BIOL 220 Problem Set 05: Review coding, data, graphics

Answer Key

Speed round (jk take your time)

Plotting

1. Fix the code below, you're not trying to make the plot look nicer, just trying to correct errors (hint: try running it in R to see if you fixed it all the way; what type of plot should we use for these kinds of data?) [1 point]

```
library(palmerpenguins)
# library(ggplot2)

ggplot(penguins,
       aes(x = bill_depth_mm, y = bill_length_mm,
           color = "species")) +
    geom_boxplot() +
    scale_color_viridis_c()
```

Answer

```
library(palmerpenguins)
library(ggplot2)

ggplot(penguins,
       aes(x = bill_depth_mm, y = bill_length_mm,
           color = species)) +
    geom_point() +
    scale_color_viridis_d()
```

2. Match the plot to the data type (e.g. "a. goes to iii."—that is not correct, just an example!) [1 point]

a. *Response:* categorical, *Explanatory:* none
b. *Response:* numerical, *Explanatory:* categorical
c. *Response:* numerical, *Explanatory:* numerical
d. *Response:* numerical, *Explanatory:* none

i. Scatter plot
ii. Bar plot
iii. Box plot
iv. Histogram

Answer

a. *Response:* categorical, *Explanatory:* none	ii. Bar plot
b. *Response:* numerical, *Explanatory:* categorical	iii. Box plot
c. *Response:* numerical, *Explanatory:* numerical	i. Scatter plot
d. *Response:* numerical, *Explanatory:* none	iv. Histogram

Estimates

3. Fix this code to calculate the standard error (again, don't add to the code, just correct it) [2 points]

```r
gentoo <- subset(penguins, penguins$species == "Gentoo")

y <- gentoo$flipper_length_mm
y <- y[!is.na(y)]

n <- nrow(y)

ybar <- mean(y)

s2 <- 1 / n * sum((y - ybar))
s <- sqrt(s2)

se <- 1 / sqrt(n) * s
se

# compare to the below calculation (this code is correct)
sd(y, na.rm = TRUE) / sqrt(n)
```

Answer

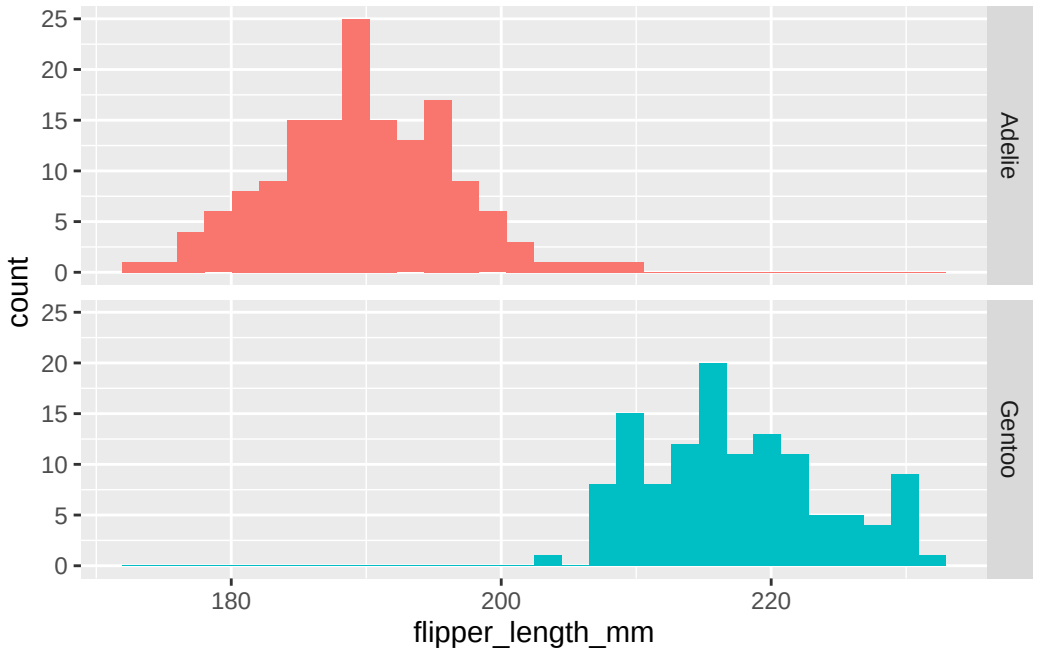
```r
gentoo <- subset(penguins, penguins$species == "Gentoo")

y <- gentoo$flipper_length_mm
y <- y[!is.na(y)]

n <- length(y)

ybar <- mean(y)

s2 <- 1 / (n - 1) * sum((y - ybar)^2)
s <- sqrt(s2)

se <- 1 / sqrt(n) * s
se

# compare to the below calculation (this code is correct)
sd(y, na.rm = TRUE) / sqrt(n)
```

Putting it all together

Let's keep using the `penguins` data. Here we'll compare the data from Gentoo and Adelie penguin flippers.

