

# Analysis of Environmental Data

Data, Sampling, and Intro to Frequentism

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst  
Michael France Nelson

# What's In This Deck?

## Slides

- All about data
- Samples/Populations
- Intro to Frequentism

## Key Take-Home Concepts

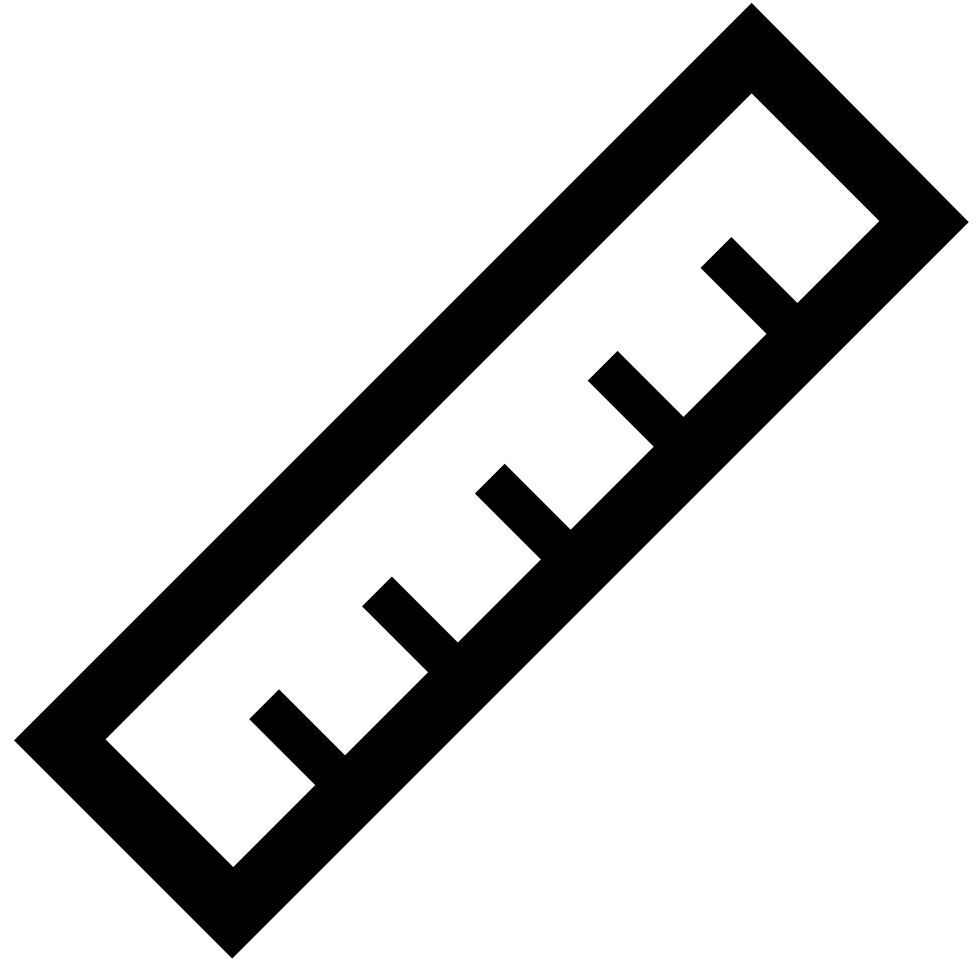
- Data: measurement and scales.
- What is the Frequentist paradigm?
- Frequentist interpretation of populations and samples
- Null and alternative hypotheses
- Frequentism may challenge your intuition.

# Data and Measurement

# Scale

## How do we measure data?

- What do we mean by *scale*?
- Measurement scale terms:
  - Discrete, continuous
  - Numeric, categorical
  - ratio, interval



# Scale

The word *scale* has a lot of meanings... In our context *measurement scale* or *data scale* refers to:

- A measurement scheme that answers questions like
  - Is it numerical?
    - Can it have negative numbers?
    - Can it contain fractions?
    - Is there a *true zero*?
  - Is it categorical?
    - Is there a meaningful ordering of the categories?

# Scale

A measurement scale is what we use to *quantify* a variable, i.e. an *attribute* of a sampling unit.

**The choice of scale may be context-dependent:**

- Does it reflect an intrinsic property of a variable?
- Does it reflect an intrinsic property of how we choose to measure it?

For example: age measured in years vs. age measured in age class.

# Numeric or Categorical

## Is our variable *quantitative* or *qualitative*?

Numeric measurement scales: our variable is measured as a numeric quantity.

Qualitative: our variable can be classified into a *category*.

## Does our *qualitative* variable have a sensible *order*?

- Categories that have a meaningful order are called *ordinal*
  - There may be an order, but the inter-category intervals cannot be directly compared.
- If there is not an intrinsic order they are *nominal*.

