Review of Statistical Estimation

STA4173: Biostatistics

Spring 2025

Introduction

- In this lecture, we will review estimation
  - Continuous variables
    - Mean
    - Median
    - Percentiles / quartiles
    - Variance and standard deviation
    - Interquartile range
  - Categorical variables
    - Count
    - Percentage
- We will also discuss exploring data graphically

R: Introduction

- In this course, we will review formulas, but we will use R for computational purposes
  - → Remember to refer to the lecture notes for specific code needed
  - → Code is also available on this course's GitHub repository
- You can install R and RStudio if you wish; both are free.
  - → We have access to the Posit Workbench ("the server") through HMCSE.
- I know that this is probably the first time you are seeing R (or any sort of programming).
  - → That is why we have "R lab" time built in to our course.
  - → Remember that I am not looking for perfection, but instead for competency.

Today's Data: Palmer Penguins

- Today we will be demonstrating the basics using the Palmer Penguins dataset, available through R.

```
1 penguins <- palmerpenguins::penguins
```

species <fct>	island <fct>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>	body_mass_g <int>	sex <fct>	year <int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

1-10 of 344 rows

Previous 1 2 3 4 5 6 ... 35 Next

Types of Variables

Continuous Variables

A continuous variable is a variable that can has an infinite set of possible values.

- Between any two possible values, there are an infinite number of possible values.
- These typically arise from measurement. (Height, weight, etc.)

Discrete Variables

A discrete variable is a variable that can only take on a finite set of possible values.

- The possible values can usually be listed.
- These typically arise from categorizing (work vs. home) or counting.

Types of Variables

species <fct>	island <fct>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>	body_mass_g <int>	sex <fct>	year <int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

1-10 of 344 rows

Previous 1 2 3 4 5 6 ... 35 Next

Types of Continuous Variables

Ratio Variables

A ratio variable is a variable that has a meaningful zero point, allowing comparisons of magnitude.

- True zero point indicates the absence of the quantity being measured.
- All arithmetic operations (addition, subtraction, multiplication, division) are meaningful.

Interval Variables

An interval variable has an arbitrary zero point and differences between values are meaningful.

- The zero point does not indicate a true absence.
- A 1 unit difference always represents the same amount.

Types of Discrete Variables

Ordinal Variables

An ordinal variable has a meaningful order of responses; the exact differences between responses are not necessarily equal.

- We understand which value is “greater” or “less,” but not by how much.
- Arithmetic is not meaningful.

Nominal Variables

A nominal variable has is no intrinsic order among the categories.

- Categories are used merely as labels or names.
- No arithmetic or ordering operations are meaningful.

Measures of Centrality: Mean

Sample Mean

The sample mean provides a single number that can represent a “typical” or central value in your data.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- R syntax:

```r
dataset_name %>% summarize(mean(variable_name, na.rm = TRUE))
```

Measures of Centrality: Mean

- Let's find the average weight (*body_mass_g*) of the penguins.

```
1 penguins %>% summarize(mean(body_mass_g, na.rm = TRUE))
```

mean(body_mass_g, na.rm = TRUE) <dbl>
4201.754

1 row

- Let's find the average flipper length (*flipper_length_mm*) of the penguins.

```
1 penguins %>% summarize(mean(flipper_length_mm, na.rm = TRUE))
```

mean(flipper_length_mm, na.rm = TRUE) <dbl>
200.9152

1 row

Measures of Centrality: Median

Sample Median

The sample median is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger.

- If n is odd, the median is the single middle value.
- If n is even, the median is the average of the two middle values.

- R syntax:

```
dataset_name %>% summarize(median(variable_name, na.rm = TRUE))
```

Measures of Centrality: Median

- Let's find the median weight (*body_mass_g*) of the penguins.

```
1 penguins %>% summarize(median(body_mass_g, na.rm = TRUE))
```

median(body_mass_g, na.rm = TRUE) <dbl>
4050

1 row

- Let's find the median flipper length (*flipper_length_mm*) of the penguins.

```
1 penguins %>% summarize(median(flipper_length_mm, na.rm = TRUE))
```

median(flipper_length_mm, na.rm = TRUE) <dbl>
197

1 row

Measures of Spread: Variance and Standard Deviation

Sample Variance

The sample variance measures how "widely spread" the data points are around the mean.

