Cary REU R/stats workshop #2: working with data in R

SE Bowden

June 19, 2017

Preparing your data for analysis

Think about how your data need to be organized *before* you collect data

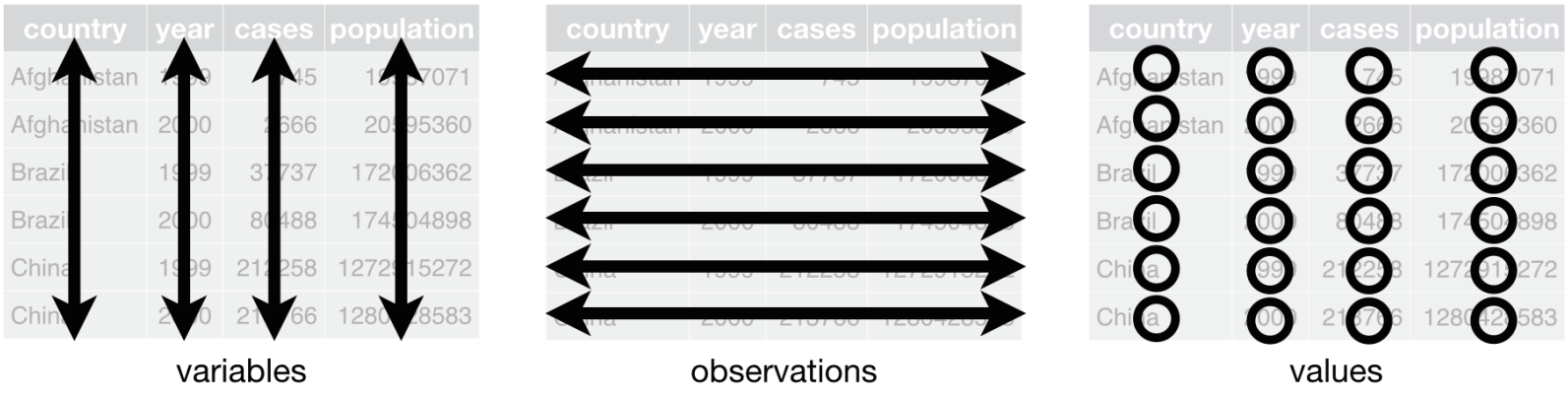


Figure 1: Tidy data

1. Each **variable** should have its own *column*
2. Each **observation** should have its own *row*
3. Each **value** should have its own *cell*

Learn more about **Tidy Data** here: http://r4ds.had.co.nz/tidy-data.html

Reading and inspecting your data

When your data are **tidy** and ready for analysis, the first step is to "read" your data into R. You'll need to set your working directory before you run the next line of code.

Go to Session > Set working directory > choose directory. This tells R where to find the files you want to use.
Note: The folder you select will appear empty (no files listed). This is because we're only browsing for a *folder* and not a *file*

Set your working directory using the instructions above. You must do this every time you open a new R or RStudio session. A line of code similar to the one below will print to the console when your directory is set

```
setwd("C:/Users/Sarah/Dropbox/CIES postdoc/Cary REU workshops 2017")
```

We read in data using the function `read.csv()`. You need to specify the FULL file name, including spaces/underscores and file extension, in quotations. `header=TRUE` is an arguement in the function that tells R if you have column names (TRUE) or not (FALSE).

```
mosq_data <- read.csv("workshop2_data.csv", header = TRUE)
```

Reading and inspecting your data

Now that we've read our data into R, we'll review some common functions that you might use to inspect your data.

Preview the first few lines of your dataset: `head(mydataset)`

```
# this prints the first six lines of your dataset to the console
head(mosq_data)
##   Species Sex Abd_length Dev_time
## 1      AL   M     0.0463        8
## 2      AL   F     0.0491       14
## 3      AL   M     0.0509        8
## 4      AL   M     0.0512        8
## 5      AG   M     0.0544        7
## 6      AL   F     0.0550        9
```

Preview the last few lines of your dataset: `tail(mydataset)`

```
# this prints the last six lines of your dataset to the console
tail(mosq_data)
##      Species Sex Abd_length Dev_time
## 3638      CQ   M     0.1468        8
## 3639      CQ   M     0.1473        9
## 3640      CQ   F     0.1477        7
## 3641      CQ   M     0.1503        9
## 3642      CQ   F     0.1534        9
## 3643      CQ   F     0.1760        9
```