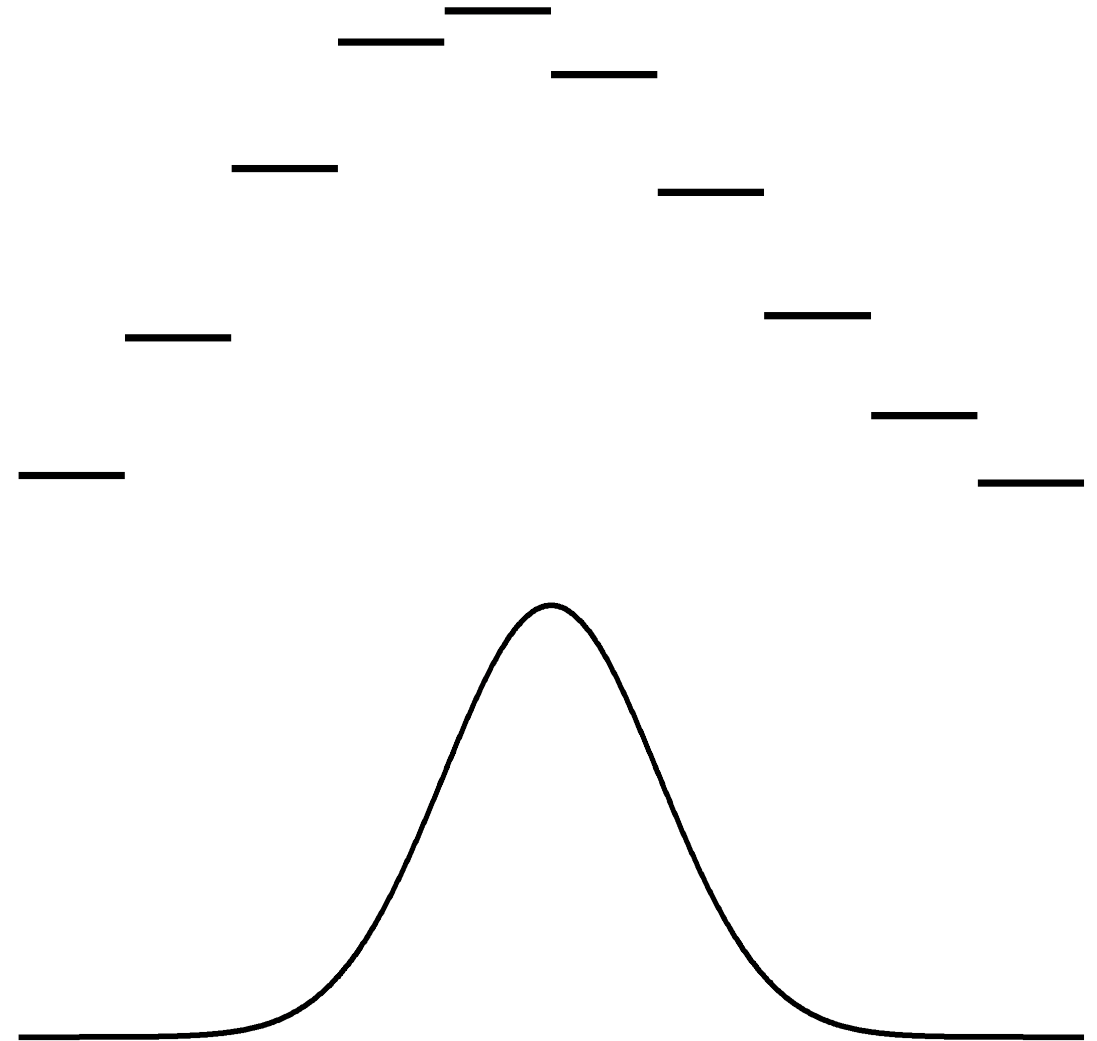
# Numeric Scales: Discrete and Continuous

## Discrete cannot take on intermediate values

- Counts, presence/absence, etc

## Continuous variables can *in principle* take on any intermediate value:

- They are *real* or *rational* numbers.
- Our ability to measure may not capture intermediate values, but they are still continuous.





# Announcement: Read the Announcements!

## **PowerPoint's Stock Art!**

I try to include items of high relevance

Answers to many of your questions are found in the announcements!

For real: pay attention to the announcements!



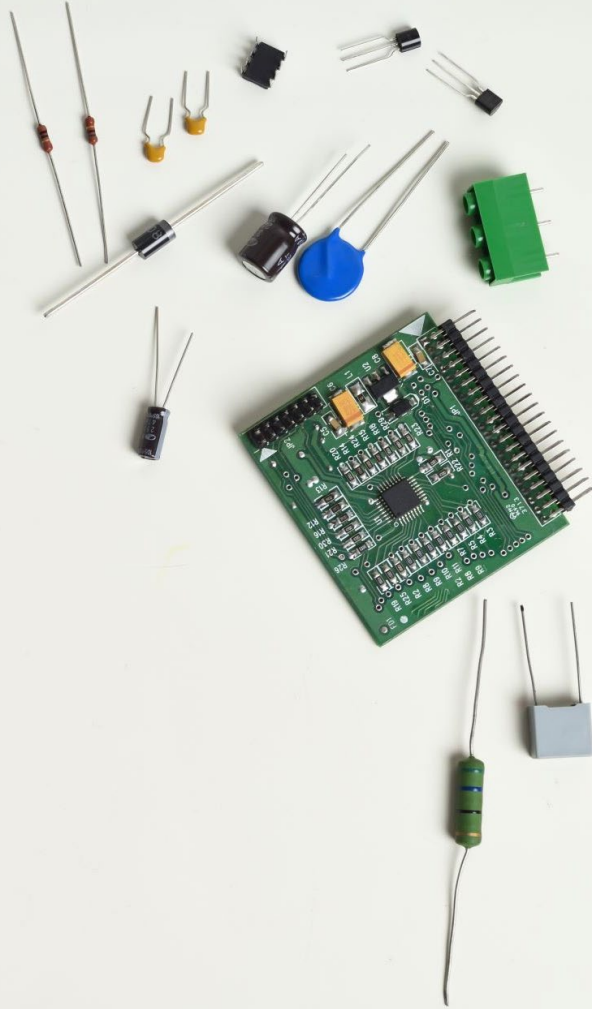
# Announcements

- Updated slide deck 2 posted!
- Default group: make sure you joined your in-class activity group.
  - If you're in the default group, you'll receive a zero grade.

