STATS 250 Lab 04

Probability and Scatterplots

Nick Seewald

nseewald@umich.edu

Week of 09/21/2020

Reminders 💡

Your tasks for the week running Friday 9/18 - Friday 9/25 (plus an extra):

Task	Due Date	Submission
Quiz 1	Monday 9/21 11:59PM ET	Canvas
MWrite 1 Initial Draft	Wednesday 9/23 5PM ET	Canvas
Homework 3	Friday 9/25 8AM ET	course.work
Lab 4	Friday 9/25 8AM ET	Canvas

No office hours or Piazza today because of the quiz!

Lab 2 Comments

- Really great work on lab 2!
- R orders levels of categorical variables *alphabetically*
- It's a good idea to have descriptive axis labels for plots. People might not know what "ConditionTreeHealth" means!
- In Dive Deeper 2, the data were collected by the City of Ann Arbor. Only trees on city-owned land are included!
- In Dive Deeper 4, it turns out that the map only contains information about trees on city-owned land. Since the City of Ann Arbor does not own the Arboretum, the data could not be used to answer that research question.

Homework 2 Comments 📝

- A causal *statement* is any sentence that is about causation.
    - "There is not evidence to say that eating chia seeds causes weight loss" **is** a causal statement
    - "Chia seeds do not cause weight loss" **is** a causal statement
    - Causal *statements* do not require causal *relationships*
- Generalizability to a population is a result of sampling: how are data collected?
    - Sample size isn't really a big deal
    - Good (random) sampling = generalizable; bad sampling = not generalizable

Homework 2 Comments 📝

- Review your homework! Even questions you got full credit on.
- You should have comments on every question you lost points on.
- Remember we don't grade every question for correctness.

Weekly Advice

We're focusing a lot on random sampling this week.

Your mileage may vary!

Your results *will* be different from mine and from your collaborators'.

Integer Sequences in R

- A **vector** is a way to hold a collection of things in R. Think of it as a pill organizer.
- This week, we're going to create a vector that holds a sequence of consecutive integers.

```r
1:6
```

```
[1] 1 2 3 4 5 6
```

- Read the colon `:` as "through", so `1:6` is "1 through 6"

Sampling in R

- Think of `1:6` as representing a six-sided die.
- We can "roll" the die by taking a `sample()` from the vector `1:6`

```r
sample(1:6, size = 1)
```

```
[1] 5
```

- Run the `sampleDieRoll` chunk (line ~63) and type what you got in the chat!

Sampling With vs. Without Replacement

- Consider randomly selecting 6 values from the set {1, 2, 3, 4, 5, 6}.
    - Say our first pick is 3.
    - What do we do with 3? Do we take 3 out of the set (don't *replace* it), or do we put it back in (*replace* it)?

Sampling With vs. Without Replacement

- Consider randomly selecting 6 values from the set {1, 2, 3, 4, 5, 6}.
  - Say our first pick is 3.
  - What do we do with 3? Do we take 3 out of the set (don't *replace* it), or do we put it back in (*replace* it)?

```
sample(1:6, size = 6, replace = F)
```

```
[1] 1 6 5 4 3 2
```

Sampling With vs. Without Replacement

- Consider randomly selecting 6 values from the set {1, 2, 3, 4, 5, 6}.
  - Say our first pick is 3.
  - What do we do with 3? Do we take 3 out of the set (don't *replace* it), or do we put it back in (*replace* it)?

```r
sample(1:6, size = 6, replace = F)
```

```
[1] 1 6 5 4 3 2
```

```r
sample(1:6, size = 6, replace = T)
```

```
[1] 1 3 6 5 6 5
```

Sampling With vs. Without Replacement

- Consider randomly selecting 6 values from the set {1, 2, 3, 4, 5, 6}.
  - Say our first pick is 3.
  - What do we do with 3? Do we take 3 out of the set (don't *replace* it), or do we put it back in (*replace* it)?

```
sample(1:6, size = 6, replace = F)
```

```
[1] 1 6 5 4 3 2
```

```
sample(1:6, size = 6, replace = T)
```

```
[1] 1 3 6 5 6 5
```

- Which of these strategies represents die-rolling in real life?

Law of Large Numbers

As you collect more data, sample averages will get close to population averages ("*expected values*").

Roll dice

```
[1] 6
[1] 3 5
[1] 2 1 4
[1] 5 6 4 5
[1] 2 5 2 2 5
```

Average of rolls

```
[1] 6
[1] 4
[1] 2.333333
[1] 5
[1] 3.2
```

Law of Large Numbers

As you collect more data, sample averages will get close to population averages ("*expected values*").

