Analysis of Environmental Data

Frameworks: Least Squares, Likelihood, Resampling

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst
Michael France Nelson

Least Squares

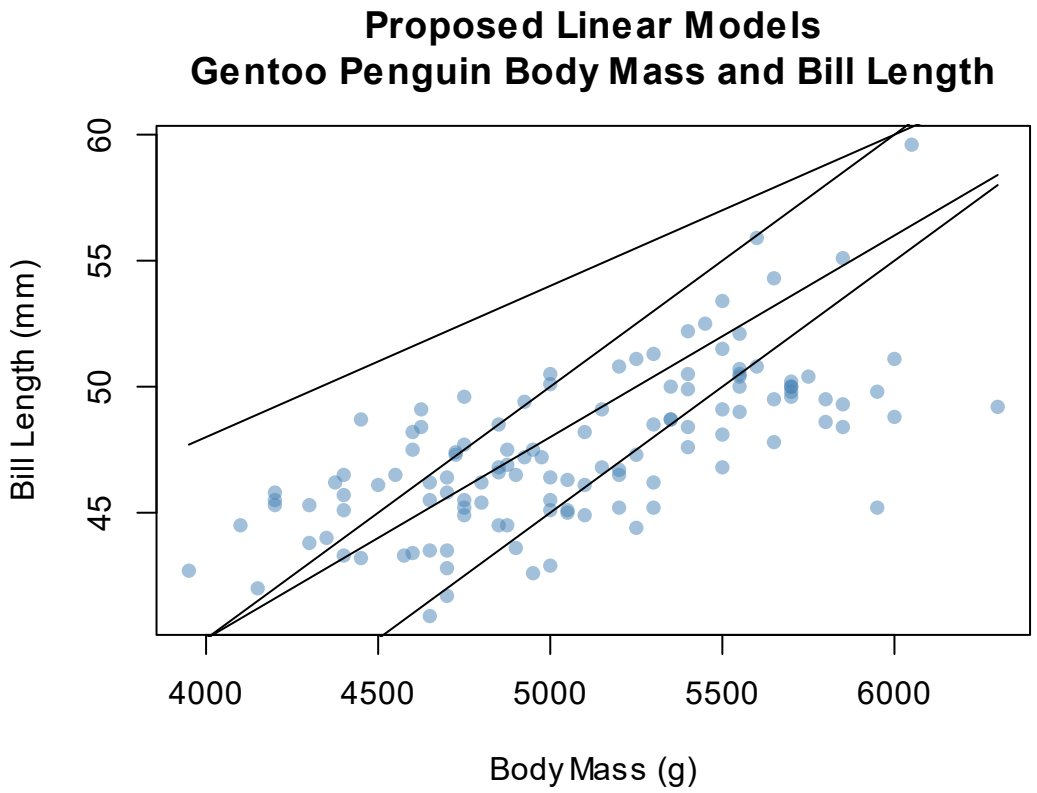
A Parametric Frequentist Approach

What's in This Section?

Take home concepts	A little floating house
• Dual model paradigm • Optimization criterion • Parametric and nonparametric inference • Least squares and likelihood	

Linear Model - Basics

Linear Function of Data



Linear Model Equation

- $Bill\ length = intercept + slope * (body\ mass)$

$$y = \alpha + \beta x + \epsilon$$

McGarigal's organization of the materials

McGarigal separated the OLS and Likelihood materials

- OLS is an optimization paradigm: minimize the sum of squared deviations (residuals).
- Likelihood is an optimization paradigm: choose model parameters that maximize the likelihood of the observed data.

For general linear models: they are equivalent. This is not necessarily the case for other model types.

$$y_0 = \beta_0 + \beta_1 x_1 + \epsilon_1$$

Dual-modeling approach: Linear Model

Linear model formula

$$y_0 = \beta_0 + \beta_1 x_1 + \epsilon_1$$

Can be decomposed:

Deterministic linear model: $y_0 = \beta_0 + \beta_1 x_1$

Stochastic model of the errors: ϵ_i

Dual-model: Parametric and Non-Parametric

Deterministic and stochastic

- Linear model: $y_0 = \beta_0 + \beta_1 x_1$
- Stochastic model, errors: ϵ_i

Deterministic Model: parametric and nonparametric are equivalent

Stochastic Model:

- Nonparametric modeling: The errors, ϵ_i, exist
- Parametric modeling: The errors, ϵ_i, exist AND we propose they are normally-distributed.

Deterministic Model

- For both types (parametric and non-parametric) the deterministic model is the same:

$$y_0 = \beta_0 + \beta_1 x_1 + \epsilon_1$$

  - We propose that linear function describes the deterministic component of the system.
  - NOTE: we could propose other deterministic models as well

Non-Parametric Inference: Stochastic Model

We make no claim about how the data are distributed in the population

- We propose that the stochastic model is: ϵ_1. We do not propose a particular distribution.

Parametric Inference: Stochastic Model

We propose that the stochastic component, ϵ_i of the system can be adequately described by a parametric distribution.

- We propose the stochastic model is ϵ_1, and that it is described by **a parametric distribution**. We frequently use a Normal distribution.

