Week 4: Data Management

Session 1

Spring 2020

iClicker quiz: Question 1

Your book mentioned several good reasons to collect information on who, what, where, when with your observations.

Which of the following is NOT a reason to collect all of the Ws?

A It means you sampled correctly and without bias

B It allows the data to be used for multiple purposes

C It ensures that the data you collect can be checked for accuracy

D It means that you won't forget some important aspect of the data

E It allows someone else to repeat the exercise exactly

iClicker Quiz: Question 2

Which of the following R concepts has been the most challenging or frustrating for you?

A Loading data files into R

B Strings literals vs. variables

C Sequences and series

D Navigating swirl

Announcements

- I'm experimenting with some alternative ways to build lecture slides.
- Slides may have different looks and feels over the next several weeks.
- I will ask for your feedback on the slide styles and content later in the course.
- Notetaker
- Do NOT install the Excel Analysis Pack. We will not be using Excel for data analysis. I will provide updated information and instructions for passages in the book the refer to analyses in Excel.

For today

- What did we do last week?
- Data concepts
- Recording data
- Variables
- Group activity time

Data: some key concepts

Data and *metadata* "Data without metadata is meaningless."

But. . . . what is metadata?

When we design data sheets, we should think like R:

Row format

- Each row is an *observation*.
- Each row contains columns for:
- Metadata (who, where, etc.)
- Explanatory variables
- Response variables

Sampling Units: This is a subtle concept, we'll keep returning to it throughout the course.

Important concepts in data entry

What are some important questions your data sheets should answer?

- Who, what, where, when
- Variables (there may be multiple variables in each category)
- Notes: unusual conditions, etc.

Which of these could be considered *metadata*?

It's better to record all of the data and other observations that could be relevant later, even if you don't end up using all of the information for your analyses.

Bears/poops group time

I sincerely apologize for the technical issues last class. Thank you all for bearing with me.

General questions?

Remember the peer evaluations. They will be available on Moodle starting Thursday.

Why do we need to sample?

Variables: key concepts

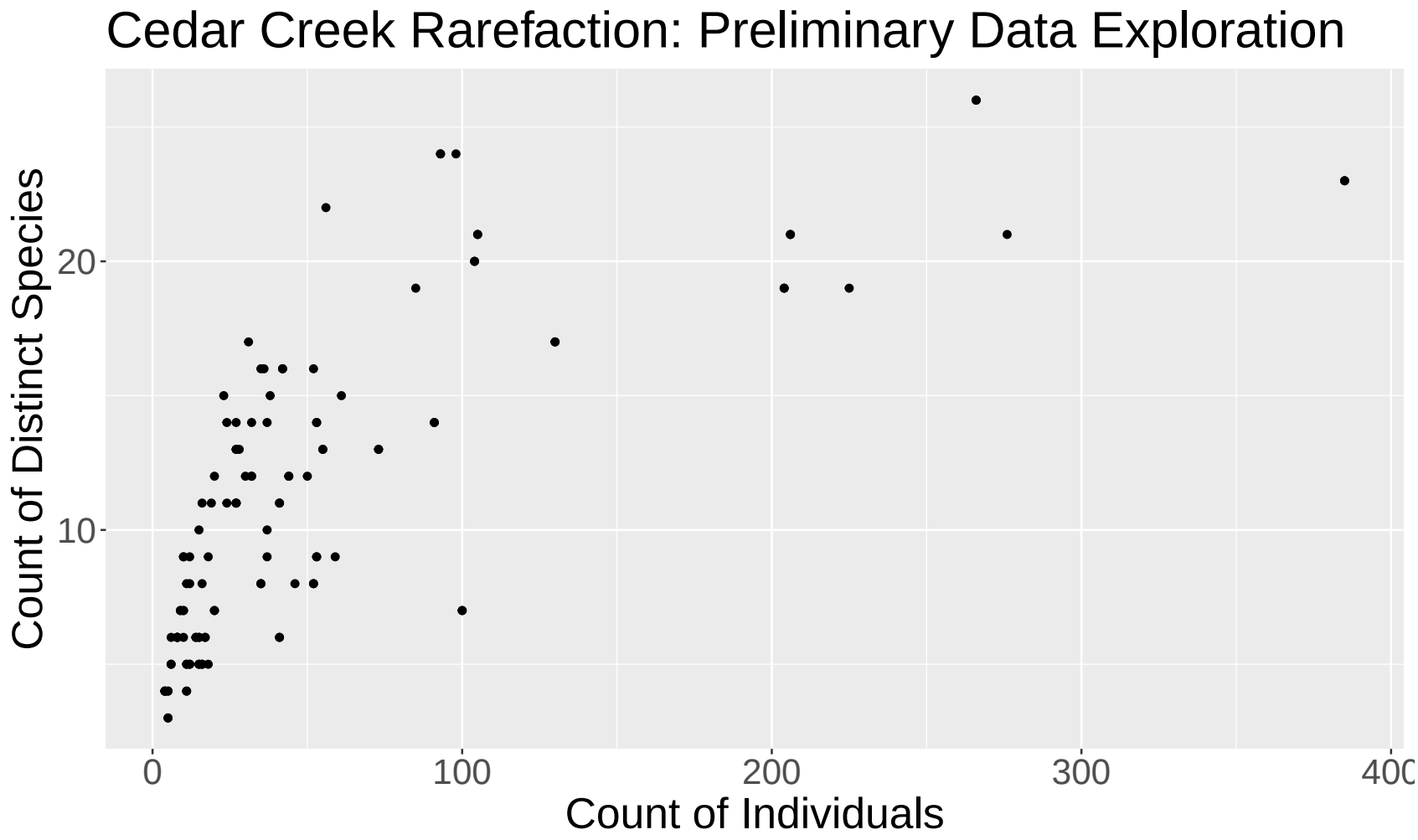
Several pairs of terms are often used to describe different types of variables:

- Independent/dependent
- Predictor/response
- Explanatory/response

What do they mean?

What are some important data types?

Variables example: Can you identify the predictor and response?



What to record?

The metadata:

1. Who recorded the data
2. What is the *sampling unit*
3. Where was the data recorded
4. When was the data recorded

What to record?

The data (i.e. the variables):

1. Response Variable
   - 1.1 the variable of interest you are trying to estimate
   - 1.2 similar to/or sometimes called the "dependent variable"
2. Explanatory Variables
   - 2.1 Variables that influence the response variable
   - 2.2 similar to/or sometimes called the "independent variables" or "predictor variable"

Some examples of variables

Can you identify the predictor and response variables in these *models*?

From high school math, you might remember the equation for a line:

$$y = mx + b \tag{1}$$

We'll learn all about statistical models of the form:

$$y_i = \alpha + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \epsilon \tag{2}$$

Why did I call these equations models?