# Announcements 2022

- I've added a non-graded data scale practice activity for today's class
  - You may have to re-load the course GitHub site to see the link.
- We made it to week 2!
- Due this Sunday:
  - Reading Questions – week 2
  - Software Setup
- Due Sunday Sep 25:
  - Reading Questions – week 3
  - Using R Notebooks
  - DataCamp: Intro to R

# Announcement: Graduate Computer Lab



## Holdsworth Room 331.

- Holdsworth door keys will unlock.
- Workstations, whiteboard, collaboration space.
- Use it to your advantage!
- We want to demonstrate the need for computers and collab space to the dept and college... this is an excellent way to do it!
- Please let me know if/when you use it, and how well it works for you.

# Tip of the Day: RMarkdown Themes

## RMarkdown Themes are Awesome!

- Themes apply a coherent look and feel to your entire RMarkdown document.
- There are several ‘built-in’ themes.
- There are lots of extended themes you can check out.
- Just specify the theme in your YAML header.

```
1 ---
2 title: "RMarkdown Themes are Awesome"
3 author: "Michael F. Nelson"
4 date: '2022'
5 output:
6   html_document:
7     toc: TRUE
8     toc_float: TRUE
9     theme: readable
10 ---
```

Note the indentation scheme

Other themes include ‘united’ and ‘darkly’

# Interval and Ratio scales

## Is there a *true zero*?

- Degrees in Kelvin: *absolute zero* is the absence of movement of particles.
  - You can't go lower than *absolute zero*.
- Degrees in Celsius: zero is centered around the freezing point of water.
  - Negative values are possible.

**Interval scales can have negative numbers**

**Ratio scales are (usually) non-negative**

# Circular Scales

## **Circular scales wrap a maximum value back to zero**

- Circular scales are not as common, but they occur when thinking about angles.
- Examples include wind direction and aspect
  - Both are measured in degrees (or radians).

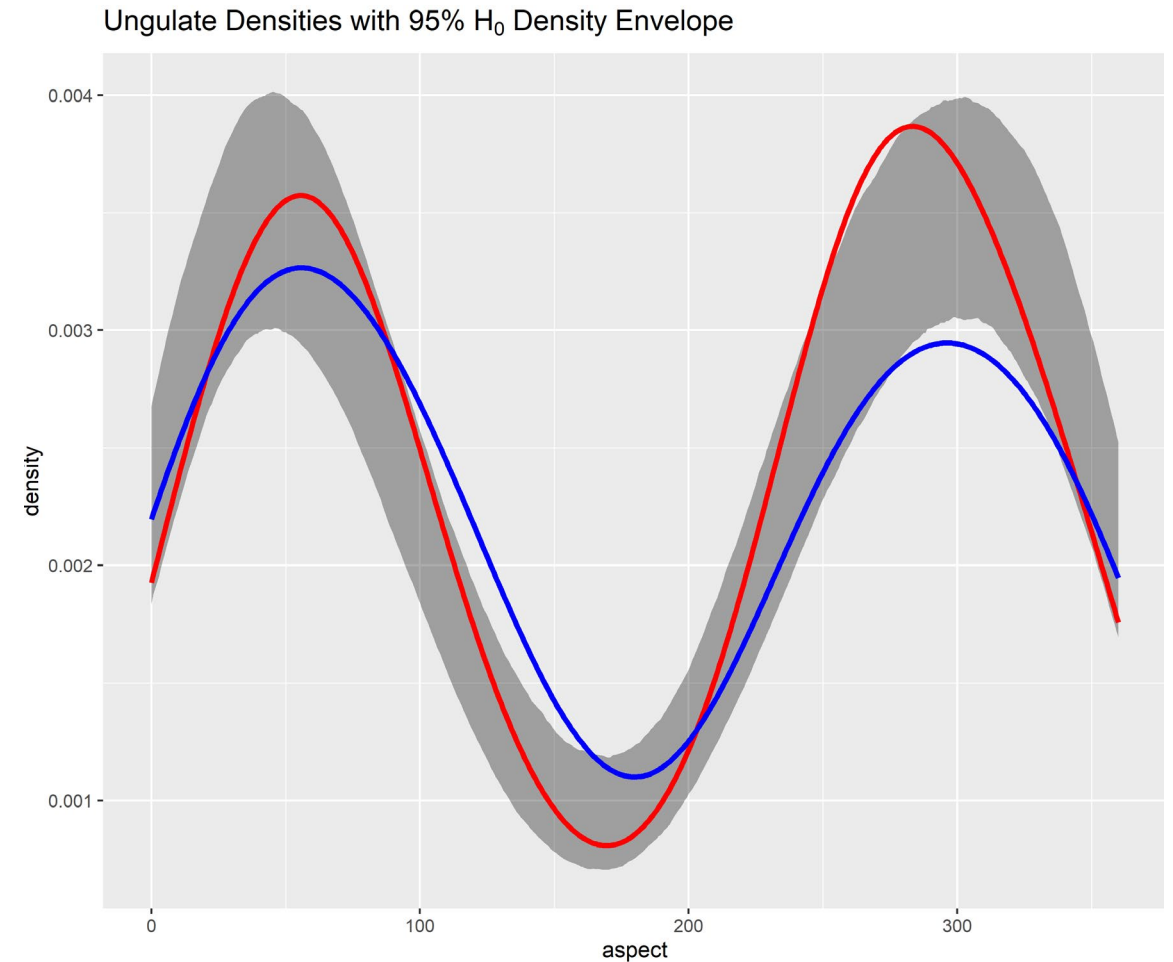
## **Modulo operator**

- The wrapping behavior of circular scales is just like addition in *modular algebra*.

