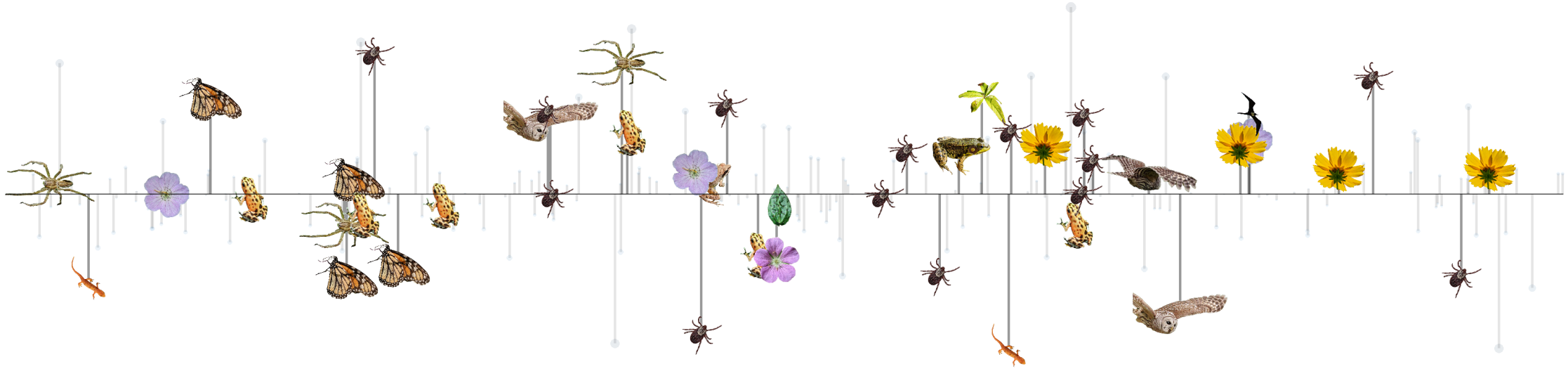
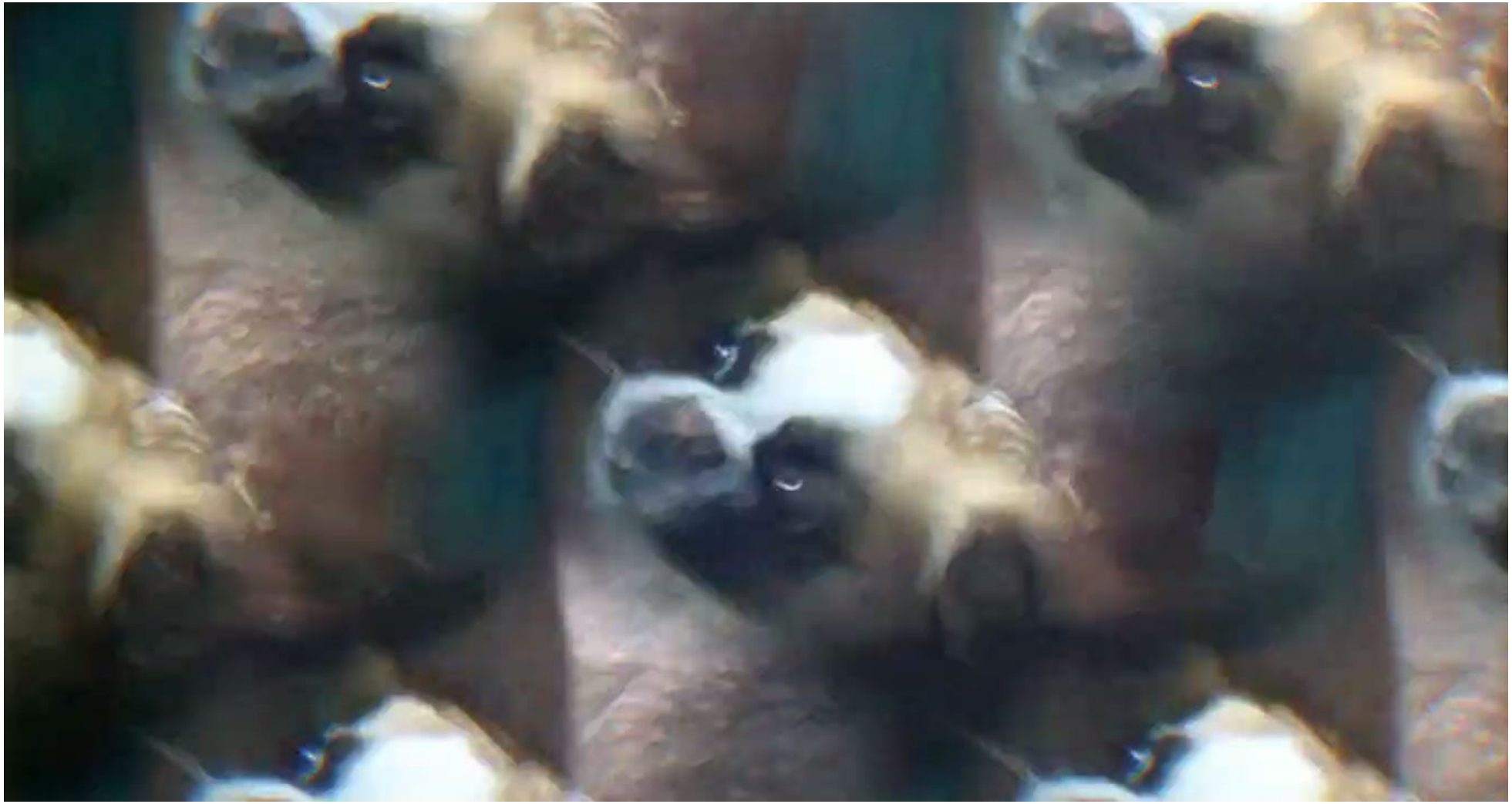
Intro to Quantitative Ecology
UMass Amherst – Michael France Nelson

Your Questions



Can you show us more photos of Betsy?!