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

- When we have a mound-shaped and symmetric distribution, most observations will fall within 2 standard deviations of the mean.
- Variance results in units^2, which typically does not make sense.

Sample Standard Deviation

The sample standard deviation also measures how "widely spread" the data points are around the mean.

$$s = \sqrt{s^2}$$

- Standard deviation is the square root of the variance, measuring spread in the *original units* of the data.
- R syntax:

```
dataset_name %>% summarize(var(variable_name, na.rm = TRUE),
                           sd(variable_name, na.rm = TRUE))
```

Measures of Spread: Variance and Standard Deviation

- Let's find the variance and standard deviation of the weight (*body_mass_g*) of the penguins.

```
1 penguins %>% summarize(var(body_mass_g, na.rm = TRUE),
2                        sd(body_mass_g, na.rm = TRUE))
```

var(body_mass_g, na.rm = TRUE) <dbl>	sd(body_mass_g, na.rm = TRUE) <dbl>
643131.1	801.9545

1 row

- Let's find the variance and standard deviation of the flipper length (*flipper_length_mm*) of the penguins.

```
1 penguins %>% summarize(var(flipper_length_mm, na.rm = TRUE),
2                        sd(flipper_length_mm, na.rm = TRUE))
```

var(flipper_length_mm, na.rm = TRUE) <dbl>	sd(flipper_length_mm, na.rm = TRUE) <dbl>
197.7318	14.06171

1 row

Measures of Spread: Interquartile Range

Sample Interquartile Range

The sample interquartile range measures the spread of the middle 50% of data.

$$\text{IQR} = P_{75} - P_{25}$$

- R syntax:

```
dataset_name %>% summarize(IQR(variable_name))
```

Measures of Spread: Interquartile Range

- Let's find the IQR of the weight (*body_mass_g*) of the penguins.

```
1 penguins %>% summarize(IQR(body_mass_g, na.rm = TRUE))
```

IQR(body_mass_g, na.rm = TRUE)
<dbl>
1200

1 row

- Let's find the IQR of the flipper length (*flipper_length_mm*) of the penguins.

```
1 penguins %>% summarize(IQR(flipper_length_mm, na.rm = TRUE))
```

