Homework Exercises Week 02

Alison Lawyer

2024-09-24

keep this chunk in all your RMarkdown scripts

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

HOMEWORK EXERCISES

For this week's homework exercises, you will continue working with the palmer penguins dataset. You will practice tidyverse functions that you have already learned about and conduct some descriptive analyses.

Part 1

Tidyverse practice

```
# Load the palmerpenguins package. You will also need the
# package dplyr from the tidyverse for the next exercises
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(palmerpenguins)

# Assign the 'penguins_raw' dataframe contained in the
# package to a new object Hint: You can just do [your
# object name here] <- penguins_raw once you load the
# package
penguins <- penguins_raw
```

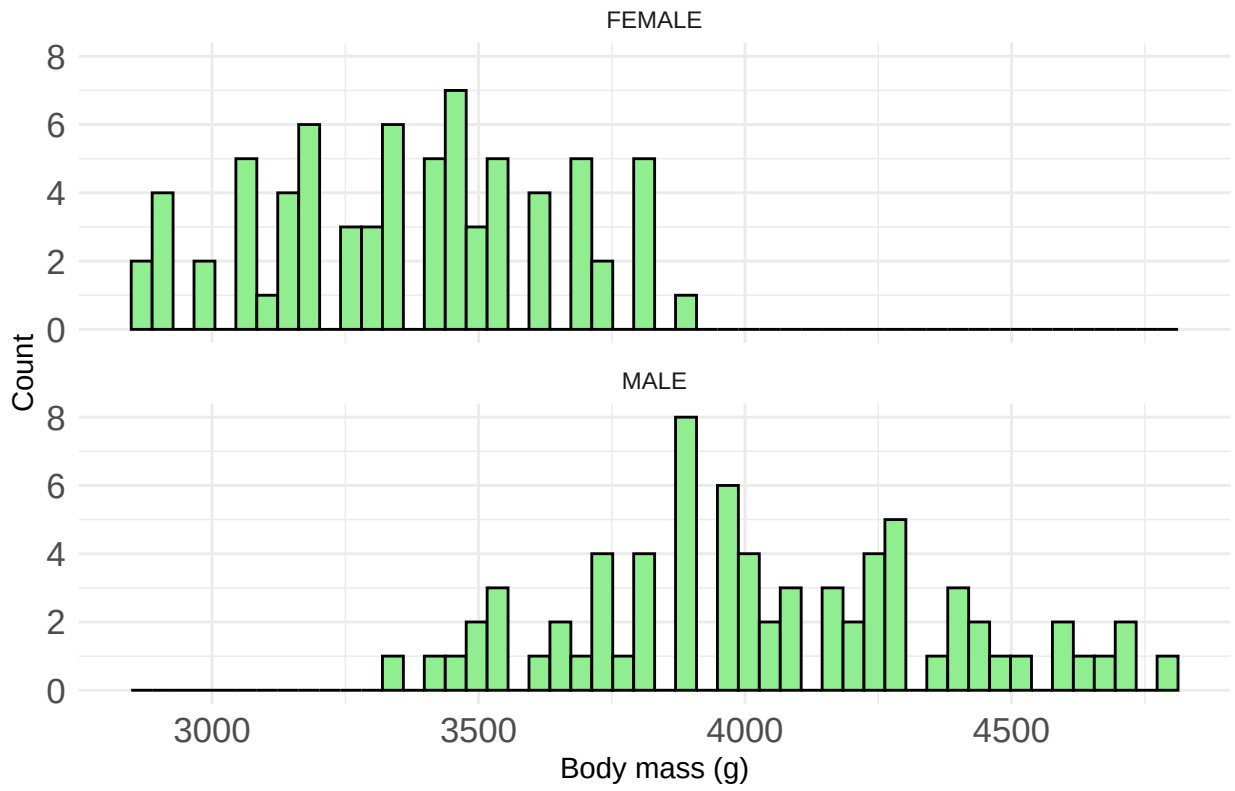
```r
### Simplify variable names.  Rename the columns
### Culmen.Length..mm. to bill_length, Culmen.Depth..mm. to
### bill_depth, Flipper.Length..mm. to flipper_length, and
### Body.Mass..g. to body_mass Use the function rename()
### from the dplyr package Save the result to a new
### dataframe object for later use
penguins_renamed <- penguins %>%
    rename(bill_length = `Culmen Length (mm)`, bill_depth = `Culmen Depth (mm)`,
        flipper_length = `Flipper Length (mm)`, body_mass = `Body Mass (g)`)


# Change all remaining column names to user lower case
# only. This is helpful to avoid confusion about
# capitalization in your code and speeds up writing code as
# well! To do this, explore the function toupper() and
# tolower(). You can rename each column manually as above
# or you can extract a vector of column names with names(),
# apply the tolower() function to that vector, and reassign
# the dataframe column names to that new vector of names.
# Try it out!

## one option: penguins_lower_columns <-
## tolower(names(penguins_renamed))
penguins_lower_columns <- penguins_renamed %>%
    names() %>%
    tolower()
penguins_renamed <- penguins_renamed %>%
    rename_with(tolower)


# Create yet another dataframe that contains a subset of
# your columns Select the columns: species, island, sex,
# bill_length, bill_depth, body_mass Note: It is easier to
# work with dataframes that have only the information you
# might need for the analyses and visualizations you want
# to do. You can always add others later if need be. When
# you do this, always work on a copy of your data!
penguins_selected <- penguins_renamed %>%
    select(c(species, island, sex, bill_length, bill_depth, body_mass))


# Filter your data to Adelie penguins using the filter()
# function from dplyr Keep all columns from your cleaned
# subset of the data created in the step before
adelie_penguins <- penguins_selected %>%
    filter(species == "Adelie Penguin (Pygoscelis adeliae)")


# Print out a descriptive summary of the Adelie penguin
# data colums with a function of your choice.  This does
# not have to be pretty!
summary(adelie_penguins)
```

```
##    species             island              sex             bill_length
##  Length:152         Length:152         Length:152         Min.   :32.10
##  Class :character   Class :character   Class :character   1st Qu.:36.75
##  Mode  :character   Mode  :character   Mode  :character   Median :38.80
##                                                           Mean   :38.79
##                                                           3rd Qu.:40.75
##                                                           Max.   :46.00
##                                                           NA's   :1
##    bill_depth      body_mass
##  Min.   :15.50   Min.   :2850
##  1st Qu.:17.50   1st Qu.:3350
##  Median :18.40   Median :3700
##  Mean   :18.35   Mean   :3701
##  3rd Qu.:19.00   3rd Qu.:4000
##  Max.   :21.50   Max.   :4775
##  NA's   :1       NA's   :1
```