# Circular Scale: Aspect

**Circular scales work well for things like days of year, direction, etc.**

An example of circular data in practice: the aspect of ungulate herd location observations in mountainous terrain:





# Converting Among Scales

**Sometimes it's convenient, or necessary, to convert a variable to a different scale.**

For example: consider count data that consist of mostly 0 and 1, with only a few values greater than 1.

- It might be useful to convert this to binary data, i.e. presence/absence.

**Aggregating into categories**

- Age, size, weight classes
- These convert numeric into ordinal scales.

# Converting Among Scales

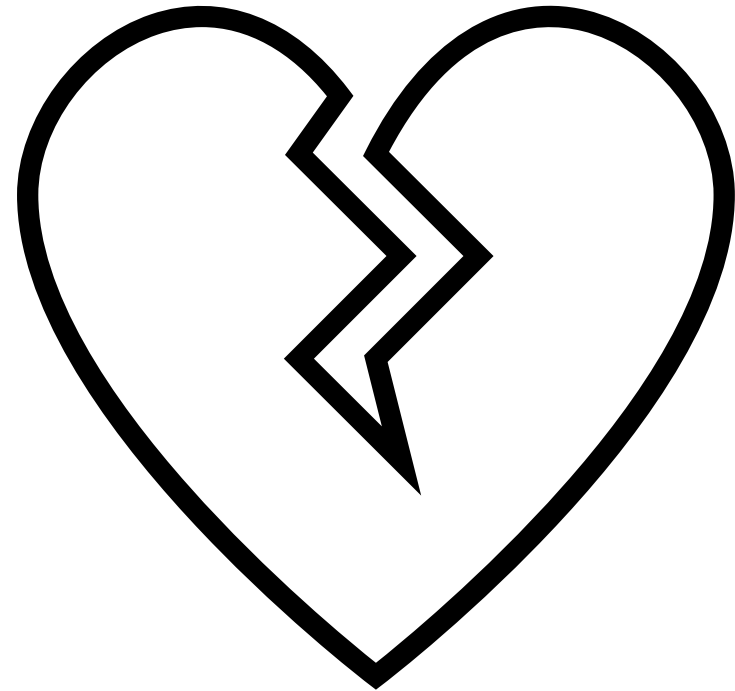
Some conversions are *destructive*: you lose information in the conversion process.


Aggregating numeric scales into categories.

e.g. converting absolute ages into age classes:

If you know the age, you can assign to an age class.

If you know only the age class, you can't make the reverse transformation.



  Baby, don't break my data!  

# Data Scales and Statistical Models

**Most of the time we build models using numbers, on either discrete (integers) or continuous scales.**

## **Theoretical (parametric) distributions as models**

- Remember that we fit models to data, not the other way around.
- Sometimes theoretical distributions are only *approximately* good models for our data.
  - For example: the Normal distribution often fits data, such as weights, very well, but its sample space includes negative numbers.
  - Some discrete distributions that are useful for count data allow for unrealistically large observations.

# The Row Data Paradigm

**Storing your data in a Row Data format will simplify your life!**

**Most of the datasets we work with can be written in a 2-dimensional table.**

## Rows

- Rows are observations
- Rows are samples
- Rows are sampling units (sometimes)
- A row is a collection of observations on a single entity.
- Rows are

## Columns

- Columns represent attributes
- Columns are variables
- Columns are properties
- Columns are fields

# Row Data Paradigm – What is it?

## A Toy Example

Species	Time	Mood	Body Mass	Ambient temperature
Crow	13:05	Sassy	413	25
Raven	10:01	Serious	980	-25
Snowy Owl	01:47	Hungry	2101	NA
Snapping Turtle	13:45	Angry	30000	15

## Why should I use it?

- Corresponds to common and convenient data structures in R and other programs.
- Data type/scale is consistent within a column.
- It's easy to look up an attribute (in a column) for an individual (a row).
- Non row formatted data will cause lots of data import headaches!

# Sample and Population

Prelude to Frequentist Thinking

# Concepts and Learning Objectives

- Key differences between population and sample
- Parameters and statistics
- Description and inference
- Statistical and ecological populations
- Sampling units

# Populations and samples

## Populations are large

- We [typically] **cannot observe all** individuals in a population
  - This is a cornerstone of Frequentist thinking
- We have to make informed guesses about the population from *samples*

## Samples are a subset of a population

- We **can observe all** sampling units in a sample
- We can completely characterize the properties of a sample

## We use the sample to make informed guesses about the population

- This is the heart of inferential statistics.





# Populations, samples, parameters, and statistics

## Population/sample and parameter/statistic are *parallel* concepts

- **Populations** have *parameters*, intrinsic characteristics of the entire population.
  - We can't calculate population parameters directly.
- We can calculate *statistics* from **samples**.

## We use statistics to infer information about population parameters

- This is the basis for *inferential statistics*.