Reading and inspecting your data

Get the dimensions (number of rows, number of columns) of your dataset: `dim(mydataset)`. **NOTE**: `dim()` only works on 2-dimensional datasets (e.g., >1 row *AND* >1 column).

```
# this prints the dimensions of your dataset in (n.rows, n.columns)
dim(mosq_data)
## [1] 3643    4
```

Print a list of the column names in your dataset: `names(mydataset)`

```
# this prints all column names in your dataset
names(mosq_data)
## [1] "Species"    "Sex"        "Abd_length" "Dev_time"
```

Look at a summary of your dataset: `summary(mydataset)`

```
summary(mosq_data)
##  Species   Sex         Abd_length        Dev_time
##  AG:1493   F:1555   Min.   :0.0463   Min.   : 6.000
##  AL:1532   M:2088   1st Qu.:0.0948   1st Qu.: 8.000
##  CQ: 618            Median :0.1046   Median : 8.000
##                     Mean   :0.1045   Mean   : 8.261
##                     3rd Qu.:0.1145   3rd Qu.: 9.000
##                     Max.   :0.1760   Max.   :16.000
```

Summarizing your data

The previous function summarizes each column in your dataset. However, we can also access and summarize each column individually. To pull out a single column in a dataset, we use the $ operator to separate the name of the dataset and the name of the column - e.g., `mydataset$mycolumnname`. We can use this format to run basic summary statistics on a single column of data.

```
# use the $ operator to access a single column
summary(mosq_data$Abd_length)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0463  0.0948  0.1046  0.1045  0.1145  0.1760

# run various summary stats column length
length(mosq_data$Abd_length)
## [1] 3643
# minimum value in column
min(mosq_data$Abd_length)
## [1] 0.0463
# maximum value in column
max(mosq_data$Abd_length)
## [1] 0.176
# range of values in column, min-max
range(mosq_data$Abd_length)
## [1] 0.0463 0.1760
# mean of column, removes NAs
mean(mosq_data$Abd_length)
## [1] 0.1044646
# standard deviation of column
```

Indexing and subsetting

Sometimes we only want to view or analyze a specific portion of our dataset. For example, with the dataset we're working with today, we're only interested in data on female mosquitoes because male mosquitoes are not involved in disease transmission.

There are two ways to filter out all of the male mosquito data:

1. Indexing

```
#mydataset[row(s) that meet this criteria, columns that meet this criteria]
mosq_females_index<-mosq_data[mosq_data$Sex=="F",c(1:4)]
#inspect/summarize
head(mosq_females_index)
##    Species Sex Abd_length Dev_time
## 2       AL   F     0.0491       14
## 6       AL   F     0.0550        9
## 7       AL   F     0.0557       10
## 8       AL   F     0.0564        8
## 9       AL   F     0.0568       14
## 13      AL   F     0.0584       12
mean(mosq_females_index$Abd_length)
## [1] 0.1065329
```

Indexing and subsetting

Sometimes we only want to view or analyze a specific portion of our dataset. For example, with the dataset we're working with today, we're only interested in data on female mosquitoes because male mosquitoes are not involved in disease transmission.

There are two ways to filter out all of the male mosquito data:

2. Subsetting

```
#subset(mydataset,row(s) that meet this criteria, select=c(column names or numbe
mosq_females_subset<-subset(mosq_data,Sex=="F",select=c(1:4))
#inspect and summarize
head(mosq_females_subset)
##    Species Sex Abd_length Dev_time
## 2       AL   F     0.0491       14
## 6       AL   F     0.0550        9
## 7       AL   F     0.0557       10
## 8       AL   F     0.0564        8
## 9       AL   F     0.0568       14
## 13      AL   F     0.0584       12
mean(mosq_females_subset$Abd_length)
## [1] 0.1065329
```

Indexing and subsetting

Here are a few more examples of ways we can index or subset our data

1. Pull out *Culex quinquefasciatus* data

```
culex<-mosq_data[mosq_data$Species=="CQ",c(1:4)]
#or
culex<-subset(mosq_data,Species=="CQ",select=c(1:4))
```