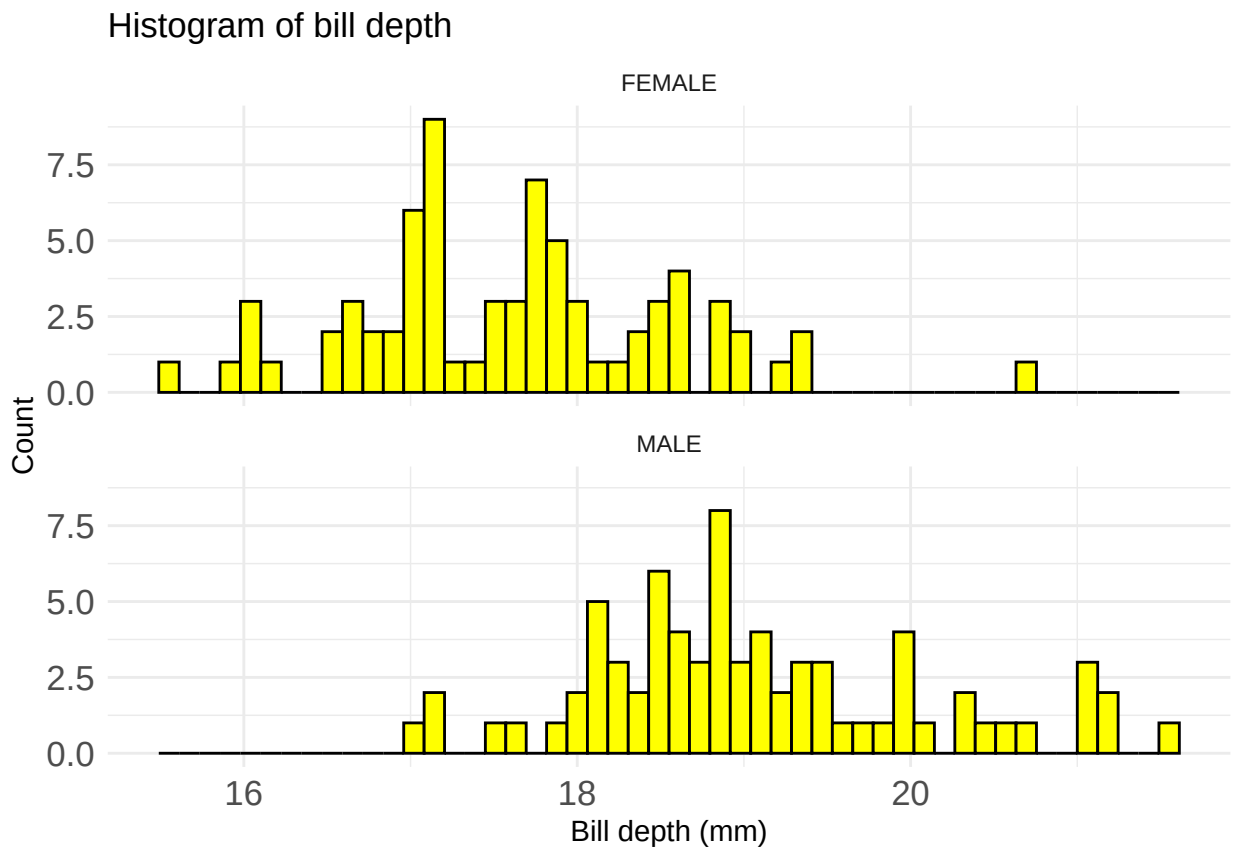
Part 2

In this part you will create histograms to assess the appropriate descriptive statistics to use on the Adelie penguin data you extracted in Part 1. Your goal is to describe the central tendency and variability in penguin bill morphology and weight, separately for males and females, ignoring observations with unknown sex.

