- floor(.2*n)
t <- o[ (g+1):(n-g) ]
tbar <- 1/(n - 2*g)*sum(t)

M <- o[ (n + 1)/2 ]
```

Verify that the sample mean, 20% trimmed mean, and median are $\bar{X} = 33.33, \bar{X}_t = 32.9, and\ M = 35$

$\bar{X} = 33.3333333$, $\bar{X}_t = 32.8571429$, $M = 35$

2.29

The largest observation in the last problem is 49. If 49 is replaced by the value 200, verify that the sample mean is now $\bar{X} = 50.1$ but the 20% trimmed mean and median are not changed.

```r
x <- c(21, 36, 42, 24, 25, 36, 35, 200, 32)

xbar <- 1/n * sum(x)

o <- sort(x)
g <- floor(.2*n)
t <- o[ (g+1):(n-g) ]
tbar <- 1/(n - 2*g)*sum(t)

M <- o[ (n + 1)/2 ]
```

$\bar{X} = 50.1111111$, $\bar{X}_t = 32.8571429$, $M = 35$

2.30

For the data in Exercise 28, what is the minimum number of observations that must be altered so that the 20% trimmed mean is greather than 1,000?

$g + 1 = 2$

2.31

Repeat the previous exercise, but use the median instead.

N = 4

What does this illustrate about the resistance of the mean, median and 20% trimmed mean?

The mean has the least resistance, followed by the trimmed mean and then the median.

2.32

For the observations,

6, 3, 2, 7, 6, 5, 8, 9, 11

Use R to verify that the sample mean, 20% trimmed mean, and median are $\bar{X} = 6.5, \bar{X}_t = 6.7\ and\ M = 6.5$, respectfully.

```r
x <- c(6, 3, 2, 7, 6, 5, 8, 9, 8, 11)

mean(x)
```

```
[1] 6.5
```

```
mean(x, trim = .2)
```

```
[1] 6.666667
```

```
median(x)
```

```
[1] 6.5
```

2.33

In general, when you have *n* observations, what is the minimum number of values that must be altered to make the 20% trimmed mean grow as larage as you want?

floor(n*.2) + 1

2.34

A class of fourth graders was asked to bring a pumpkin to school. Each of the 29 students counted the number of seeds in their pumpkin, and the results were:

250, 220, 281, 247, 230, 209, 240, 160, 370, 274, 210, 204, 243, 251, 190, 200, 130, 150, 177, 475, 221, 350, 224, 163, 272, 236, 200, 171, 98

Use R to compute the sample mean, 20% trimmed mean, median and MOM.

```
x <- c(250, 220, 281, 247, 230, 209, 240, 160, 370, 274, 210, 204, 243, 251, 190,
       200, 130, 150, 177, 475, 221, 350, 224, 163, 272, 236, 200, 171, 98)

xbar <- mean(x)
tbar <- mean(x, trim = .2)
M <- median(x)
MOM <-  mean( x[ abs( x - median(x) ) / mad(x) < 2.27 ] )
```

$\bar{X} = 229.1724138$, $\bar{X}_t = 220.7894737$, $M = 221$, $MOM = 214.12$

2.35

Compute the 20% Winsorized values for the observations:

21, 36, 42, 24, 25, 36, 35, 49, 32

```r
x <- c(21, 36, 42, 24, 25, 36, 35, 49, 32)

winsorize <- function( x, trim = .2) {
   n <- length(x)

   o <- sort(x)
   g <- floor(trim*n)

   o[1:(g+1)] <- o[(g+1)]
   o[(n-g):n] <- o[n-g]

   o
}

winsorize(x)
```

```
[1] 24 24 25 32 35 36 36 42 42
```

2.36

For the observations in the pervious problem, use R to verify that the 20% Winsorized variance is 51.36.

```r
Wvar <- var(winsorize(x))
```

$W_{var} = 51.3611111$

2.37

In the previous problem, would you expect the sample variance to be larger or smaller than 51.36?

Larger, Winsorizing pulls in the extremes.

Verify your answer.

```r
var(x)
```

```
[1] 81
```

2.38

In general, will the Winsorized sample variance, s_w^2, be less than the sample variance, s^2?

Yes.

2.39

For the observations,

6, 3, 2, 7, 6, 5, 8, 9, 8, 11

verify that the sample variance and 20% Winsorized variance are 7.4 and 1.8, respectfully.

```
x <- c(6, 3, 2, 7, 6, 5, 8, 9, 8, 11)

variance <- var(x)
```

$Var = 7.3888889, Var_w = 51.3611111$

2.40

Consider again the number of pumpkin seeds given in Exercise 34.

Compute the 20% Winsorized variance.

```
x <- c(250, 220, 281, 247, 230, 209, 240, 160, 370, 274, 210, 204, 243, 251, 190,
       200, 130, 150, 177, 475, 221, 350, 224, 163, 272, 236, 200, 171, 98)

Wvar <- var( winsorize(x) )
```

$W_{var} = 1375.6059113$

2.41

Snedecor and Cochran (1967) report results from an experiment dealing with weight gaim in rats as a function of source and amount of protein.

One of the groups was fed beef with a low amount of protein. The weight gains were:

90, 76, 90, 64, 86, 51, 72, 90, 95, 78

Compute the 20% trimmed mean and 20% Winsorized variance.

```
x <- c(90, 76, 90, 64, 86, 51, 72, 90, 95, 78)

tbar <- mean(x, trim = .2)
wvar <- var(winsorize(x))
```

$\bar{X}_t = 82, Var_w = 69.1555556$