NEON Questions

When do you think you will be able to get the feedback out for the NEON group project?

- All slides that were submitted as of early last evening have feedback!

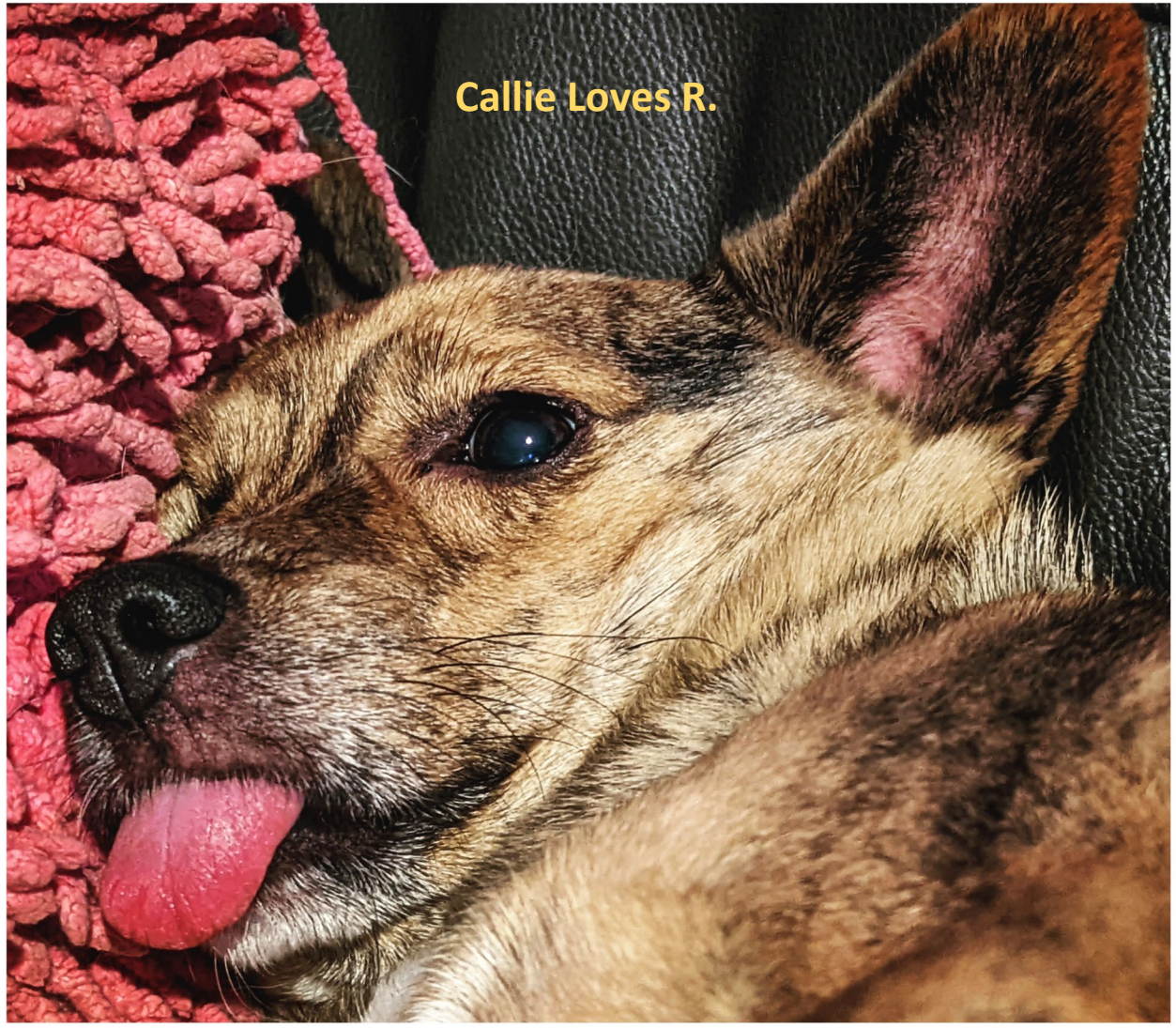
NEON Questions

- When are we presenting the NEON bird data? How are the presentations on the last day of class going to work out?
  - Last day of classes (next Tuesday).
  - We have too many groups (at least 12) for everybody to present. Presentations will be optional, but I'd love for about 4-5 groups to present their work!

Beyond This Course

What other relevant classes can we take that use R programming? What other courses are good to take in conjunction with this one if you want to learn more?

- Lots! Courses that I'm aware of in the ECO/NRC world:
  - Analysis of Environmental Data + Lab: ECO 602/634
  - Applied Ecological Statistics: ECO 636
  - Forest Measurements: NRC 534



Beyond This Course

Professor Lacey (RIP) especially likes Mixed Models.

- Landscape Ecology: ECO 621
- Data Visualization: ECO 690STB
- Advanced Statistical Ecology: ECO
- Spatial Data Analysis in R: ECO 637
- Special topics courses
- Independent study and practicum courses (watch out for the opportunities emails from Deb)
- Many of these involve data collection and analysis. You can practice your R skills in a real-life setting.

Beyond This Course

How much harder does R get (is what we've learned so far a good foundation)?

- This course is meant to give you the basics you need to understand R syntax, basic data visualization, standard frequentist (general linear model) analyses, and do basic data manipulation.
- The collection of basic syntax elements you need to know is surprisingly small.
- Archie is happy because he just learned how to import data from a csv flie!

Beyond This Course

He's feeling even better about R now!