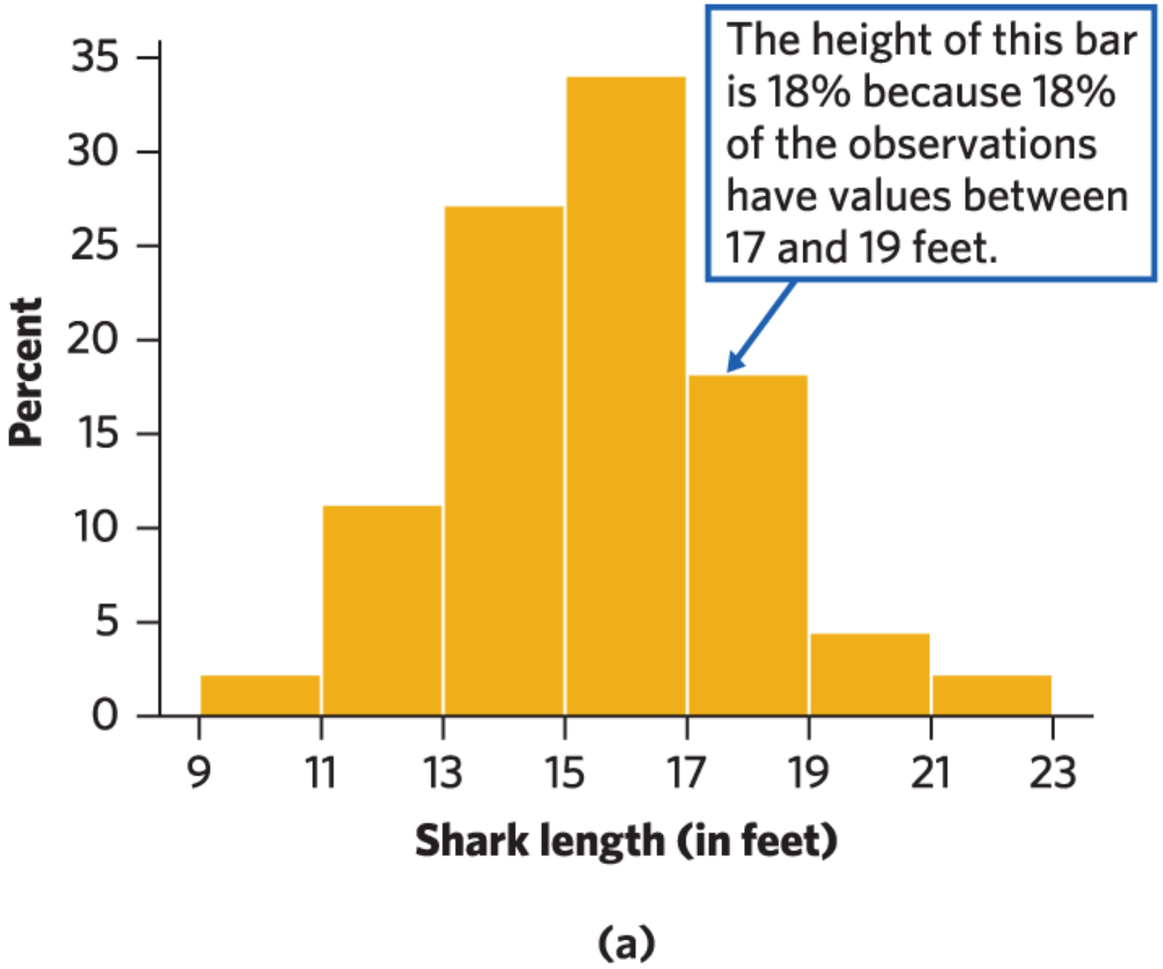
IQR(flipper_length_mm, na.rm = TRUE)
<dbl>
23

1 row

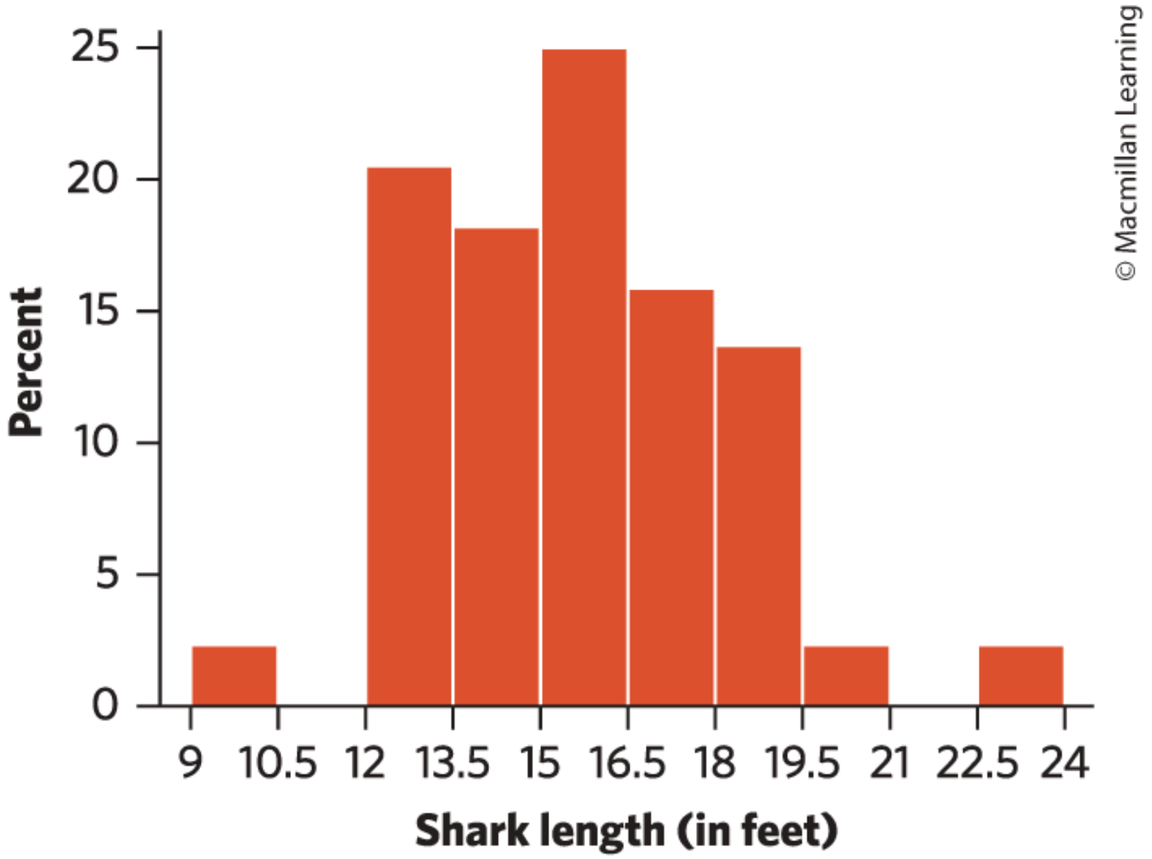
Mean & Standard Deviation vs. Median & IQR