2. Pull out *Culex quinquefasciatus* **female** data

```
culex_females<-mosq_data[mosq_data$Species=="CQ"&mosq_data$Sex=="F",1:4]
#or
culex_females<-subset(mosq_data,Species=="CQ"&Sex=="F",select=c(1:4))
```

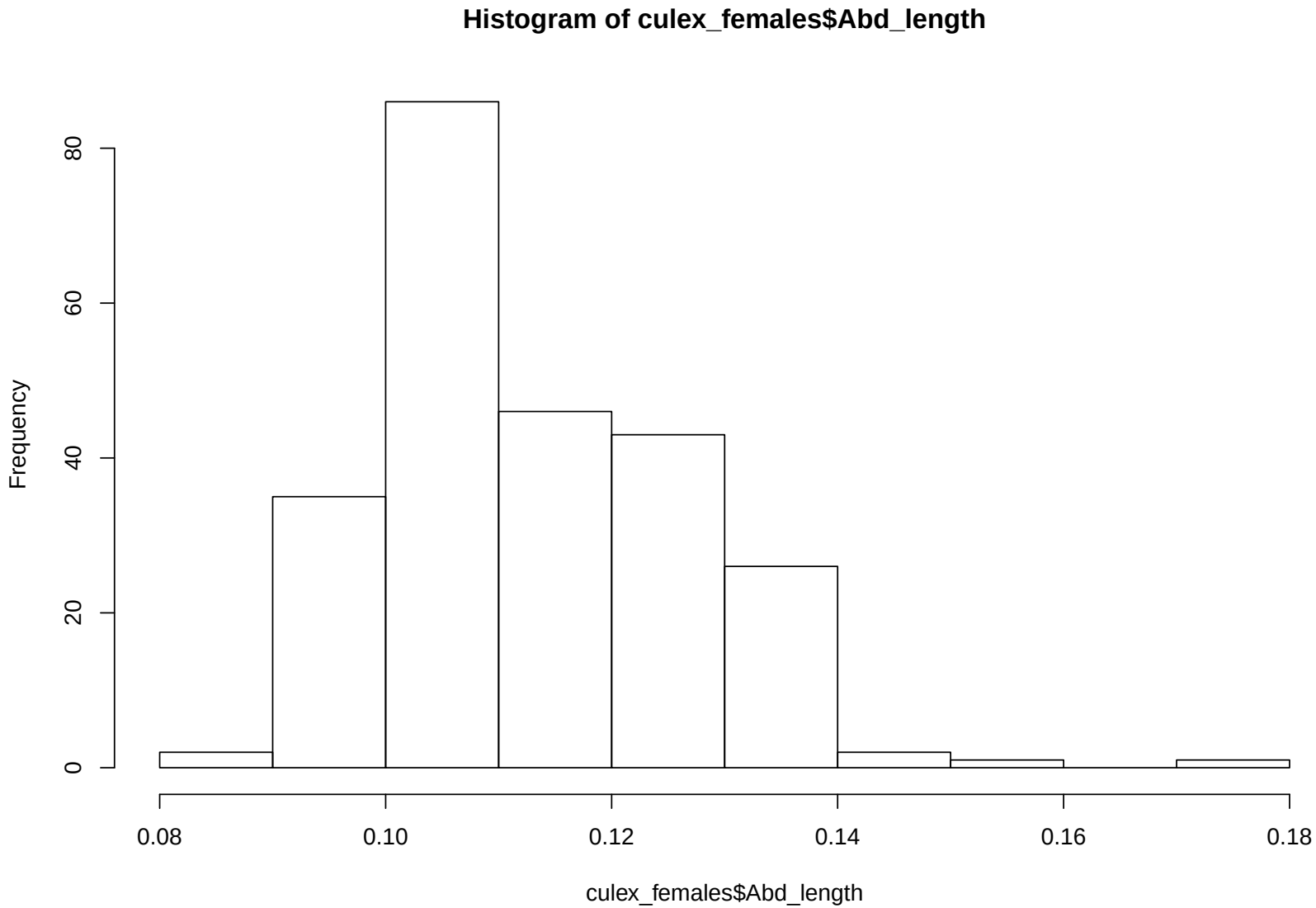
3. Pull out all individuals that took at least 10 days to develop

```
devtime_10days<-mosq_data[mosq_data$Dev_time>=10,c(1:4)]
#or
devtime_10days<-subset(mosq_data,Dev_time>=10,select=c(1:4))
```

