Chapter 3

3.1

Based upon a sample of 100 individuals, the values 1, 2, 3, 4, 5 are observed with relative frequencies 0.2, 0.3, 0.1, 0.25, 0.15, respectfully.

Compute the mean, variance and standard deviation.

```
values <- c(1, 2, 3, 4, 5)
frequencies <- c(0.2, 0.3, 0.1, 0.25, 0.15)

stopifnot(sum(frequencies) == 1)

n <- 100

xbar <- sum(values * frequencies)
variance <- n/(n - 1) * sum( ((values - xbar)^2) * frequencies )
stdDev <- sqrt(variance)
```

$\bar{X} = 2.85, \sigma^2 = 1.9469697, \sigma = 1.3953386$

3.2

Fifty individual are rated on how open-mided they are. The ratings have the values 1, 2, 3, 4, and the corresponding relative frequencies are 0.2, 0.24, 0.4, 0.16, respectively.

Compute the mean, variance and standard deviation.

```
n <- 50
x <- c(1, 2, 3, 4)
f <- c(0.2, 0.24, 0.4, 0.16)

stopifnot(sum(f) == 1)

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

$\bar{X} = 2.52, \sigma^2 = 0.9893878, \sigma = 0.9946797$

3.3

For the values 0, 1, 2, 3, 4, 5, 6, the corresponding relative frequencies based on a sample of 10,000 observations are 0.015625, 0.093750, 0.234375, 0.312500, 0.234375, 0.312500, 0.234375, 0.093750, 0.015625, respectively.

Determine the mean, variance and standard deviation.

```
n <- 10e3
x <- 0:6
f <- c(0.015625, 0.093750, 0.234375, 0.312500, 0.234375, 0.093750, 0.015625)

stopifnot(sum(f) == 1)

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

$\bar{X} = 3, \sigma^2 = 1.50015, \sigma = 1.2248061$

3.4

For a local charity, the donations in dollars recieved during the last month were 5, 10, 15, 20, 25, 50, having the frequencies 20, 30, 10, 40, 50, 5, respectively.

Compute the mean, variance and standard deviation.

```
x <- c(5, 10, 15, 20, 25, 50)
d <- c(20, 30, 10, 40, 50, 5)
n <- sum(d)
f <- d / n

stopifnot( sum(f) == 1)

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

$\bar{X} = 18.3870968, \sigma^2 = 85.0439883, \sigma = 9.2219297$

3.5

The values 1, 5, 10, 20 have the frequencies 10, 20, 40, 30.

Compute the mean, variance and standard deviation.

```
x <- c(1, 5, 10, 20)
c <- c(10, 20, 40, 30)
n <- sum(c)
```

```
f <- c / n

stopifnot(sum(f) == 1)