- When should we use the mean vs. the median to describe the center of the distribution?
  - → Mound-shaped and symmetric $\rightarrow \bar{x}$ & s.
  - → Not mound-shaped and symmetric $\rightarrow M$ & IQR.
- … How do we know the shape of the distribution?
- We will explore histograms.

Graphs: Histograms

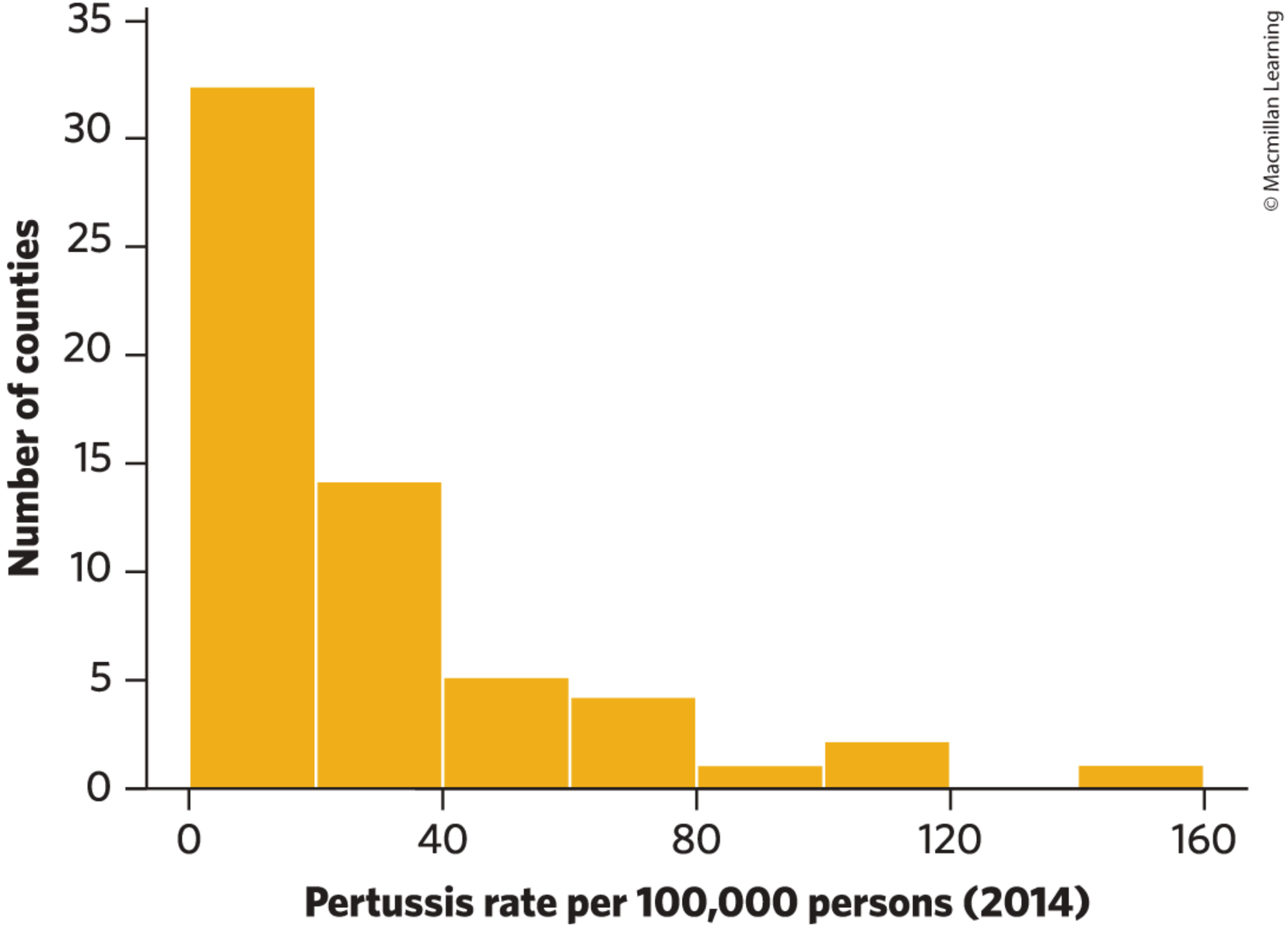


(a)

