Practice Solutions to Getting Started with R and RStudio

Jessica Minnier, PhD & Meike Niederhausen, PhD

OCTRI Biostatistics, Epidemiology, Research & Design (BERD) Workshop

2019/09/24

Practice questions 1

1. Open a new R script and type code/answers for next tasks in it. Save as `Practice1.R`
2. Create a vector of all integers from 4 to 10, and save it as `a1`.
3. Create a vector of *even* integers from 4 to 10, and save it as `a2`.
4. What is the sum of `a1` and `a2`?
5. What does the command `sum(a1)` do?
6. What does the command `length(a1)` do?
7. Use the commands to calculate the average of the values in `a1`.
8. The formula for the first n integers is $n(n+1)/2$. Compute the sum of all integers from 1 to 100 to verify that this formula holds for $n = 100$.
9. Compute the sum of the squares of all integers from 1 to 100.
10. Take a break!

Answers to practice questions 1

#2 Create a vector of all integers from 4 to 10, and save it as `a1`.

#3 Create a vector of *even* integers from 4 to 10, and save it as `a2`.

```
a1 <- 4:10
a2 <- c(4, 6, 8, 10)
# the following works as well:
a2 <- 2*(2:5)
# or
a2 <- seq(4, 10, by=2)
```

#4 What is the sum of `a1` and `a2`?

```
a1+a2
```

```
Warning in a1 + a2: longer object length is not a multiple of shorter
object length
```

```
[1]  8 11 14 17 12 15 18
```

Note that instead of giving an error, the terms of `a1` are repeated as needed since `a2` is longer than `a1`

#5 What does the command `sum(a1)` do?

```
sum(a1)
```

```
[1] 49
```

`sum` adds up the values in the vector

#6 What does the command `length(a1)` do?

```
length(a1)
```

```
[1] 7
```

`length` is the number of values in the vector

#7 Use the commands to calculate the average of the values in `a1`.

```
sum(a1) / length(a1)
```

```
[1] 7
```

```
# this is equivalent
mean(a1)
```

```
[1] 7
```

#8 The formula for the first n integers is $n(n+1)/2$. Compute the sum of all integers from 1 to 100 to verify that this formula holds for $n = 100$.

```
sum(1:100)
```

```
[1] 5050
```

```
# verify formula for n=100:
n=100
n * (n+1) / 2
```

```
[1] 5050
```

#9 Compute the sum of the squares of all integers from 1 to 100.

```
# The following code creates a vector of the squares of all integers from 1 to 100
(1:100)^2
```

```
  [1]     1     4     9    16    25    36    49    64    81   100   121
 [12]   144   169   196   225   256   289   324   361   400   441   484
 [23]   529   576   625   676   729   784   841   900   961  1024  1089
 [34]  1156  1225  1296  1369  1444  1521  1600  1681  1764  1849  1936
 [45]  2025  2116  2209  2304  2401  2500  2601  2704  2809  2916  3025
 [56]  3136  3249  3364  3481  3600  3721  3844  3969  4096  4225  4356
 [67]  4489  4624  4761  4900  5041  5184  5329  5476  5625  5776  5929
 [78]  6084  6241  6400  6561  6724  6889  7056  7225  7396  7569  7744
 [89]  7921  8100  8281  8464  8649  8836  9025  9216  9409  9604  9801
[100] 10000
```

```
# Now add the squares:
sum((1:100)^2)
```

```
[1] 338350
```

Practice 2

1. Create a new script and save it as `Practice2.R`
2. Create data frames for males and females separately.
3. Do males and females have similar BMIs? Weights? Compares means, standard deviations, range, and boxplots.
4. Plot BMI vs. weight for each gender separately. Do they have similar relationships?
5. Are males or females more likely to be bullied in the past 12 months? Calculate the percentage bullied for each gender.

Practice 2 Answers

#2 Create data frames for males and females separately.

```
boys <- mydata[mydata$sex == "Male", ]
dim(boys)
```

```
[1]  8 11
```

```
girls <- mydata[mydata$sex == "Female", ]
dim(girls)
```

```
[1] 12 11
```

Check number of boys & girls:

```
summary(mydata$sex)
```

```
Female   Male
    12      8
```

#3 Do males and females have similar BMIs? Weights? Compares means, standard deviations, range, and boxplots.

```
summary(boys$bmi); sd(boys$bmi)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.18   19.57   20.90   20.63   21.58   22.46 
```

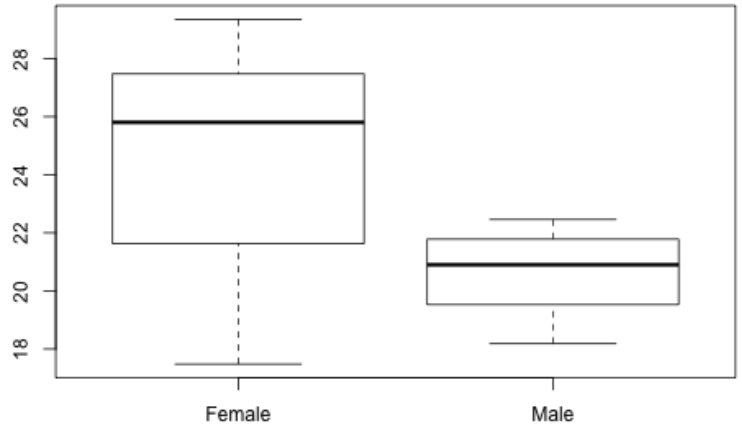
```
[1] 1.466896
```

```
summary(girls$bmi); sd(girls$bmi)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  17.48   21.95   25.80   24.59   27.47   29.35 
```