Additional foundation concepts

Arithmetic operations, logical expressions, assigning variables

Importing data from files

Subsetting data frames (and similar data structures)

Difference between functions and variables

Knowledge of different data structures, and the functions that work with them.

Finding help: R help, google, peers.etc

Why is R so picky?

Why do we have to be so specific about commas and parentheses, isn't R smart enough to interpret what we mean?

- Nope!
- R is very literal. Our brains can 'get the gist' of poorly formed statements, but not R
- Human language is imprecise, ambiguous. Are these qualities we want in our analyses?

Neon Bird Data

Blue Group

Research Question

- Were different detection methods used at different observation distances?

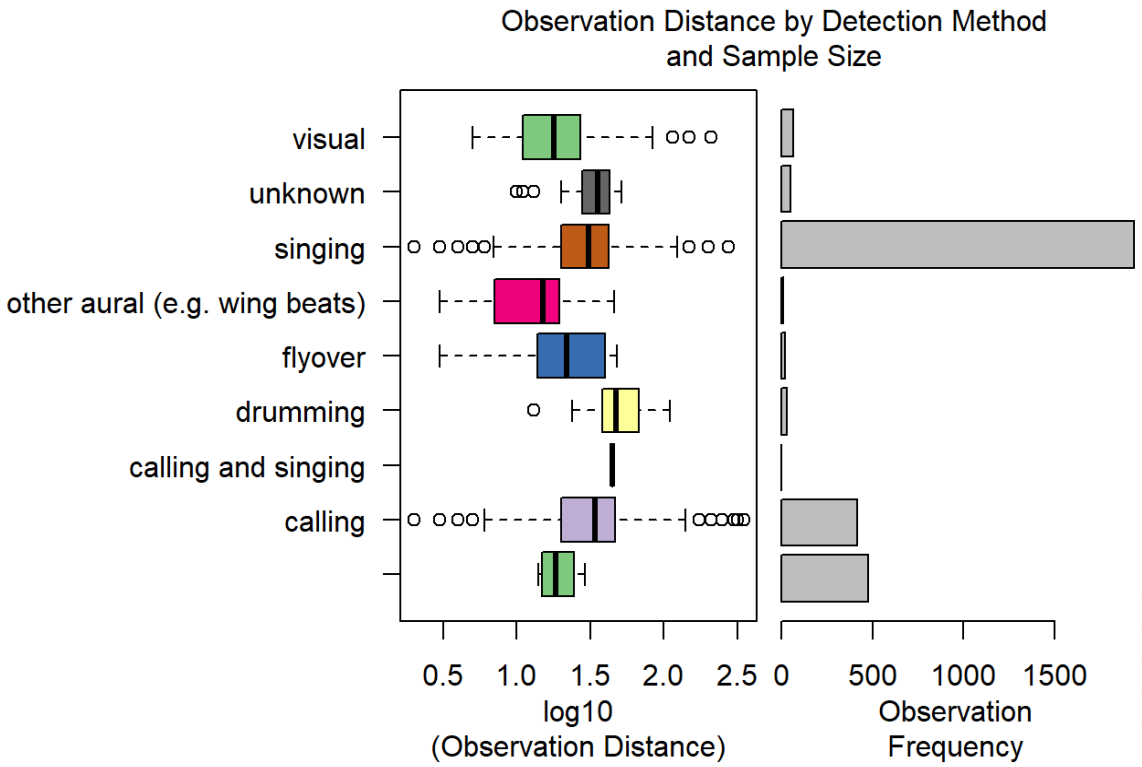
Null Hypothesis:

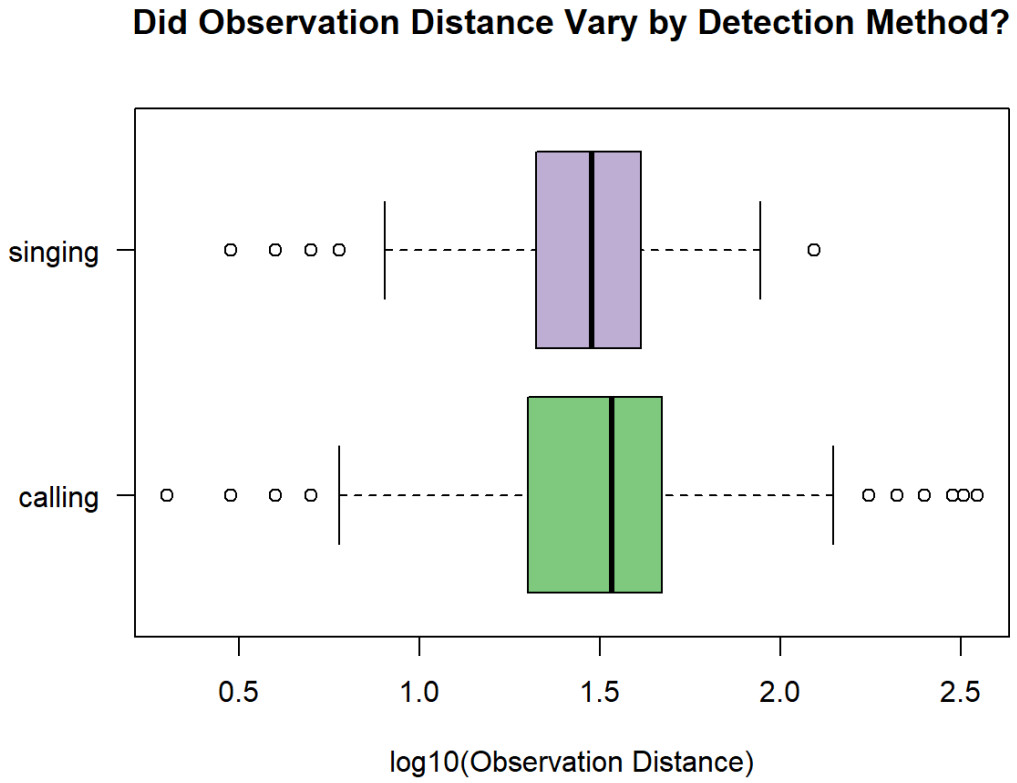
➨There is no difference in observation distance for each detection method.