Roll dice

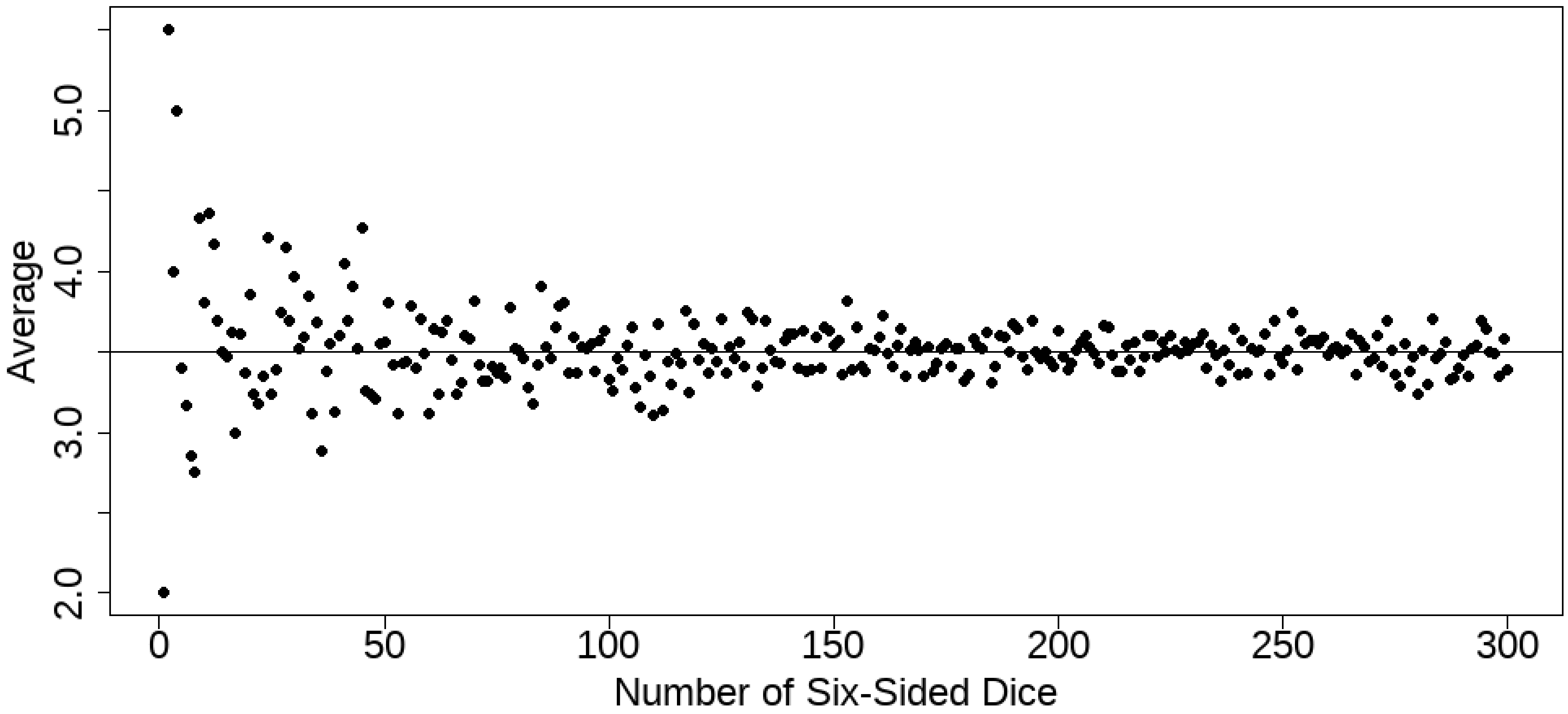
```
[1] 6
[1] 3 5
[1] 2 1 4
[1] 5 6 4 5
[1] 2 5 2 2 5
```

Average of rolls

```
[1] 6
[1] 4
[1] 2.333333
[1] 5
[1] 3.2
```

The mean seems like it's trying to do something, but it's too variable to really see what's happening.

Law of Large Numbers



Expected Value

We can compute the value that the sample averages will converge to!

$$\sum_{i=1}^{n} x_i \cdot p_i$$

- Σ means "summation" (addition)
- x_i is the value (in our case, 1, 2, 3, 4, 5, or 6)
- p_i is the *probability* of observing the value

For the six-sided die, the expected value is

Expected Value

We can compute the value that the sample averages will converge to!

$$\sum_{i=1}^{n} x_i \cdot p_i$$

- Σ means "summation" (addition)
- x_i is the value (in our case, 1, 2, 3, 4, 5, or 6)
- p_i is the *probability* of observing the value