# Samples, Sampling Units, and Variables



- A sample is a **group of observations** taken from a larger *population*.
- A sampling unit (SU) is the unit/entity/thing of interest for the research question.
- A variable is an attribute of the SU

# Populations, Samples, Sampling Units, and Variables

## These concepts form a *nested* structure:

- A statistical population is [usually] a subset of an ecological population
- A sample is a subset of a statistical population
- One sampling unit is a subset of a sample
- A Variable is a quantity measured on a single sampling unit



# Statistical and ecological populations

**We'll use Bullheads to illustrate the differences.**



**Brown Bullhead illustration by Duane Raver  
(USFWS)**

# What is an ecological, or biological, population?

## **The collection of all possible sampling units.**

- The scale of the research question may or may not encompass the entire ecological population.
- *A statistical population* is usually a subset of the *ecological population*.
- An ecological population does not generally vary based on the scope of a research question.

## **The bullhead ecological population**

- All individual fishes across the entire species range

# What is a statistical population?

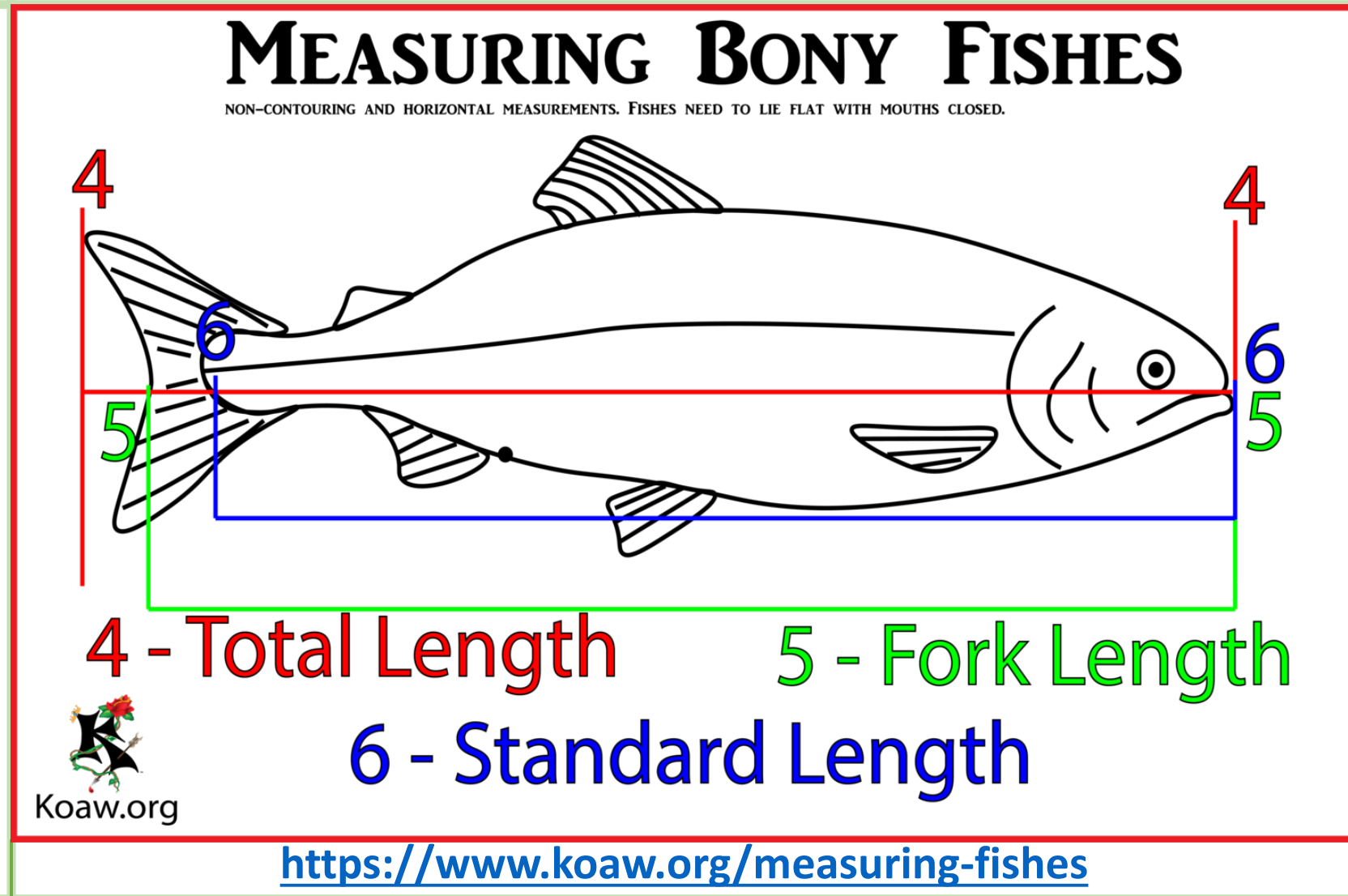
## **The statistical population depends on the scope of the *research question***

- Suppose we were studying bullhead in a single lake:
  - ecological population: entire species range
  - statistical population: the lake
- What about bullhead in Massachusetts?
  - ecological population: entire species range
  - statistical population: all bullhead within MA
- Note that the ecological population did not change.



# Bullhead sampling units and variables

Which bullhead attributes might we want to measure or observe?



# Bullhead sampling units and variables

## Both the sampling unit and variables are context-dependent

- Suppose we were studying bullhead in a single lake:
  - sampling unit could be individual fishes
  - variable might be total length
- Suppose we wanted to compare average bullhead size in multiple lakes
  - The sampling unit might be *individual fishes* or *individual lakes*
  - The choice of SU would depend on our question
  - Variable might be average length in a lake



# Sampling units are context dependent: McGarigal testimonies

## **Some interrelated questions to ask yourself for each testimony:**

- What are the spatial and temporal scales?
- What is the statistical population?
- What are the sampling units?