Alternative Hypothesis:

➨There is a difference in observation distance between some detection methods.

Were different detection methods used at different observation distances?






```
##
##  Welch Two Sample t-test
##
## data:  log10(observerDistance) by detectionMethod
## t = 1.4555, df = 752.57, p-value = 0.1459
## alternative hypothesis: true difference in means between group calling and group singing is not
equal to 0
## 95 percent confidence interval:
##  -0.009640942  0.064932962
## sample estimates:
## mean in group calling mean in group singing
##              1.486771              1.459125
```

Miscellanea

- Can you show us more photos of Betsy?! Yesssss!
- How long have you been teaching this course? 4 years

What other kinds of functions can R do that are beyond the scope of this course?

Custom Functions

Can't find a function that does what you need?

You can write your own R functions!

```
cube_root = function(x)
{
  return(x ^ (1/3))
}
```

```
> betsy = 27
> cube_root(betsy)
[1] 3
```

Functions from Packages

There are thousands of R packages out there, each with their own set of functions.

Fortunately, you know the basic syntax:

- Invoke a function using its name and a set of parentheses.
- Each function accepts zero or more arguments.
- Check the help entry for the list of required arguments.,

RMarkdown

How fancy can you get with .Rmd knitted visuals?

- Quite fancy!
- You can write raw html code (if you know how)
- You can use CSS
- You can use LaTex for math expressions
- There are lots of functions and packages for formatting tables.
  - These are not simple, consider the effort vs. payoff tradeoff before obsessing about formatting the perfect table.

RMarkdown

How do you change the knit outcome besides just changing it on the top bar?

- You can adjust options in the YAML header. Some helpful links are:
  - RMarkdown Crash Course – YAML Headers
  - RMarkdown section of R for Data Science

```
---
title: |
 Introduction to Quantitative Ecology at University of Massachusetts, Amherst
 ![](`r here::here("docs", "images", "banner_all_4.png")`){width=100wv}
author: "[Michael France Nelson](https://michaelfrancenelson.github.io/){target='_blank'}"
date: "Spring 2023"
output:
  html_document:
    theme: united
    css: !expr here::here("data", "formatting", "styles_mfnelson.css")
    toc: TRUE
    toc_float: TRUE
    includes:
      in_header: !expr here::here("data", "formatting", "base_target_blank.html")
editor_options:
  chunk_output_type: console
---
```

RMarkdown

How do you change the knit outcome besides just changing it on the top bar?

- You can manually call the render() function. You can specify lots of custom options via the many arguments. This is the approach I use for rendering webpages for this course:

```
if (FALSE)
{
  require(here)
  require(rmd.utils)
  rmarkdown::render(
    input = here::here("2023_spring", "index_2023.Rmd"),
    output_file = here::here("docs", "index")
  )
}
```

Plot Questions: Data Dimensionality

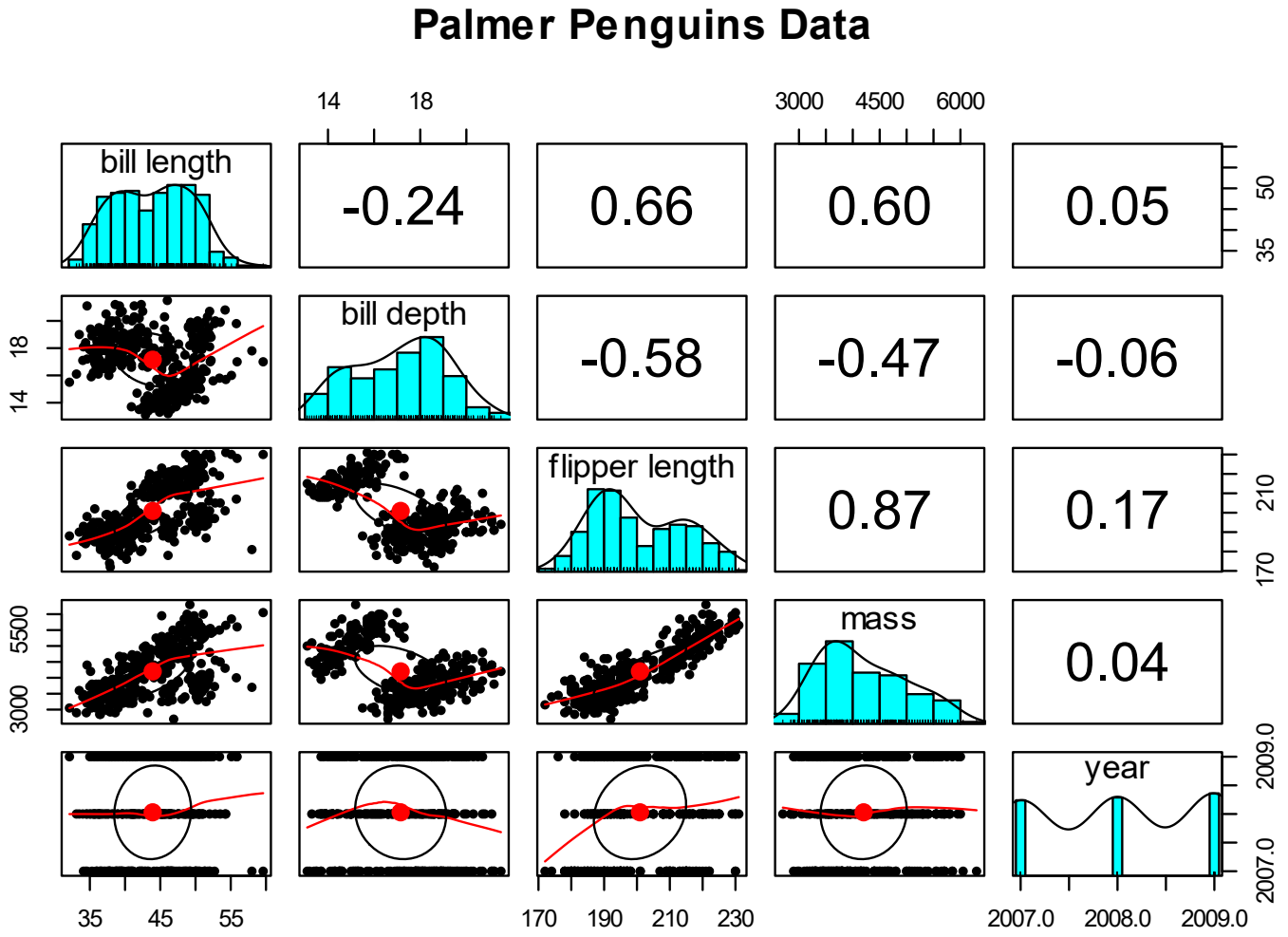
Is there a way to plot more than three variables in a scatter plot in R?