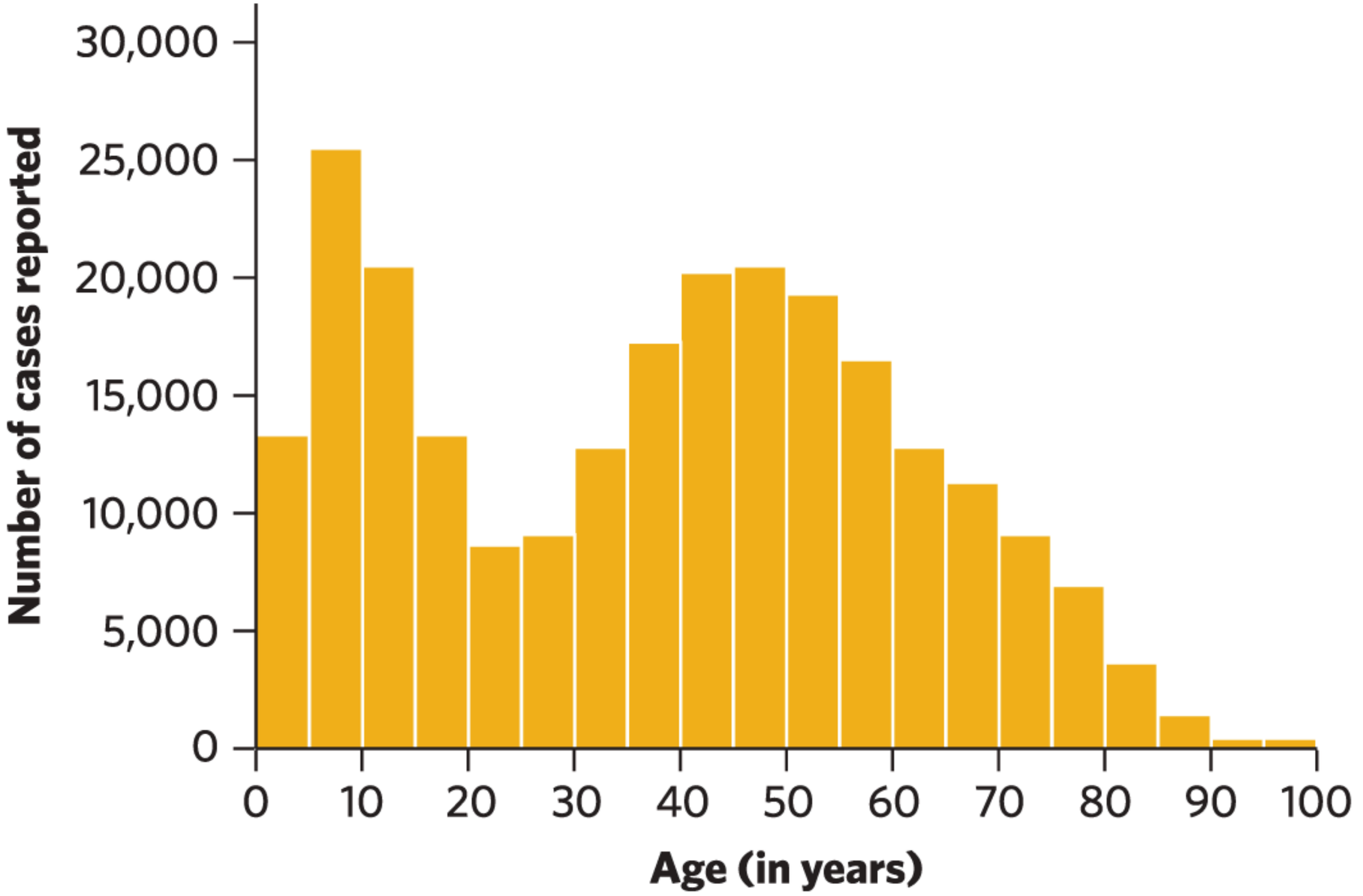


(b)