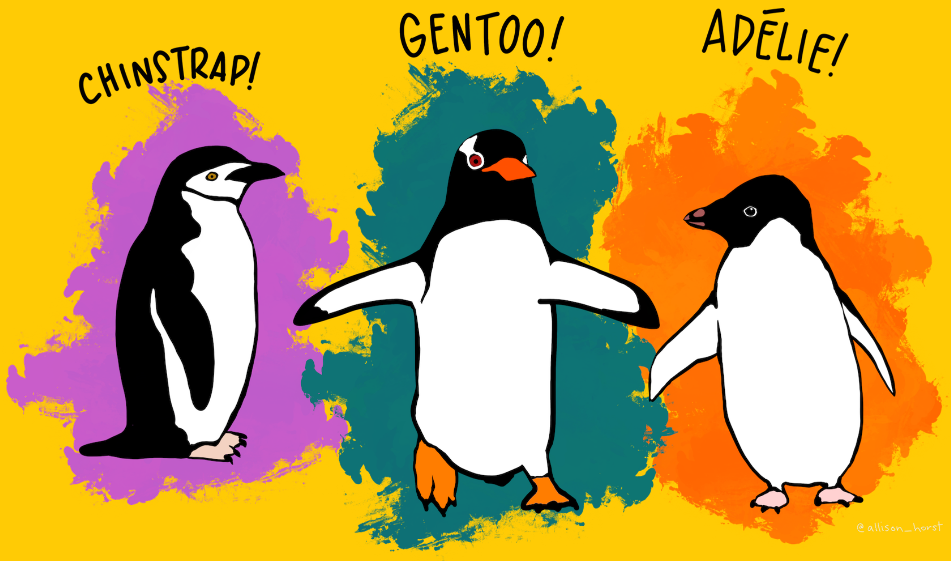
For the six-sided die, the expected value is

$$1 \cdot \left(\frac{1}{6}\right) + 2 \cdot \left(\frac{1}{6}\right) + 3 \cdot \left(\frac{1}{6}\right) + 4 \cdot \left(\frac{1}{6}\right) + 5 \cdot \left(\frac{1}{6}\right) + 6 \cdot \left(\frac{1}{6}\right) = 3.5$$

Penguins!

```
penguins <- read.csv(url("https://raw.githubusercontent.com/STATS250SBI/palmerpenguins/master/inst/e
```

```
str(penguins)
'data.frame':    333 obs. of  8 variables:
 $ species          : chr  "Adelie" "Adelie" "A
 $ island           : chr  "Torgersen" "Torgers
 $ bill_length_mm   : num  39.1 39.5 40.3 36.7
 $ bill_depth_mm    : num  18.7 17.4 18 19.3 20
 $ flipper_length_mm: int  181 186 195 193 190
 $ body_mass_g      : int  3750 3800 3250 3450
 $ sex              : chr  "male" "female" "fem
 $ year             : int  2007 2007 2007 2007
```



Scatterplots

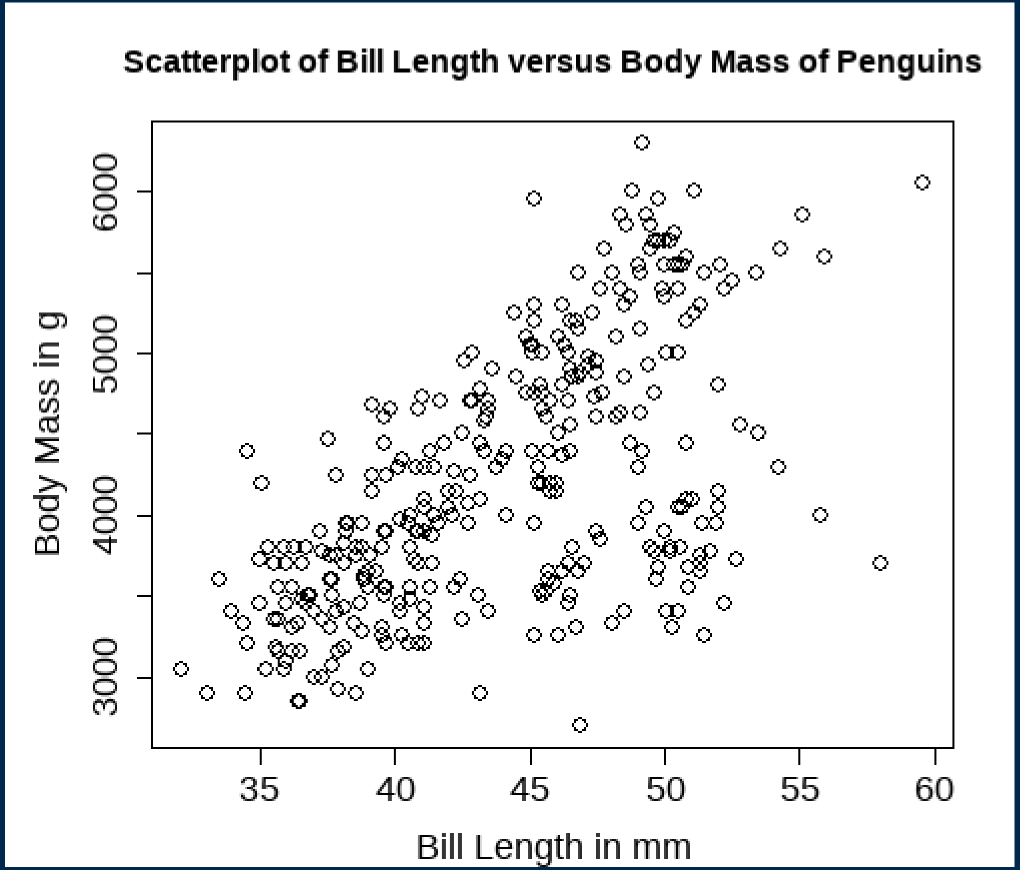
A **scatterplot** is a way to display the relationship between quantitative explanatory (x) and response (y) variables.