A preview of data visualization: histograms

Use `hist()` to look at a histogram of your data

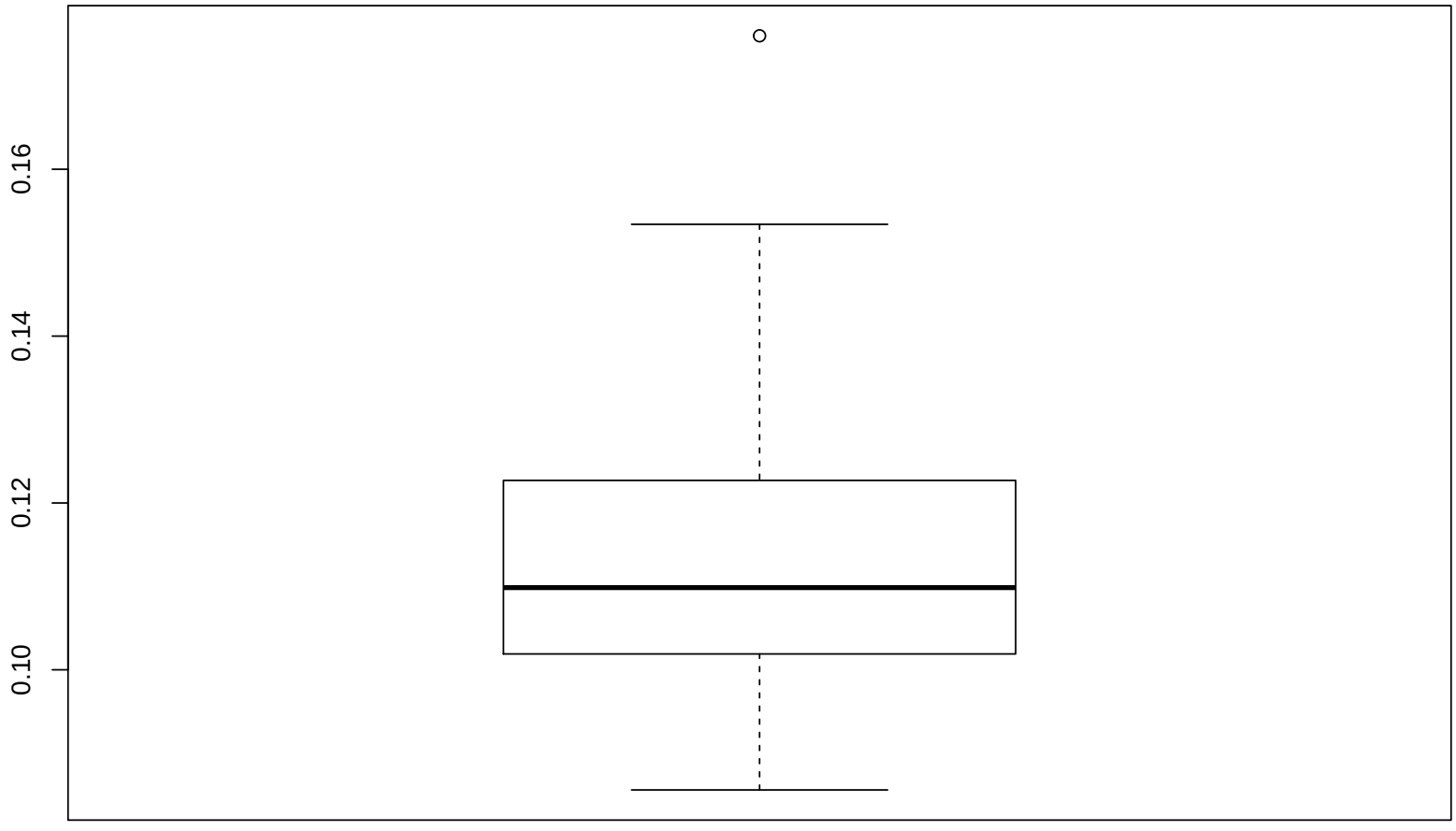
```
hist(culex_females$Abd_length)
```



A preview of data visualization: box-and-whisker plots

Use `boxplot()` to draw a box-and-whisker plot of your data

```
boxplot(culex_females$Abd_length)
```