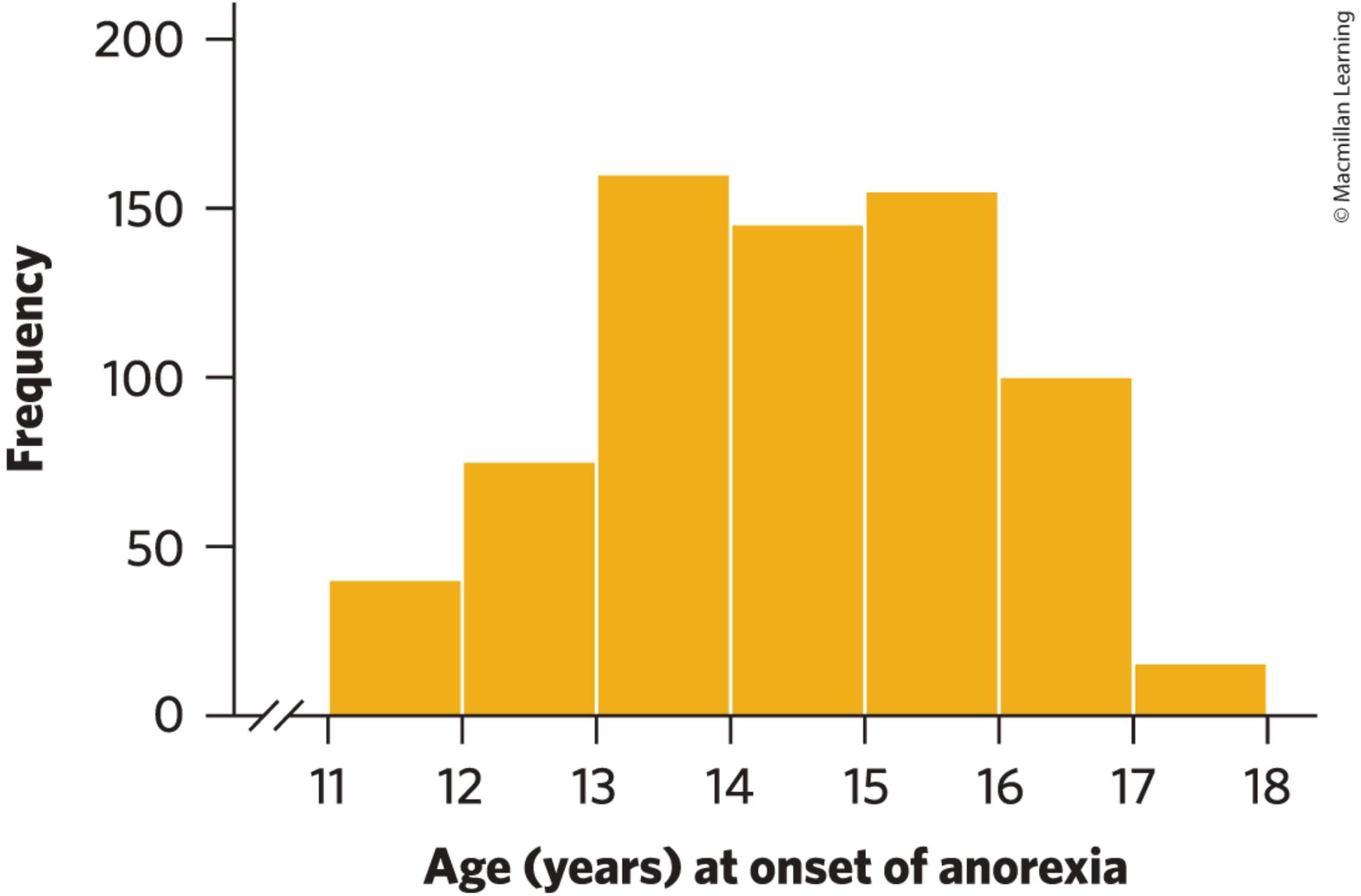
Graphs: Histograms



Graphs: Histograms



Graphs: Histograms



Graphs: Histograms (R code)

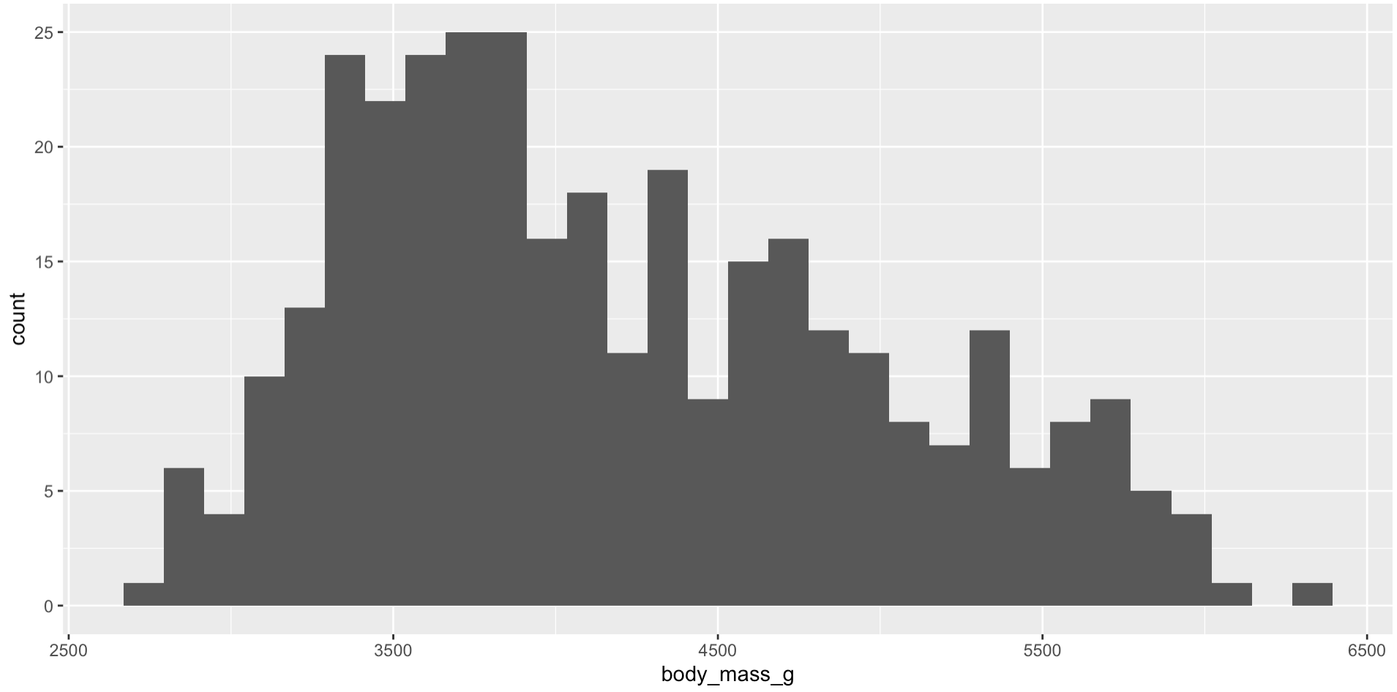
- We are using the `ggplot2` package for graphing.
  - → It will always start with `ggplot()`.
  - → We will then layer elements on top.
- R syntax:

```
dataset_name %>%
  ggplot(aes(x=variable_name)) +
  geom_histogram()
```