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

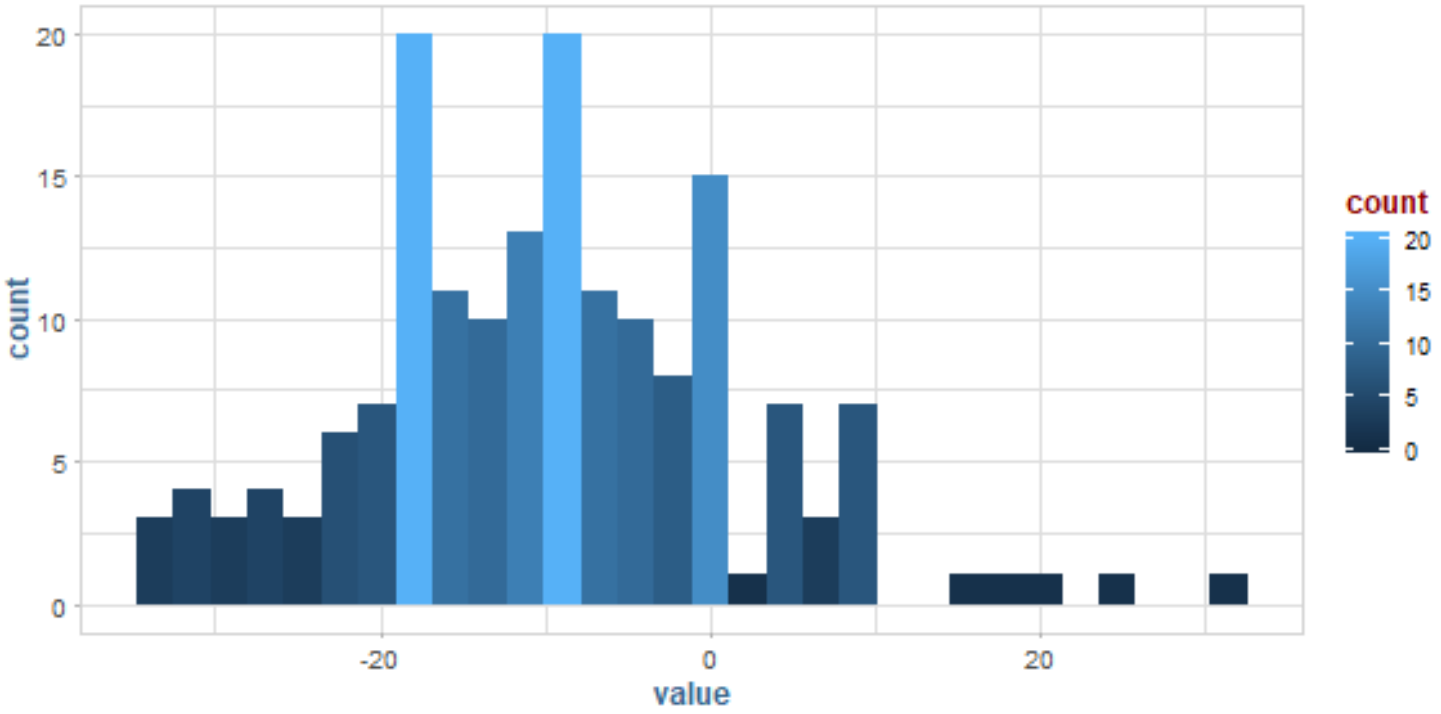
$\bar{X} = 11.1, \sigma^2 = 42.3131313, \sigma = 6.5048544$

3.6

For the data in Table 2.1, dealing with changes in cholesterol levels, create a histogram with R.

```
data <- data.table::fread(paste0(data.dir, "ibtable2_1_dat.txt"), fill = T)

suppressWarnings({print(
   ggplot(data.table(value = as.vector(as.matrix(data))), aes(value)) +
      geom_histogram(aes(fill = ..count..), bins = 30)
)})
```