Practice exercises

1. Working from the original data file, **subset the data by sex**
2. Report the number of females and males **for each species**
3. Report the mean and standard deviation of body length **for each sex**
4. Create a histogram and a bocplot of body length **for each sex**
5. Which sex has more **body length outliers**?
6. Do female body size outliers tend to be **much longer or much shorter** than average?
7. Do male body size outliers tend to be **much longer or much shorter** than average?

Hint: for last three questions, type `?boxplot` *into the console to pull up the help page in the bottom right pane*

Practice exercise: answers

1. Working from the original data file, **subset the data by sex**

```
males<-mosq_data[mosq_data$Sex=="M",c(1:4)]
females<-mosq_data[mosq_data$Sex=="F",c(1:4)]

#or

males<-subset(mosq_data,Sex=="M",select=c(1:4))
females<-subset(mosq_data,Sex=="F",select=c(1:4))
```

Practice exercise: answers

2. Report the number of females and males **for each species**

```
dim(males[males$Species=="CQ",c(1:4)])
## [1] 376   4
dim(males[males$Species=="AL",c(1:4)])
## [1] 932   4
dim(males[males$Species=="AG",c(1:4)])
## [1] 780   4
```

```
dim(females[females$Species=="CQ",c(1:4)])
## [1] 242   4
dim(females[females$Species=="AL",c(1:4)])
## [1] 600   4
dim(females[females$Species=="AG",c(1:4)])
## [1] 713   4
```

Practice exercise: answers

3. Report the mean and standard deviation of body length **for each sex**

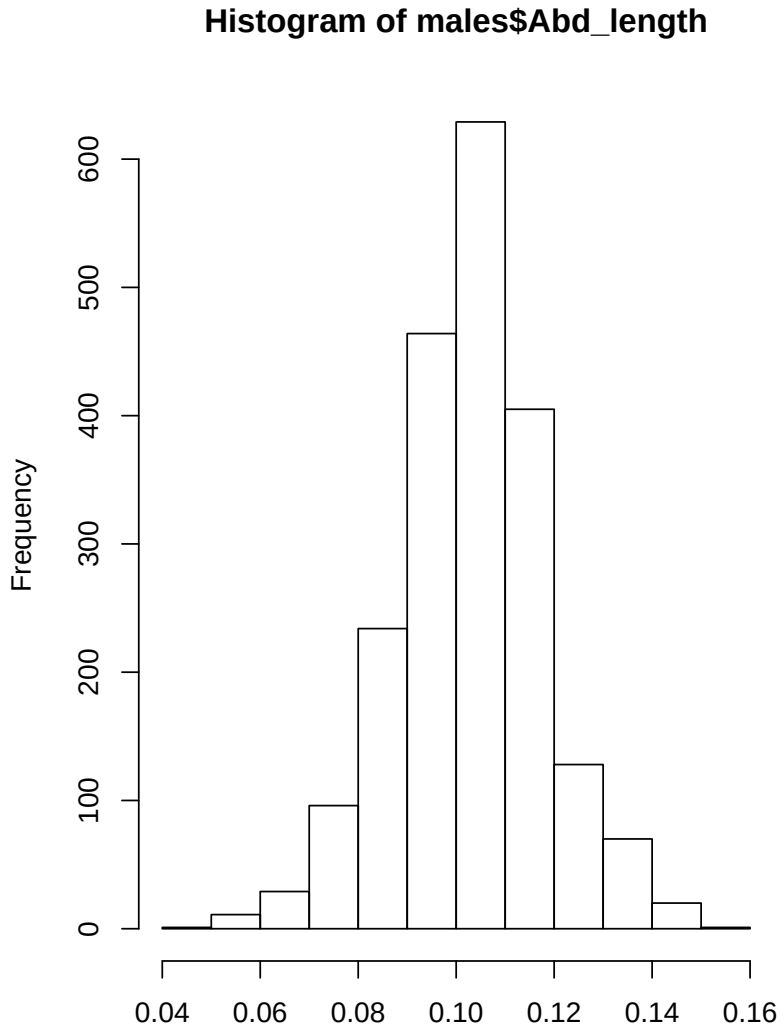
```
mean(males$Abd_length)
## [1] 0.1029243
sd(males$Abd_length)
## [1] 0.01496534
```