# Testimony 1: Spatio-temporal scales

- **Temporal scale:**

- Observations were taken yearly for 10 years.

- **Geographic scale:**

- A single mountaintop in the White Mountains National Forest

# Testimony 1: Variables

## Which quantities were measured?

1. year

2. 'upper elevational distribution'

- This is vague in the text...
  - Is it the elevation of the highest observed nest?
  - Is it an average elevation of all nests of a set of high elevation species?

# Testimony 1: Populations and sampling units

## Populations

- Statistical: Collection of nesting sites on one peak
- Ecological: All possible nesting sites of the bird species considered.

## Sampling units

- The SU appears to be individual nesting sites.
  - But recall the ambiguity from the previous slide

# Testimony 3: Scales

- **Temporal scale:**
  - Observations were taken yearly for 10 years.
- **Geographic scale:**
  - Entire White Mountains National Forest
- Same variables as before: year, 'upper elevational distribution'

# Testimony 3: Populations and sampling units

## Populations

- Statistical population: Collection of nesting sites on all measured peaks in the White Mountains
- Ecological population: All possible nesting sites of the bird species considered.

## Sampling units

- Appears to be individual nesting sites.
  - But recall the ambiguity from testimony 1
- The SU could also be individual mountain tops within the White Mountains in this testimony

# Recap

- Key differences between population and sample
- Parameters and statistics
- Description and inference
- Statistical and ecological populations
- Sampling units

# In-Class Data Scales Practice

- [Find the instructions on GitHub.](#)
- This is non-graded, but it's good practice for the assignments.



# Note to self: Re-Start Recording

# Intro to Frequentist Thinking

# Concepts and Learning Objectives

Brief introduction to Frequentism

Frequentist interpretation of populations and samples

Null and alternative hypotheses

Frequentism will challenge your intuition

# What is Frequentism?

## Inferential framework

- Most widely used framework.
  - It has many pros and cons
- Requires assumptions.
  - They are often reasonable, but sometimes not
- Many tools are robust to violations of assumptions
- Powerful theoretical basis and sophisticated analytical tools
- Frequentism focuses on the modeling *process*, not the outcomes of a particular experiment

**Other frameworks exist...** We'll only briefly discuss them in this course.

# Frequentism essentials

## Key Frequentist assumptions include

- Population exists, is infinite.
- Population parameters are true, but unknowable quantities.
- When we specify a model, there exist **true** model parameters (but they are unknowable).

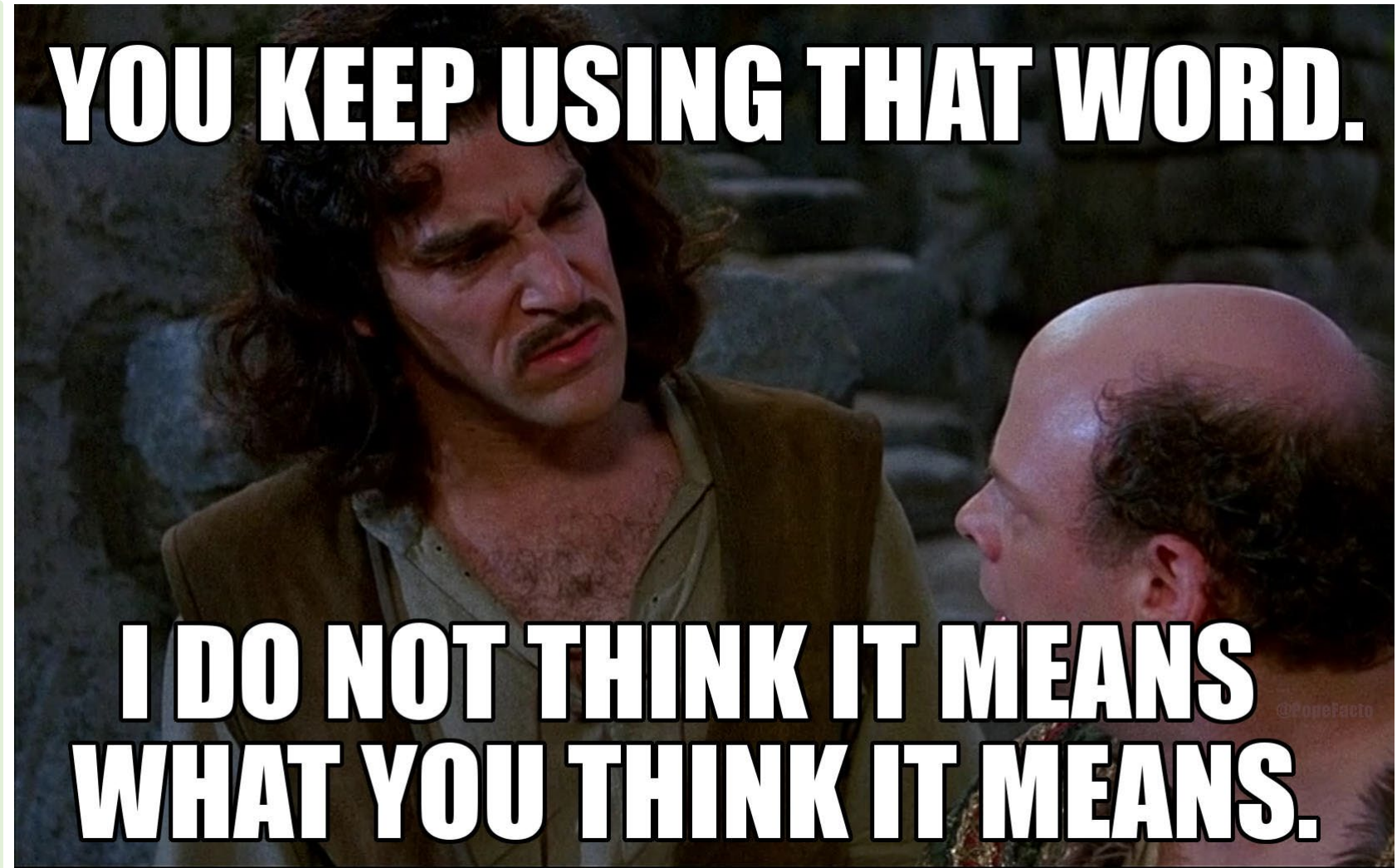
## Frequentism is based upon hypothetical infinite resampling