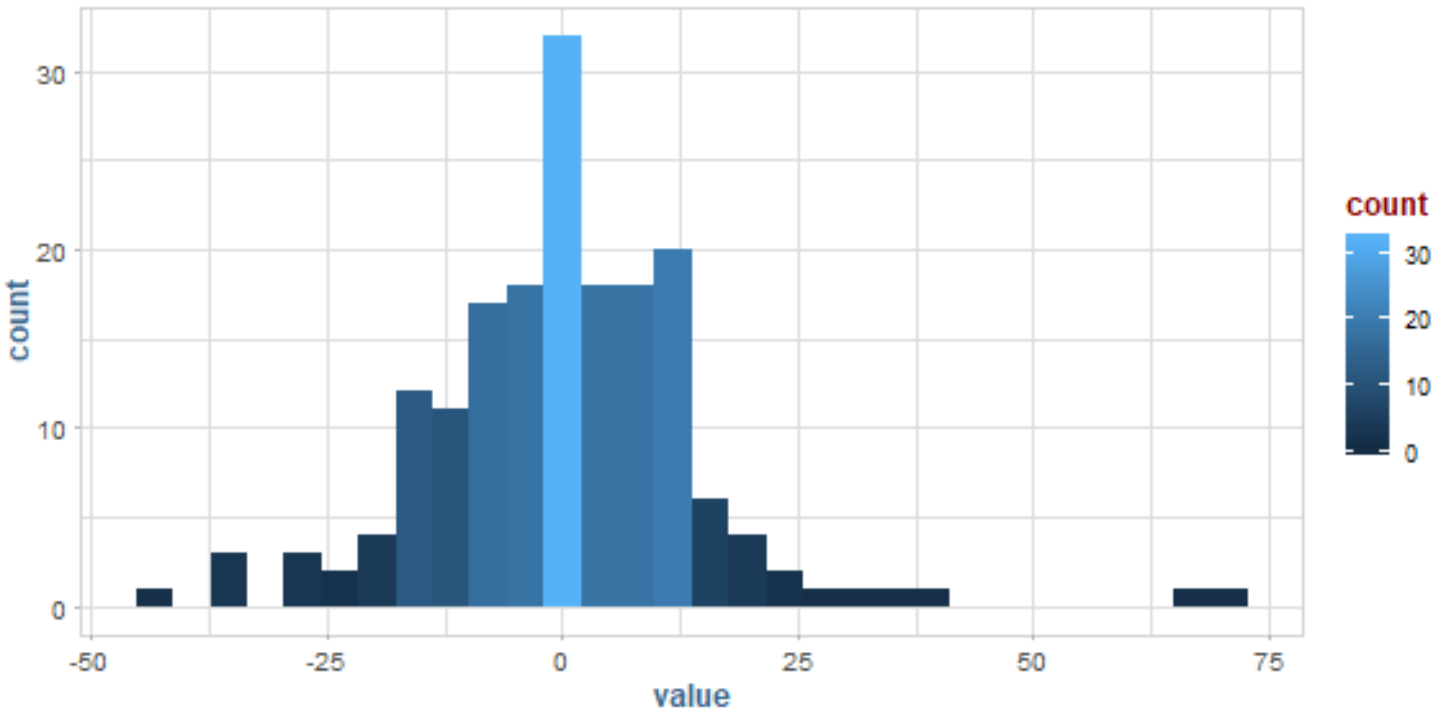
3.7

For the data in Table 2.2, create a histogram using R, and speculate about whether values less than zero are outliers.

```
data <- data.table::fread(paste0(data.dir, "ibtable2_2_dat.txt"), fill = T)

values <- data.table(value = as.vector(as.matrix(data)))
values <- values[!is.na(values$value)]

suppressWarnings({print(
   ggplot(values, aes(value)) +
      geom_histogram(aes(fill = ..count..), bins = 30)
)})
```



Below -45 & above 50 appear to be outliers.

Then compare your answer to the result obtained using *outbox*.

```
# MAD outliers

values[ abs(values$value - median(values$value)) / mad(values$value) > 2.27 ]
```

```
   value
1:    68
2:    30
```

```
 3:   -36
 4:   -27
 5:   -35
 6:   -28
 7:    71
 8:    34
 9:   -43
10:   -27
11:    39
12:   -36
```

```
# Classical Outliers

values[ abs(values$value - mean(values$value)) / sd(values$value) > 2]
```

```
   value
1:    68
2:    30
3:   -36
4:   -35
5:    71
6:    34
7:   -43
8:    39
9:   -36
```

3.8

The heights of 30 male Egyptian skulls from 4,000 BC are reported by Thomson and Randall-Maciver (1905) to be:

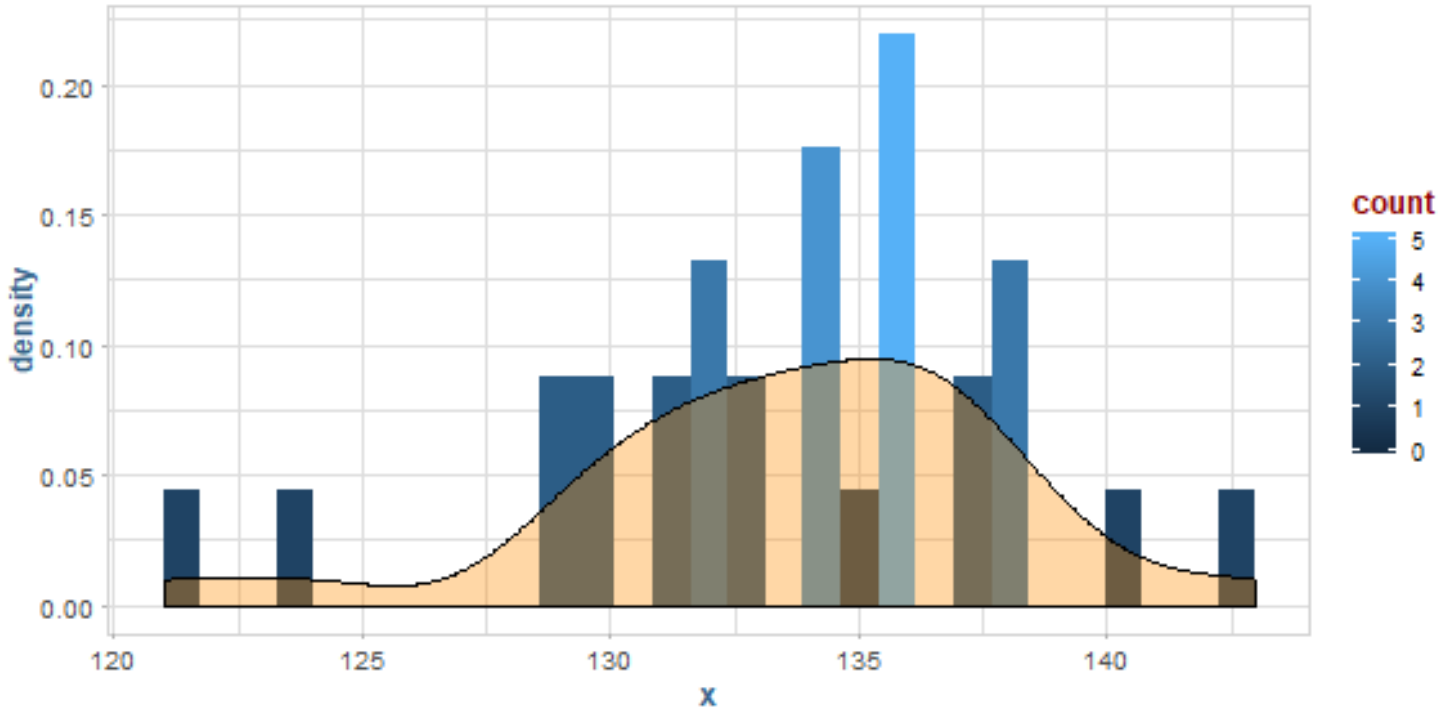
121, 124, 129, 130, 130, 131, 131, 132, 132, 132, 133, 133, 134, 134, 134, 134, 135, 136, 136, 136, 136, 137, 137, 138, 138, 138, 140, 143.

```
x <- c(121, 124, 129, 129, 130, 130, 131, 131, 132, 132, 132, 133, 133, 134, 134, 134,
134, 135, 136, 136, 136, 136, 136, 137, 137, 138, 138, 138, 140, 143)

n <- length(x)

stopifnot(n == 30)

ggplot(data.table(value = x), aes(x)) +
   geom_histogram(aes(y = ..density.., fill = ..count..), bins = 30) +
   geom_density(fill = "darkorange", alpha = .35)
```



Create a histogram.

Find outliers with classic rule and MAD-median rule.

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

```
[1] 121 124 143
```

```
# Classical Outliers

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

```
[1] 121 124 143
```

3.9

For the data in the previous exercise, verify that the classic outlier detection rule, given by *2.6*, finds three outliers, which match the values flagged as outliers by the MAD-median rule.

Confirmed 3 outliers with both rules

Despite this result, what are the concerns with the classic outlier detection rule?

There appears to be masking going on, as there are clearly more than 3 outliers.

3.10

What do Exercises 6 and 7 suggest about using a histogram to detect outliers?

Histograms are not always useful in detecting outliers

3.11

Table 3.5 shows the exam scores for 27 students.

```
scores <- c(83, 69, 82, 72, 63, 88, 92, 81, 54,
            57, 79, 84, 99, 74, 86, 71, 94, 71,
            80, 51, 68, 81, 84, 92, 63, 99, 91)
```

Create a stem-and-leaf display.

```
stem(scores)
```

```
  The decimal point is 1 digit(s) to the right of the |

  5 | 14
  5 | 7
  6 | 33
  6 | 89
  7 | 1124
  7 | 9
  8 | 0112344
  8 | 68
  9 | 1224
  9 | 99
```

3.12

If the leaf is the hundredths digit, what is the stem for the number 34.679?

34.6

3.13

Consider the values 5.134, 5.532, 5.869, 5.809, 5.268, 5.495, 5.142, 5.483, 5.329, 5.149, 5.240, 5.823.

If the leaf is taken to be the tenths digit, why would this make an uninteresting stem-and-leaf display?

```
x <- c(5.134, 5.532, 5.869, 5.809, 5.268, 5.495, 5.142, 5.483,
5.329, 5.149, 5.240, 5.823)
```

There would only be one stem:

```
stem(x, scale = 1/10)
```

```
  The decimal point is at the |

  4 | 111233555889
```

3.14

For the boxplot in Figure *3.11*, determine approximately, the quartiles, the interquartile range, and the median. *M = 80, Q1/Q2 = 50/121, IQR = 71*

Approximately how large is the largest value not declared an outlier? *215*

3.15

In Figure *3.11*, about how large must a value be to be declared an outlier? *x > 227.5*

How small? *x < -56.5*

3.16

Use R to create both a boxplot and a plot of the relative frequencies using the film data in Table 3.1.

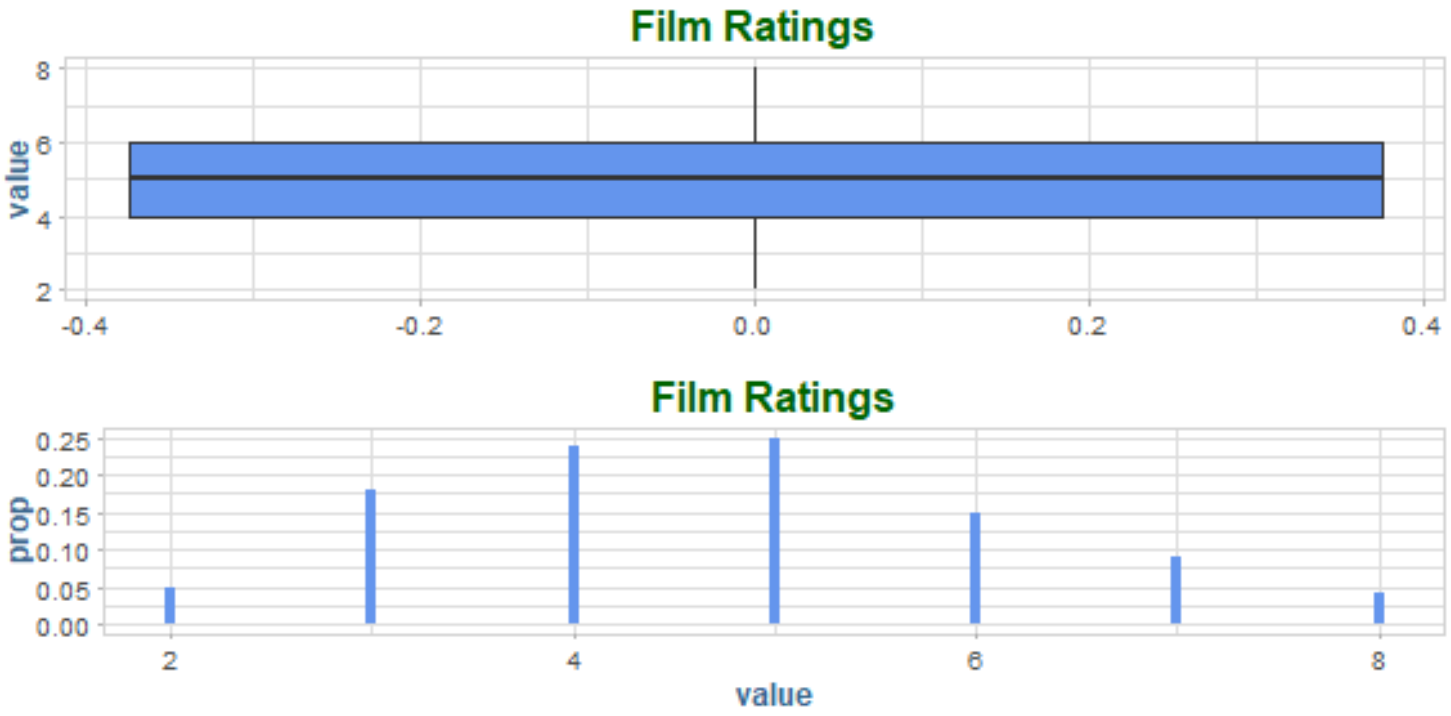
```
data <- data.table::fread(paste0(data.dir, "film_dat.txt"), fill = T)

values <- data.table(value = as.vector(as.matrix(data)))
values <- values[!is.na(values$value)]

p1 <- ggplot(values, aes(y = value)) +
   geom_boxplot(fill = "cornflowerblue") +
   labs(title = "Film Ratings")

p2 <- ggplot(values, aes(x = value)) +
   geom_bar(aes(y = ..prop..), fill = "cornflowerblue", width = 0.05) +
   labs(title = "Film Ratings")