That's a hard question, and each approach has tradeoffs.

- Symbolize using multiple graphical attributes:
  - Size, shape, color, transparency.
  - The ggplot paradigm is very helpful for this approach
- Coplots
- Pairplots
- 'slices' through 3D data space.

Pairplot

Copper loves a good pairplot!



2D Slices in 3D Space

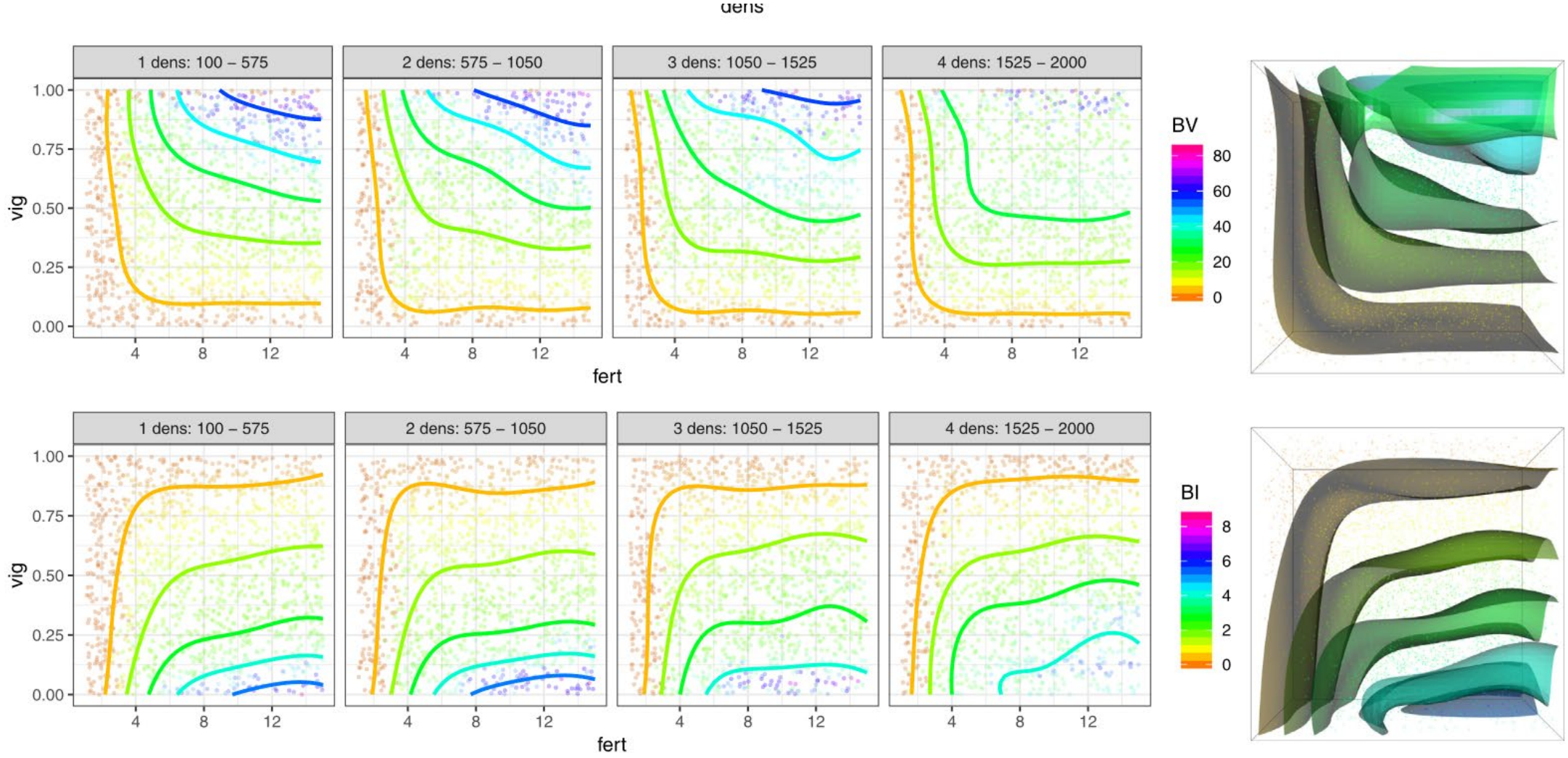


Fig 4 from Nelson et al. 2018

Plot Questions: Data Dimensionality

Is there a way to plot more than three variables in a scatter plot in R?