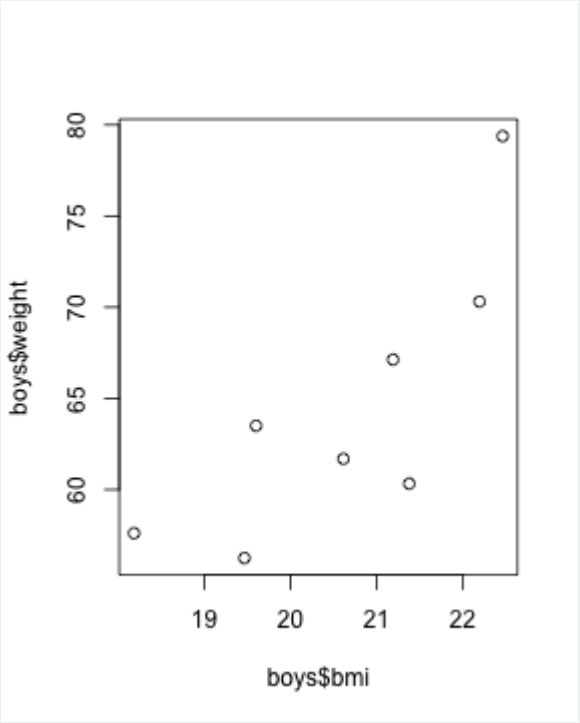
```
[1] 3.70739
```

```
boxplot(mydata$bmi ~ mydata$sex)
```

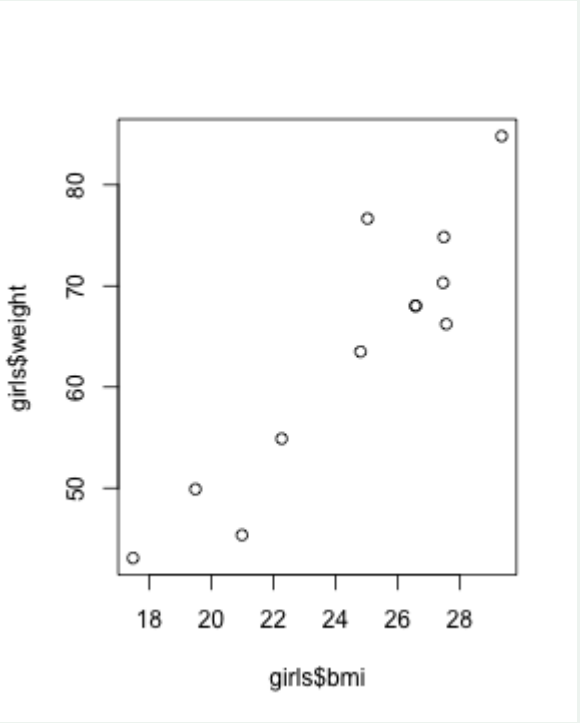


#4 Plot BMI vs. weight for each gender separately. Do they have similar relationships?

```
plot(boys$bmi, boys$weight)
```




```
plot(girls$bmi, girls$weight)
```



#5 Are males or females more likely to be bullied in the past 12 months? Calculate the percentage bullied for each gender.

```
bullied_boys <-
  boys[boys$bullied_past_12mo == TRUE,]
nrow(bullied_boys)
```

```
[1] 3
```

```
bullied_boys_prct <-
  nrow(bullied_boys) / nrow(boys) * 100
bullied_boys_prct
```

```
[1] 37.5
```

```
# alternative
mean(boys$bullied_past_12mo, na.rm=TRUE)
```

```
[1] 0.375
```

```
# Apply the same method for girls:
bullied_girls <-
  girls[girls$bullied_past_12mo == TRUE,]
nrow(bullied_girls)
```

```
[1] 6
```

```
bullied_girls_prct <-
  nrow(bullied_girls) / nrow(girls) * 100
bullied_girls_prct
```

```
[1] 50
```

```
# alternative. Answers don't match. Why???
mean(girls$bullied_past_12mo, na.rm=TRUE)
```

```
[1] 0.4
```

#5 cont'd

On the previous slide we saw that our two methods for calculating the percentage of girls that were bullied in the past 12 months did not match. What went wrong?

```
nrow(bullied_girls)
```

```
[1] 6
```

```
girls$bullied_past_12mo
```

```
 [1]    NA    NA  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
[12] FALSE
```

To get the number of girls that were bullied we need to make sure the missing values (NA) are not included.

#5 cont'd - working with NA's

```
# values of bullied_past_12mo
girls$bullied_past_12mo
```

```
 [1]    NA    NA  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
[12] FALSE
```

```
# which are missing (logical)
is.na(girls$bullied_past_12mo)
```

```
 [1]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE
```

```
# which are NOT missing (logical)
!is.na(girls$bullied_past_12mo)
```

```
 [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[12]  TRUE
```

#5 cont'd - fix girls' code

Exclude the missing values from the `bullied_girls`:

```
girls2 <- girls[!is.na(girls$bullied_past_12mo),]
nrow(girls2)
```

```
[1] 10
```

```
bullied_girls2 <- girls2[girls2$bullied_past_12mo == TRUE,]
nrow(bullied_girls2)
```

```
[1] 4
```

```
# from girls dataset, total number bullied
sum(girls$bullied_past_12mo, na.rm = TRUE)
```

```
[1] 4
```

#5 cont'd - Calculate percentage girls bullied

```
bullied_girls_prct2 <- nrow(bullied_girls2) / nrow(girls2) * 100
bullied_girls_prct2
```

```
[1] 40
```

```
# Compare to alternative
mean(girls$bullied_past_12mo, na.rm=TRUE)
```

```
[1] 0.4
```