The data are paired (x, y) and then each pair is plotted on a grid.

We can use scatterplots to look for **associations** between these quantitative variables.

Scatterplots (line ~142)

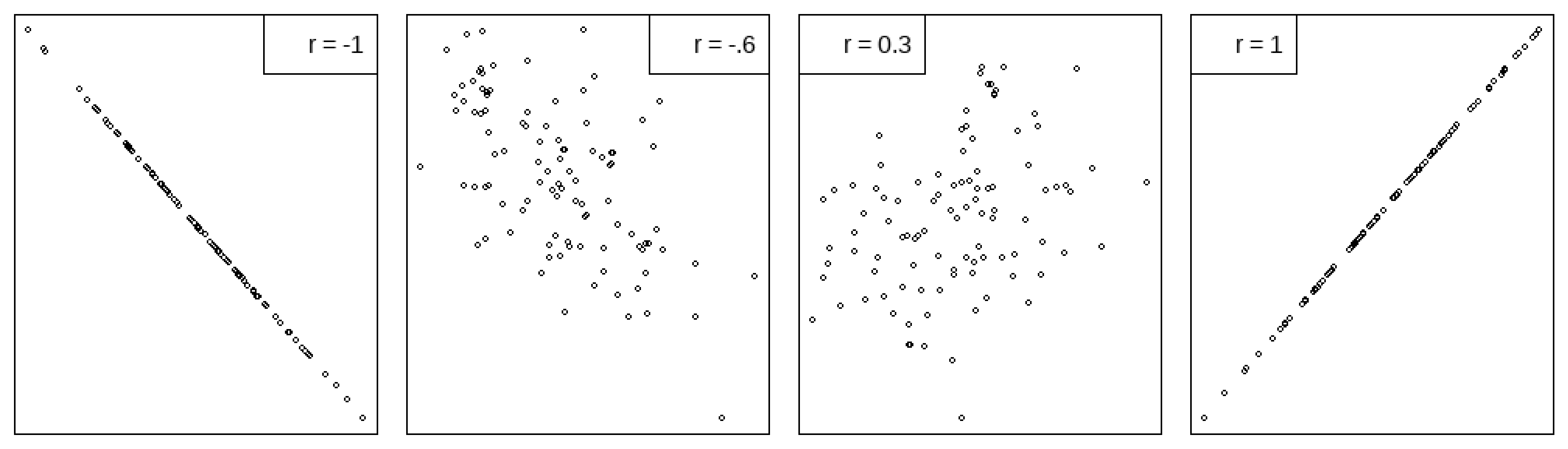
```
plot(penguins$bill_length_mm,
     penguins$body_mass_g,
     main = "Scatterplot of Bill Length versus
     xlab = "Bill Length in mm",
     ylab = "Body Mass in g")
```



- positive association
- reasonably linear
- moderately strong
- no apparent unusual points

Correlation

- The **correlation** between two quantitative variables quantifies the strength of the *linear* association.
- Denote correlation by r
- As $|r|$ gets close to 1, the linear relationship becomes stronger



Correlation (line ~155)

We can find the correlation between two quantitative variables in R using the `cor()` function.

```r
cor(penguins$bill_length_mm, penguins$body_mass_g)
```

```
[1] 0.5894511
```

Lab Project ⌨

You will be **randomly** moved to a breakout room for the rest of the lab (minus ~10 minutes)

Your tasks

1. Introduce yourself to your collaborators!
2. **Work together** to complete the "Try It!" and "Dive Deeper" portions of the lab assignment by copy/pasting and modifying appropriate code from earlier in the document.

How to get help

- I'll be floating around between breakout rooms to check on everyone
- Use the "Ask for help" button to flag me down
- Let me know when you're done

"Exit Ticket"

Please take 1-2 minutes to complete the survey at

bit.ly/njs_ticket4