- 3D scatterplots.
  - These are ok, but can be very difficult to interpret especially if there are many points.
  - My hypothesis: The whole image is in focus – we can't use our plane of focus to help us. Is this why most dogs ignore TV?
  - Interactive 3D plots are much better.
    - Surface plot example
    - 3D spider plot
    - Check out the rgl package

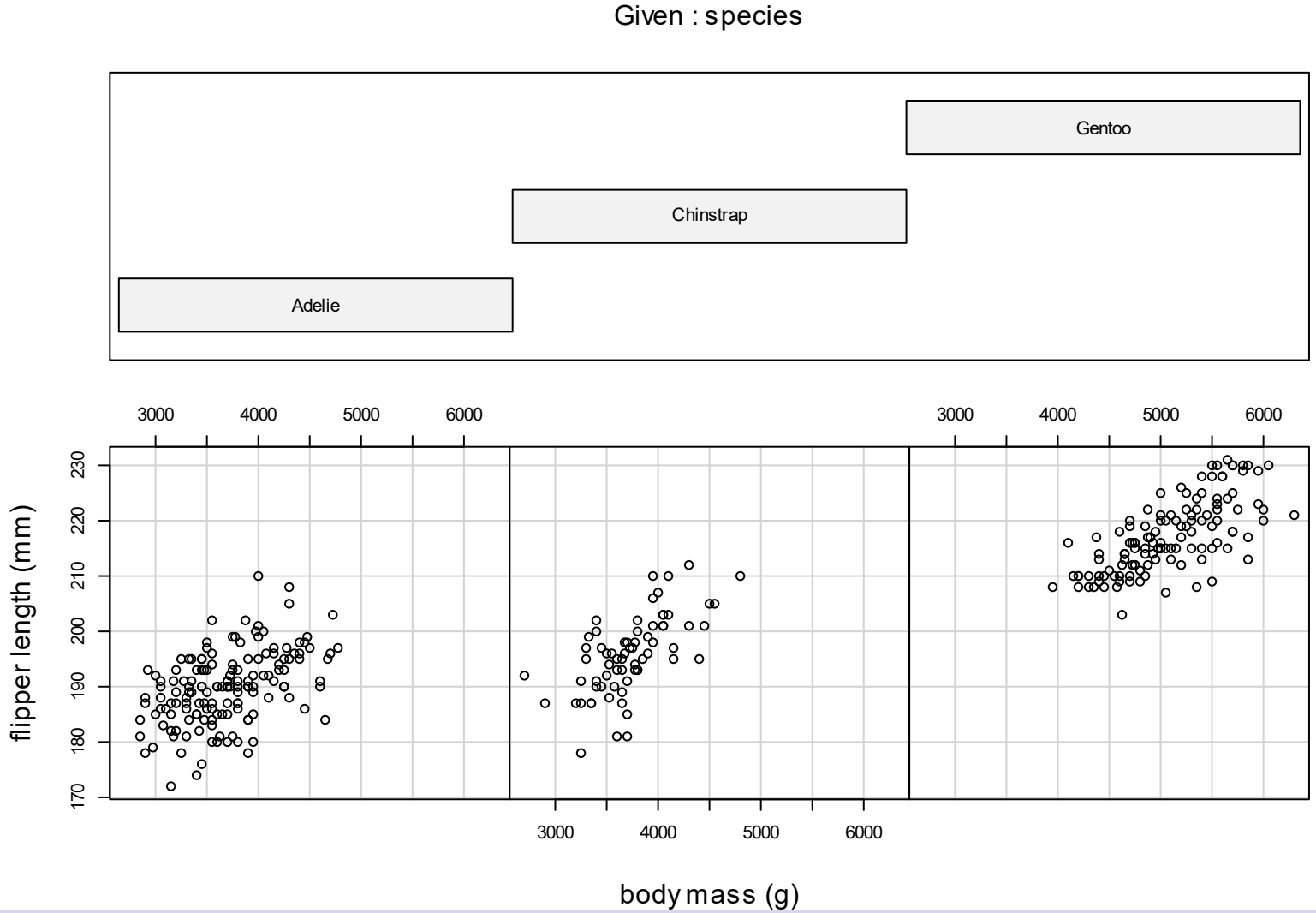
Visualizing 3D Data: Coplots

Coplot with a categorical variable

Each slice is a penguin species:

- **Adelie**
- **Chinstrap**
- **Gentoo**

What can you see?



Plot Customization: axes

How can you remove underscores from axis tick labels?

- Your best bet in base plotting is to create custom axes. The procedure depends on the type of plot.
- axis() function works for scatterplots and histograms (and possibly others?).
  - First suppress the plotting of axes in your main call to plot(), then create a vector of the desired labels.
- You can use text manipulation functions like gsub(), paste(), etc.
  - These take some practice.
- You can create a column of factor level names with the desired formatting in your original data frame. This works great for barplots and boxplots. This often requires the least effort for boxplots and barplots.
- Consult Dr. Google.

Plotting Statistical Tests – Data Exploration is Key

Can any type of test be visualized, or can it only be certain types?

- Great question! The answer depends on what kind of data you have and what you want from your plot.
- A thorough graphical exploration will accomplish a lot.
  - Scatterplots can help show the presence and form of bivariate relationships.
  - Boxplots can help with ANOVA categorical predictors.
  - Pairplots are useful for multiple variables.

Plotting Statistical Tests – Model Validation Plots

Can any type of test be visualized, or can it only be certain types?

Sometimes you want to plot components of a fitted model.