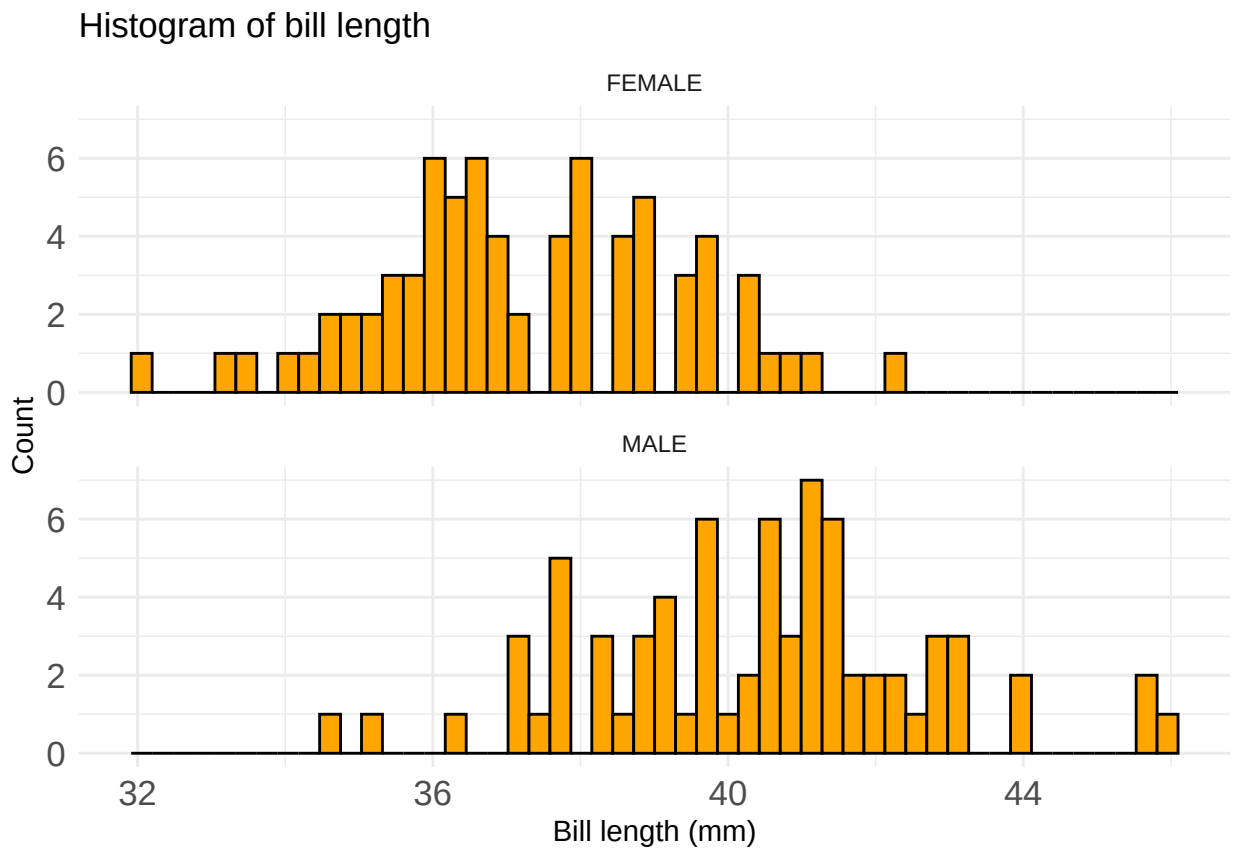
```
# Determine how to calculate appropriate descriptive
# statistics for your data.  For this, first create a
# histogram of the distribution of all your continuous
# variables For each variable, create a histogram for males
# and females separately.  You can use the hist() function
# or explore the use of ggplot(). The latter will allow you
# to split the output by sex without having to create
# separate subsets of the data by using the facet_wrap() or
# facet_grid() functions within ggplot. See lecture code
# for examples.
adelie_penguins_no_na <- adelie_penguins %>%
    na.omit()
adelie_penguins_no_na %>%
    ggplot(aes(x = body_mass)) + geom_histogram(color = "black",
    fill = "lightgreen", bins = 50, na.rm = TRUE) + theme_minimal() +
    labs(title = "Histogram of body mass", x = "Body mass (g)",
        y = "Count") + theme(axis.text = element_text(size = 13)) +
    facet_wrap(~sex, nrow = 3)
```




```
adelie_penguins_no_na %>%
    ggplot(aes(x = bill_depth)) + geom_histogram(color = "black",
    fill = "yellow", bins = 50, na.rm = TRUE) + theme_minimal() +
    labs(title = "Histogram of bill depth", x = "Bill depth (mm)",
        y = "Count") + theme(axis.text = element_text(size = 13)) +
    facet_wrap(~sex, nrow = 3)
```




```
adelie_penguins_no_na %>%
    ggplot(aes(x = bill_length)) + geom_histogram(color = "black",
    fill = "orange", bins = 50, na.rm = TRUE) + theme_minimal() +
    labs(title = "Histogram of bill length", x = "Bill length (mm)",
        y = "Count") + theme(axis.text = element_text(size = 13)) +
    facet_wrap(~sex, nrow = 3)
```




```