grid.arrange(p1, p2, nrow = 2)
```



3.17

Use R to create a boxplot and a kernel density estimate using the data in table 3.2.

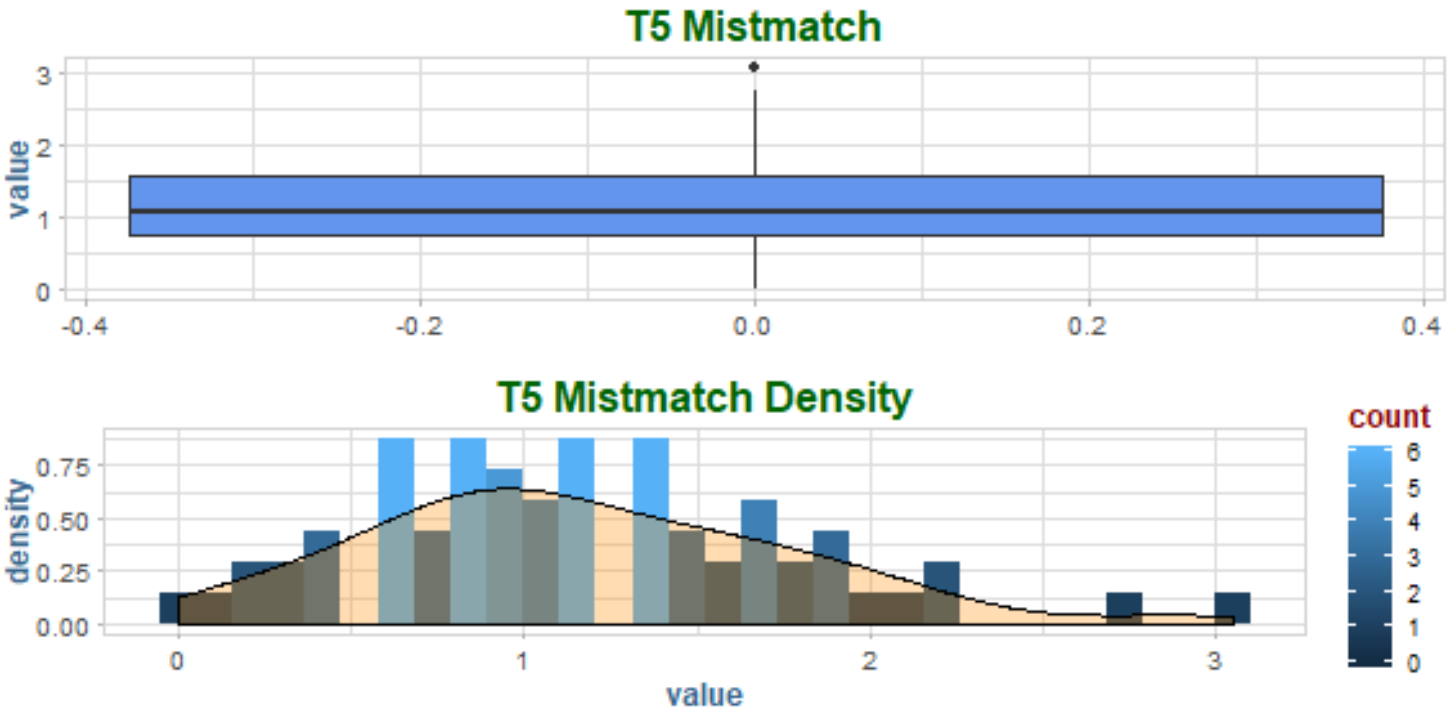
```
data <- data.table::fread(paste0(data.dir, "mismatch_dat.txt"), fill = T)

values <- data.table(value = as.vector(as.matrix(data)))
values <- values[!is.na(values$value)]

p1 <- ggplot(values, aes(y = value)) +
   geom_boxplot(fill = "cornflowerblue") +
   labs(title = "T5 Mistmatch")

p2 <- ggplot(values, aes(x = value)) +
   geom_histogram(aes(y = ..density.., fill = ..count..), bins = 30) +
   geom_density(fill = "darkorange", alpha = .3) +
   labs(title = "T5 Mistmatch Density")

grid.arrange(p1, p2, nrow = 2)
```



3.18

Describe a situation where the sample histogram is likely to give a good indication of the population histogram based on 100 observations.

The population follows a symmetric distribution and outliers are rare.

3.19

Comment generally on how large of a sample size is needed to ensure that the sample histogram will likely provide a good indication of the population histogram?

There are numerical estimators avalaible using bootstrapping techniques, however, a general rule of thumb is > 100.

3.20

When trying to detect outliers, discuss the relative merits of using a histogram vs a boxplot.

A boxplot is generally better at showing outliers.

3.21

A sample histogram indicates that the data are highly skewed to the right.

Is this a reliable indication that if all individuals of interest could be measured, the resulting histogram would also be highly skewed?

Not always. A population could be symetric and a given sample could be skewed.