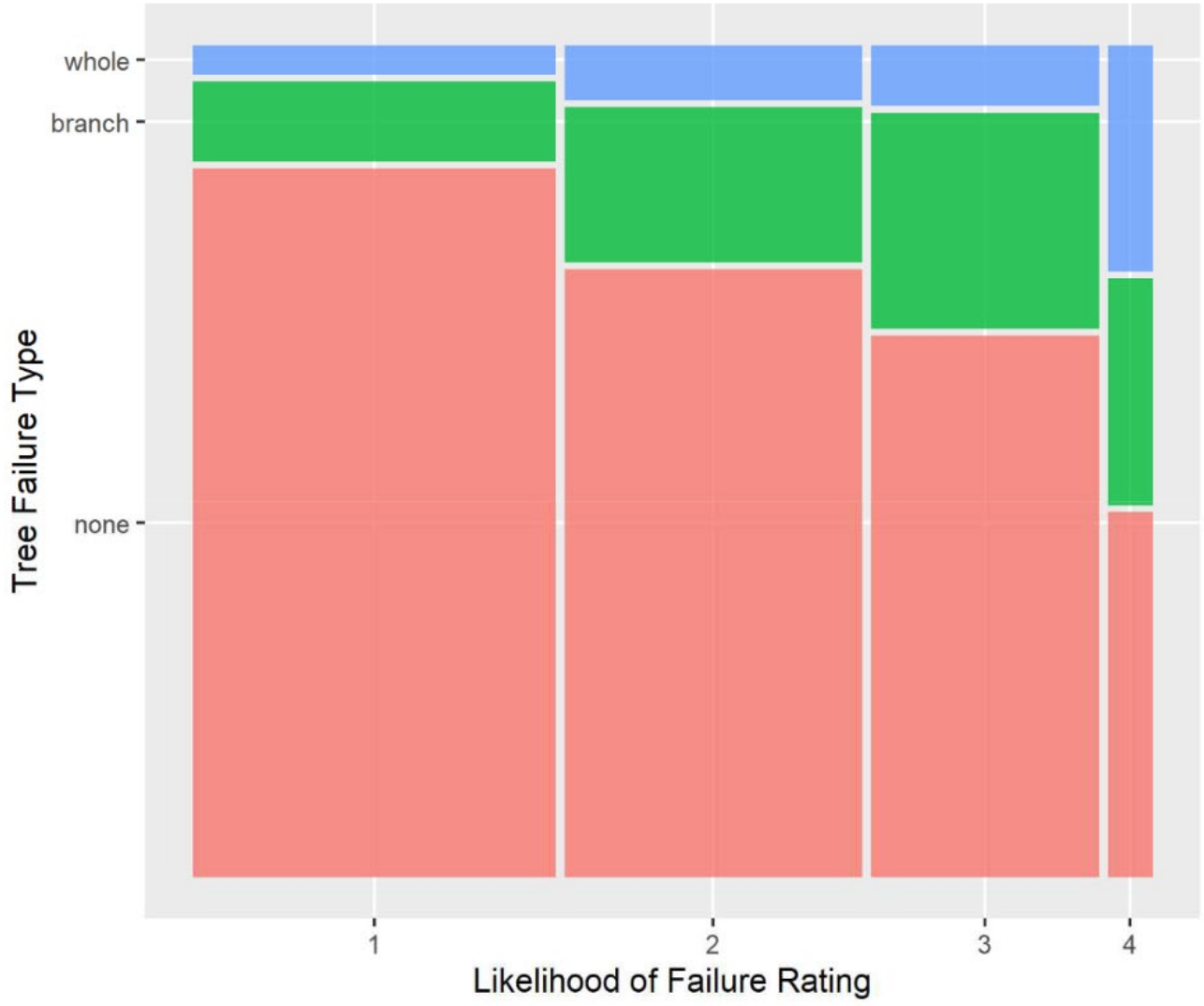
- Plots of residuals vs predicted values tell you a lot. For example, they can reveal unequal variance, AKA heterogeneity and non-independence. They can also indicate if your deterministic model is appropriate.
- Residuals vs. predictors: can reveal non-independence.
- QQ plots help visualize normality of the residuals.
- Leverage plots can identify influential points.

Plotting Statistical Tests – Mosaic Plots: proportion or chi-square tests.

Can any type of test be visualized, or can it only be certain types?

Mosaic plots help visualize associations

[Figure 2 from Nelson et al. 2022]



Plotting Statistical Tests

Can any type of test be visualized, or can it only be certain types?

A good graphical exploration will help you detect potential problems early in the process and can suggest appropriate models.

Way way way way way too often I see people (students and researchers of all levels) applying models to their data without even plotting it! Don't be that person.

Archie will judge you if you don't plot your data.

R in the real world

- What kinds of jobs are you most likely to use R in?
- What are some jobs that require the use of R?

Lots!

Which distribution?

Is there any research or guide that tells you what distribution you should run on a set of data?

- There are several things you need to consider.
- We've talked about numerous distributions in this class:
  - Poisson – good for counts
  - Bernoulli – good for presence/absence
  - Binomial – good for counts of successes in a fixed number of trials.
  - Normal – good for many continuous variables
- You need to first become familiar with the broad classes and properties of distributions. Discrete, continuous, bounded. This will help you understand properties of specific distributions.
  - I've given you a base understanding upon which you can build.

Which distribution?

Is there any research or guide that tells you what distribution you should run on a set of data?

- You can consult more advanced texts like Bolker and Zuur. These are tough reads, but they have good guides for what kinds of models and distributions work with different scenarios.
- Consult previous research in your field. You can often get a good idea of standard distributions used for work like yours. Beware, however, not all published research uses appropriate statistical methods!
- Schedule a QSG session!

Betsy likes the f-distribution. She thinks it stands for ‘food’.

Test statistics, and more!