# Answer the question: What is the appropriate measure of
# central tendency and spread to use for these variables
# and why?

## Using mean should be a sufficient measure of central
## tendency for all values, since there isn't a dramatic
## visible skew in either direction on the graphs. If we
## use mean, then we should use standard deviation to
## measure spread.
```

Part 3

Now that you have decided on the appropriate descriptive statistics for your Adelie penguins, you will calculate them and summarize the results in a new dataframe.

```
# Calculate the mean and standard deviation for each of the
# three continuous variables and arrange them in a summary
# dataframe that contains species, sex, means and standard
# deviations for each measure (total of 8 columns).
adelie_penguins_no_na %>%
    group_by(sex, species) %>%
    summarize(mean_bill_depth = mean(bill_depth), sd_bill_depth = sd(bill_depth),
        mean_bill_length = mean(bill_length), sd_bill_length = sd(bill_length),
        mean_body_mass = mean(body_mass), sd_body_mass = sd(body_mass))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

```
## # A tibble: 2 x 8
## # Groups:   sex [2]
##   sex    species   mean_bill_depth sd_bill_depth mean_bill_length sd_bill_length
##   <chr>  <chr>               <dbl>         <dbl>            <dbl>          <dbl>
## 1 FEMALE Adelie P~            17.6         0.943             37.3           2.03
## 2 MALE   Adelie P~            19.1         1.02              40.4           2.28
## # i 2 more variables: mean_body_mass <dbl>, sd_body_mass <dbl>
```

```
# Answer the question: What is more variable, bill length
# or bill depth?

## Bill length has a larger SD value for both sexes than
## the SD value for bill depth, so bill length is more
## variable.
```