- Frequentist assumptions are often **asymptotically** true.
- Source of misconceptions about terminology

## Hypothesis testing: $H_0$ and $H_a$

# Null and Alternative models

- Hypothesis testing: allows for quantification of *confidence* and *significance*.
- ‘Confidence’ and ‘significance’ are tricky terms in statistics... They don’t have the same meaning as in everyday language



[bigmedium.com/ideas/mvp-does-not-mean-what-you-think-it-means.html](https://bigmedium.com/ideas/mvp-does-not-mean-what-you-think-it-means.html)

# Note on terminology

## ***Confidence or significance:***

**The Frequentist definition is a little hard to explain to non-statisticians, but we're often called on to do so!**

### **What Confidence Interval Does Not Mean**

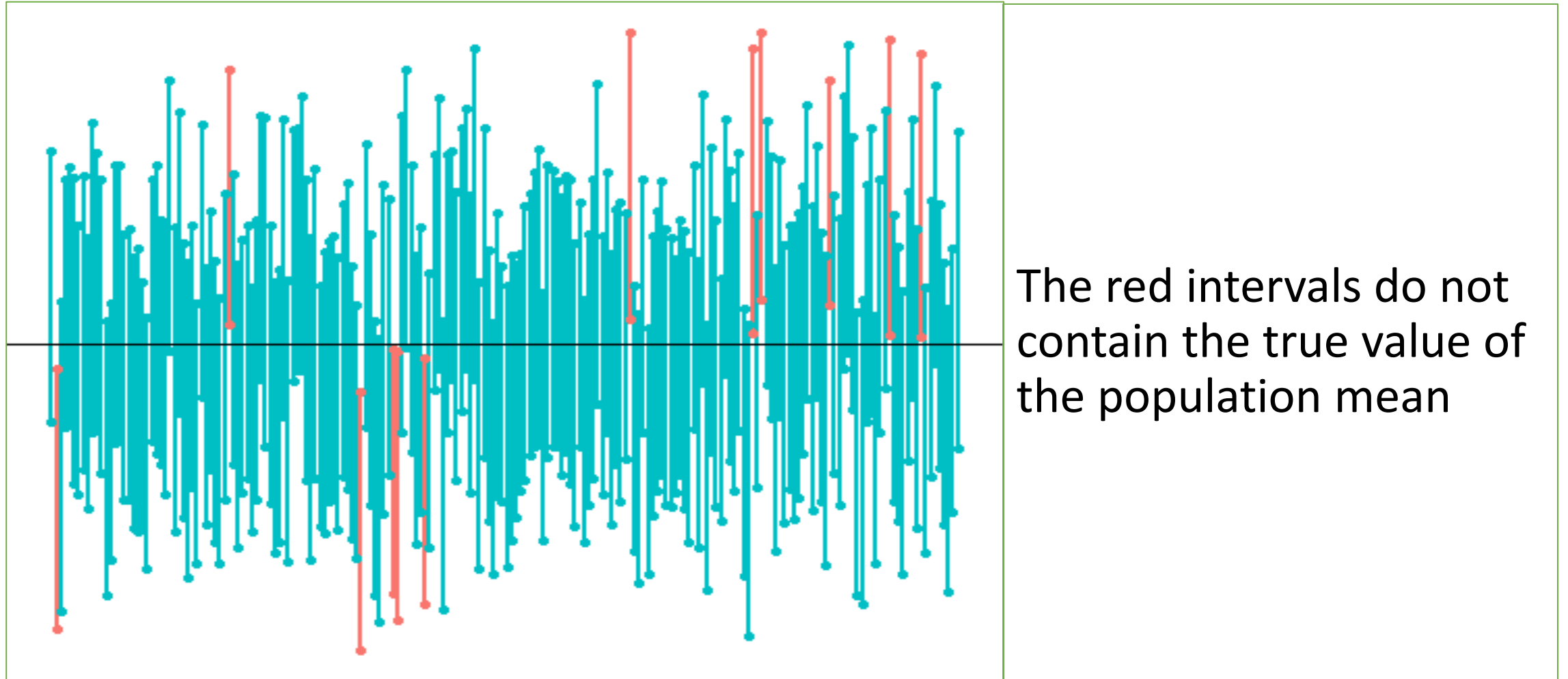
“I’m 95% sure my CI contains the true value.”

No: It either does or does not, but you can't know.