What are the differences between F, P, and T values?

- F and t are test statistics. They are calculated from data.
- The p-value is calculated from the test statistic (and the distribution parameters).

What is the history of statistics?

Uffff, it's complicated and I'm hardly an expert!

- Lots of early work concerned probabilities. Blaise Pascal and Pierre de Fermat were interested in probability in the context of gambling. They began formalizing probability theory!
- The romanticized origin of Bayesian statistics is Thomas Bayes' famous theorem.
  - The statistical framework around it was mostly developed others later on.
  - Fun video that develops some intuition about Bayes' theorem and core philosophy of Bayesian stats.
    - Some of the banter gets tedious upon repeated watchings, but it's a great video.
- Gauss first solved the Euler-Poisson integral, which would later become the normal distribution.
- Ronald Fisher formalized lots of concepts in Frequentists statistics that we use today.
  - He was… well, it's complicated and not all good. He was a Eugenicist, which we now consider a form of white supremacy.

What is the history of statistics?

- Newer developments include:
  - Spatial statistics – dealing with spatial autocorrelation
  - Time-series modeling
  - Simulation methods: Bootstrapping, Monte-Carlo simulation, agent-based modeling, random walks, others.
- Huge theoretical and practical developments in Bayesian stats have been facilitated by increased computational power.

Figure 1. Callie is estimating the parameters for a binomial distribution and its conjugate beta prior distribution.

Correlation Coefficients

- Correlation coefficients help us characterize the coordinated variation between two continuous variables. They are standardized to fall in the range -1 to 1 for easy interpretation
  - Pearson correlation: assumes a linear relationship. Has other assumptions like bivariate normality, but we often ignore those. Calculation is based on covariances and variances, which use sums of squares.
  - Spearman correlation: calculation is rank-based. Sometimes called non-parametric. Only assumes a monotonic relationship
  - Correlations can only capture monotonic relationships between two variables.

Sample sizes and confidence

Why do error bars become longer at smaller sample sizes, and smaller at larger sample sizes?

- Let's talk this through!

How does sample size affect confidence intervals?

- Recall the sampling distribution. Standard error is adjusted by the square root of sample size: it decreases as n increases.

Degrees of Freedom

- Small sample sizes tend to underestimate uncertainty, the n-1 helps correct for that.
  - The n-1 is in the denominator, so it increase the value of the calculation.

Betsy is very uncertain how she feels about bathtime!

Unequal Group Sizes

What is the best way to account for datasets that are unequal when utilizing an ANOVA/ANCOVA?

- If possible, design your experiment with equal sampling. As long as you only have a few deaths, it's not a big problem!
- You can use different types of sums of squares in your ANOVA
  - An R-bloggers post about this
  - A more technical blurb
  - Check out the R package 'car'
- If your groups have approximately equal variance, then a regular ANOVA is OK.
- You are always safe to use a Kruskal-Wallace analysis.

Modeling

Can you explain the theoretical and coding differences between a stochastic model and mechanistic one?

- Remember the dual model paradigm? It has two parts, the deterministic and stochastic models.
- Deterministic models come in two flavors (really a continuum with these two extremes).

Mechanistic model: A model usually based on theory. Attempts to explain the mechanism by which a system works. For example, a predator-prey model based on holding time. Many mechanistic models have corresponding mathetmatical functions. For example the Michaelis Menten function for enzyme kinetics or the Ricker function for population growth or predator/prey relationships.

Phenomenological model: A model that attempts to fit a mathematical function to data. A Phenomenological model makes no claim about the underlying mechanism, it is purely meant to describe a pattern.

Modeling

Can you explain the theoretical and coding differences between a stochastic model and mechanistic one?

- Remember the dual model paradigm? It has two parts, the deterministic and stochastic models.
- **Stochastic models** describe the randomness in the system. Probability distributions are often used for the stochastic model.

Modeling

Can you explain the theoretical and coding differences between a stochastic

The deterministic model is often coded in the model formula in R

```
lm(bill_length_mm ~ body_mass_g, data = penguins)
```

For lm() the normal distribution is the stochastic model, you don't need to explicitly code it.

For other types of models, you might need to specify both the forms of the deterministic and stochastic models:

```
glm(counts ~ outcome + treatment, family = poisson())
```

Significance and p-values

At what point is a value considered “close to” a value (like a p-value) vs. very far away? For example, would 0.02 be considered a small or large deviation from 0.05?

- That’s not an easy question. We like it when our p-values are much less than 0.05. If they are close, we have to think about the value of an arbitrary cutoff in the first place.
- If your p-value is less than 0.05, you can safely characterize your result as significant. If it is above 0.02, I’d usually call it ‘marginally significant’. As long as you acknowledge that your p-values were close to 0.05, then you can rest with a clean conscience.

Correlations and Associations

What are the key differences between correlation and association?

- Correlation is a measure of coordinated variation between two numeric (preferably continuous) variables.
  - Is there a positive relationship between biomass and nitrogen addition?
  - Correlations can be visualized in a scatterplot.
- Association is a measure of how often the possible combinations of the levels of two factor variables co-occur.
  - Do owls tend to occur near forest edges more often than expected by chance?
  - Associations can be represented by counts in a contingency table. You can visualize them with a mosaic plot.

Categorical Data

What does ordinal and nominal mean again?

- Both refer to categorical data.
- Original has categories with an intrinsic quantitative ordering. For example, size or age classes.
- Nominal has categories with no intrinsic ordering. For example brand names of paper towels, or species of lizard.

What is the best way to examine relationships between nonnumerical data?

- Contingency tables and chi-square tests are good for categorical data.
- Logistic regression models can handle presence/absence.
- Multinomial logistic regression can handle ordinal and nominal data.
  - But the interpretation is not easy.

That’s all folks!