Graphs: Histograms

- Let’s look at the histogram of penguin weight (*body_mass_g*):

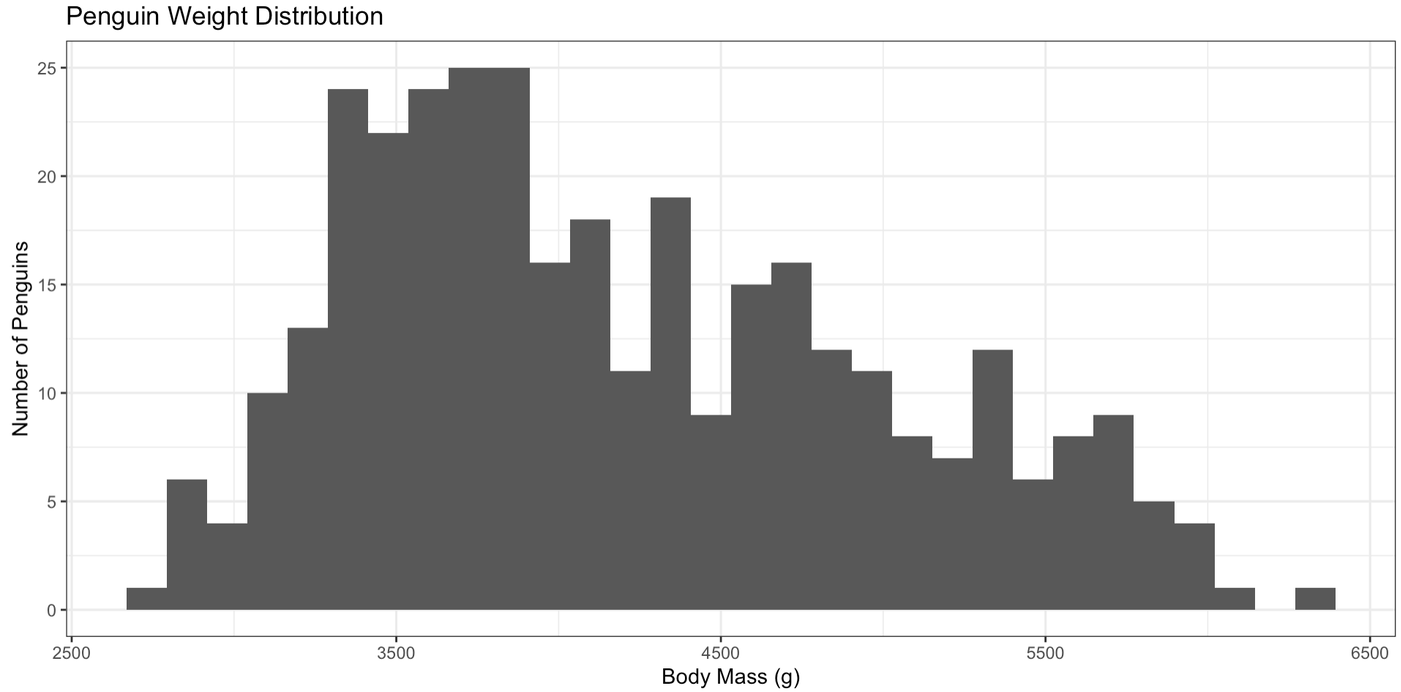
```
penguins %>%
  ggplot(aes(x=body_mass_g)) +
  geom_histogram()
```



Graphs: Histograms

- Let's look at the histogram of penguin weight (*body_mass_g*):

```
penguins %>%
  ggplot(aes(x=body_mass_g)) +
  geom_histogram() +
  labs(x = "Body Mass (g)",
       y = "Number of Penguins",
       title = "Penguin Weight Distribution") +
  theme_bw()
```



Wrap Up

- Today we reviewed estimation.
- Next week, we will review statistical inference.
  - → Confidence intervals
  - → Hypothesis testing
- Get to know you quiz - complete with RStudio.
  - → .qmd ⟶ Quarto
  - → .R ⟶ R script
- Join the Discord server!
  - → If you are already a Discord user, this is a friendly reminder that you can change your display name...