To start, I'll show you a faceted histogram (of an *exploratory* quality) of flipper lengths across these two species:

```
ggplot(subset(penguins, penguins$species != "Chinstrap"),
              aes(x = flipper_length_mm, fill = species)) +
    geom_histogram() +
    facet_grid(rows = vars(species)) +
    theme(legend.position = "none")
```



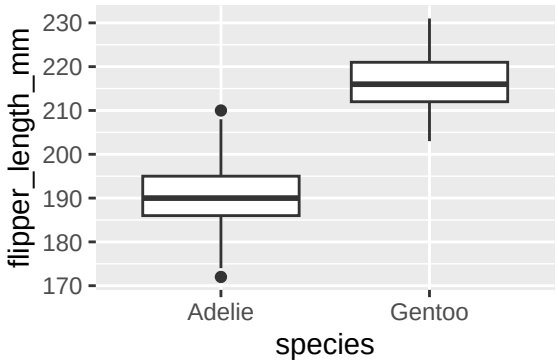
Use the `penguins` data (and refer to the above graph) to answer questions 4–6

4. Make a boxplot of the Adelie and Gentoo penguins
    a. use *ggplot* code to make the plot and upload the image to Google Forms (e.g. take a screenshot and upload that) [1 point]
    b. what does the boxplot tell you that the histrogram does not; what does the histogram tell you that the boxplot does not? [1 point]

Answer

4a.

```
ggplot(subset(penguins, penguins$species != "Chinstrap"),
              aes(y = flipper_length_mm, x = species)) +
    geom_boxplot()
```



4b. The boxplot tells us that Adelie has two outliers. The boxplot more easily tells us that the interquartile region for both species *looks* about the same width. The histogram more clearly shows us that the distribution for Adelie is more smooth, perhaps because there are more data points, and also that the overlap between the two groups is mostly cause by the extreme larger values for Adelie overlapping with more average values for Gentoo.

5. Now we will look numerically at the spread of these the data on these two species
    a. calculate the mean and standard deviation of flipper length of both species, report your answer with 1 decimal place [1 point]
    b. calculate the standard errors of the means for both species, again, 1 decimal place [1 point]

Answers

5a. and 5b.

```
library(dplyr)

dat <- subset(penguins, penguins$species != "Chinstrap" &
                  !is.na(penguins$flipper_length_mm))

dat <- group_by(dat, species)

estimates <- summarize(dat, mean = mean(flipper_length_mm),
                       sd = sd(flipper_length_mm),
                       se = sd(flipper_length_mm) / sqrt(n()))
estimates
```

```
# A tibble: 2 x 4
  species  mean    sd    se
  <fct>   <dbl> <dbl> <dbl>
1 Adelie   190.  6.54 0.532
2 Gentoo   217.  6.48 0.585
```

6. Explain the standard deviations and standard errors

   a. which species has the bigger standard deviation, which has the bigger standard error? Based on the graphs of the data, did you expect this answer or were you surprised? Why or why not? [1 point]
   b. Compare the standard **error** of the Gentoo versus the Adelie, what is more likely to explain why one is larger than the other: the sample sizes for the two different species, or the standard deviations for the two different species? Explain why you picked your answer [1 point]

Answers

6a. Adelie has the bigger SD while Gentoo has the bigger SE. Given that graphs, I expected them to have approximately the same SD, with Adelie possibly having a bigger SD because of the outliers. The fact that Gentoo has a larger SE is also not really surprising given their SDs are close in value and Gentoo has a much smaller sample size

6b. The sample size is responsible for Gentoo having a larger SE than Adelie because the SD of Adelie is in fact bigger than the SD for Gentoo. That means it can only be the sample size that causes the SE of Gentoo to be bigger