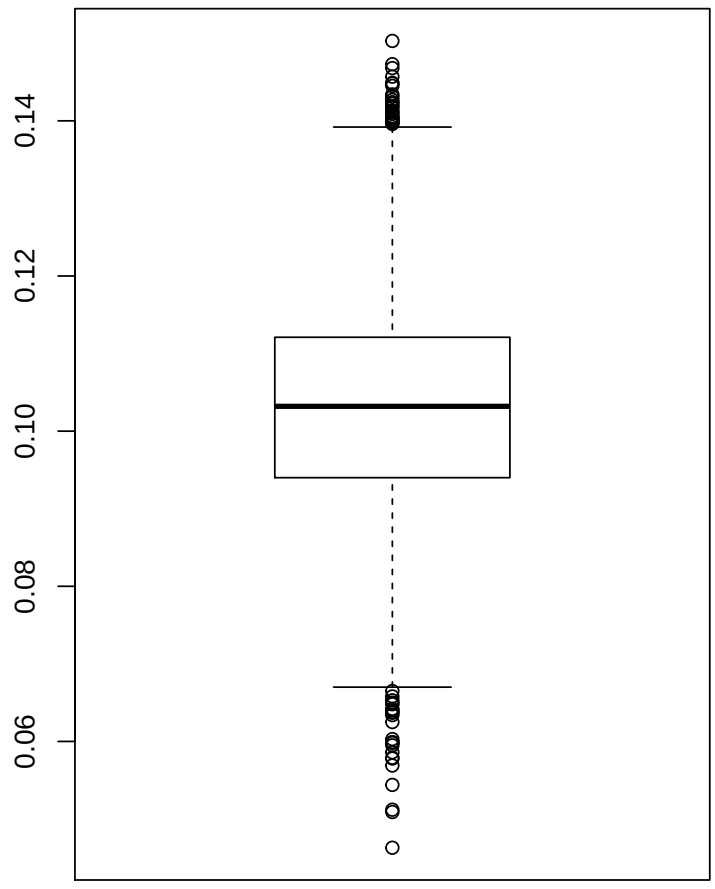
```
mean(females$Abd_length)
## [1] 0.1065329
sd(females$Abd_length)
## [1] 0.01632038
```

Practice exercise: answers

4. Create a histogram and a boxplot of body length **for each sex**

```
par(mfrow=c(1,2))
hist(males$Abd_length)
boxplot(males$Abd_length)
```





Practice exercise: answers

By perusing the help page for `boxplot`, we can see that there is a function `boxplot.stats` that does the computations necessary to make the boxplot. If we go to the help page for `boxplot.stats`, we can see that the output of this function contains a list of the outliers (`$out`), which allows us to answer the remaining questions.

5. Which sex has more **body length outliers**? **Males**

```
boxplot.stats(males$Abd_length)$out
##  [1] 0.0463 0.0509 0.0512 0.0544 0.0569 0.0578 0.0579 0.0586 0.0595 0.0598
## [11] 0.0599 0.0599 0.0603 0.0625 0.0634 0.0637 0.0638 0.0640 0.0642 0.0648
## [21] 0.0650 0.0653 0.0658 0.0665 0.1396 0.1397 0.1399 0.1399 0.1400 0.1402
## [31] 0.1404 0.1404 0.1406 0.1409 0.1411 0.1413 0.1417 0.1419 0.1422 0.1424
## [41] 0.1427 0.1431 0.1434 0.1445 0.1448 0.1449 0.1457 0.1468 0.1473 0.1503

boxplot.stats(females$Abd_length)$out
##  [1] 0.0491 0.0550 0.0557 0.0564 0.0568 0.0584 0.0594 0.0601 0.0625 0.0632
## [11] 0.0634 0.0636 0.1534 0.1760
```

6. Do female body size outliers tend to be **much longer or much shorter** than average?

Shorter

7. Do male body size outliers tend to be **much longer or much shorter** than average?

Approximately the same on either end