### **What it actually means**

“If I were to repeat the experiment many times, approximately 95% of the CIs I construct would contain the true population parameter”

# A set of 200 95% Confidence Intervals



The red intervals do not contain the true value of the population mean



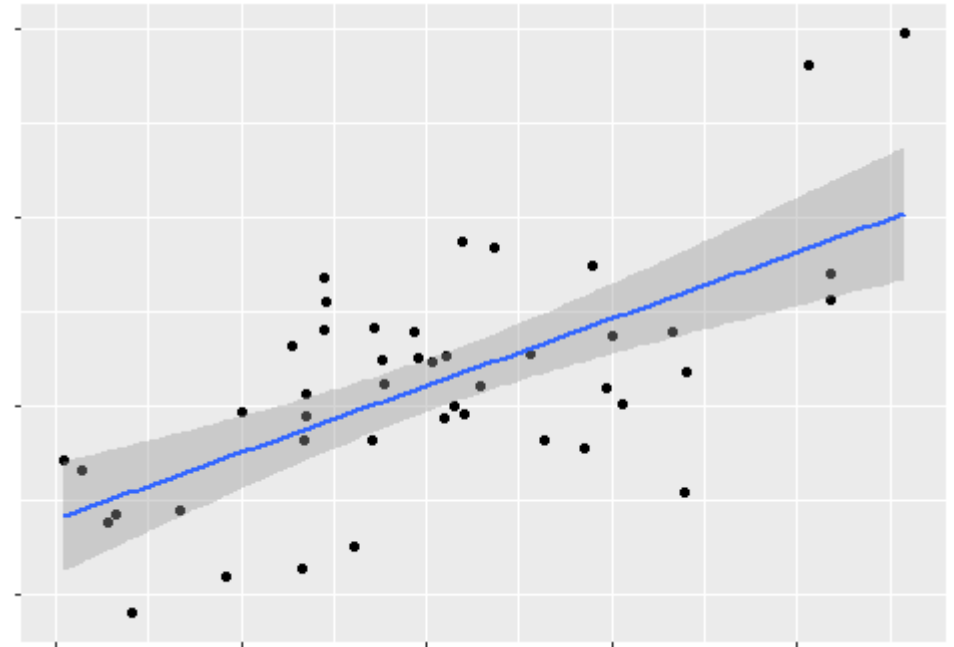
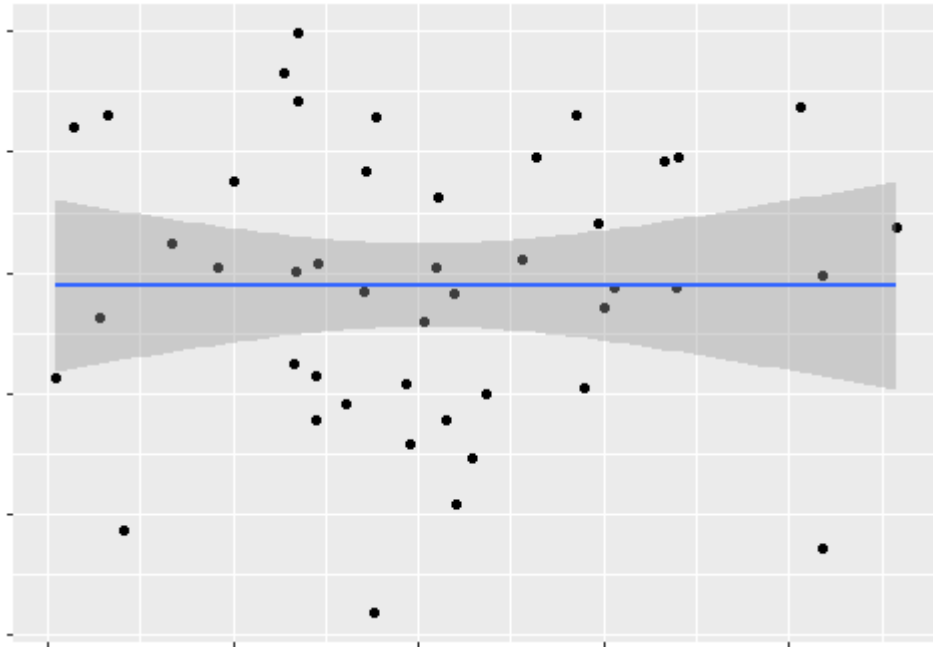
# Frequentist Hypotheses

**Null Model:** This is what we should see if there is no relationship between  $x$  and  $y$ .

**False Positives:** Sometimes the Null is true, but by chance we observe a pattern like the plot on the right.

**Alternative Model:** This is what we want to observe if there is a relationship between  $x$  and  $y$

**False Negatives:** Sometimes there is a true relationship, but by chance we observe a pattern like the plot on the left



# Modeling in a Frequentist world

## Frequentist modeling implements the Dual Model Paradigm

- Do you remember the dual model paradigm???

## Focus is on modeling process

- A particular parameterization of a model is just one of infinitely many possible realization of a **stochastic process** process.
- Model realizations are our best guess about the **true but unknowable** model parameters.

# Frequentist conceptual difficulties

**Frequentist confidence is based on hypothetical infinite repeated sampling.**

- Frequentist *confidence* and *significance* refer to repeating the process many times.

**A particular sampling/modeling process may or may not be a good approximation of the real population.**

- A particular CI either *does* or *does not* contain the true value. In the Frequentist world we can never know.

**These concepts are difficult.**

- One of the best ways to gain intuition is to explain to non-scientists.

# For Thursday

- Start Deck 3
- In-Class R: read the assignment description before class