We want to optimize the fit of our deterministic model

- Since we've proposed that a *parametric* distribution describes the stochastic model, we can find model parameters that make our model the most likely.

Parametric + Non-Parametric Optimization

We want to optimize the fit of our deterministic model for both types, but to do so we need to the noise…

- **Non-Parametric:** what are our options?

- **Parametric**: we are willing to assume a distribution, what are our options?

Parametric + Non-Parametric Optimization

We want to optimize the fit of our deterministic model for both types, but to do so we need to the noise…

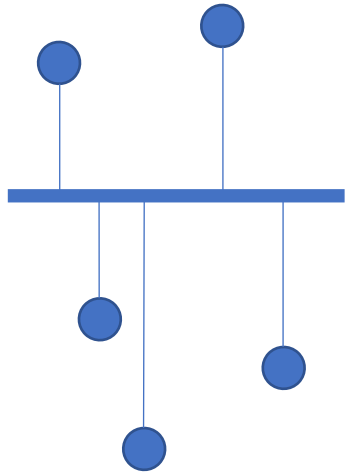
- **Non-Parametric:** what are our options?
  - Least Squares as our optimization criterion! It doesn't assume any particular distribution.
  - Simulation?
- **Parametric**: we are willing to assume a distribution, what are our options?
  - Least squares or simulation? No distributional assumptions
  - Likelihood methods: we assume an error distribution

Simple Optimization: the Mean

The mean is a statistic that describes the *center* of data.

What properties would we like a measure of center to have?

- **Lie between the minimum and maximum?**
- Have equal counts of values above and below?
- Minimize the sum of residuals?

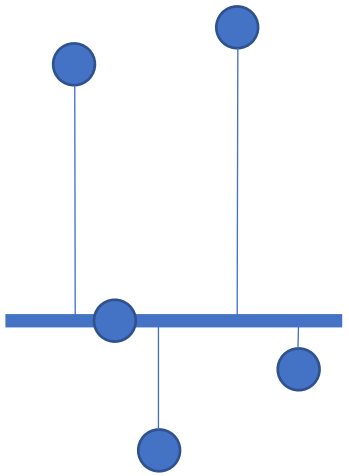


Simple Optimization: the Mean

The mean is a statistic that describes the *center* of data.

What properties would we like a measure of center to have?

- Lie between the minimum and maximum?
- **Have equal counts of values above and below?**
- Minimize the sum of residuals?

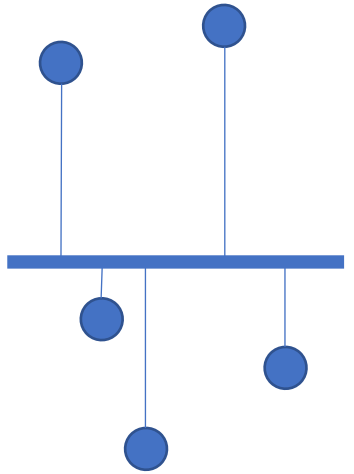


Simple Optimization: the Mean

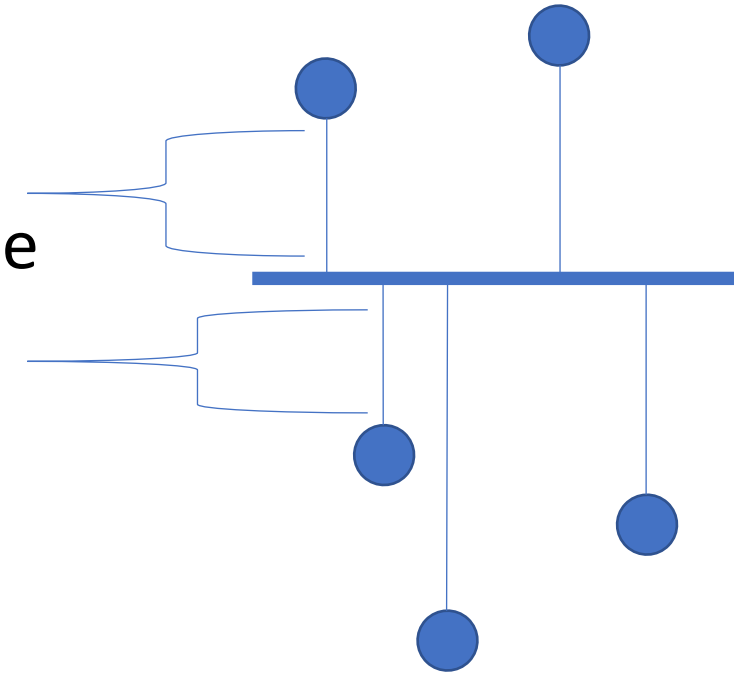
The mean is a statistic that describes the *center* of data.

What properties would we like a measure of center to have?

- Lie between the minimum and maximum?
- Have equal counts of values above and below?
- **Minimize the sum of residuals?**



Least Squares

The Idea	Optimization criterion
• Minimize the squared differences between observed and expected values. • Minimize the squared residuals, a.k.a. errors	• Squared difference between mean and observed. Square these distances 

Simple Optimization: the Mean

The mean optimizes 2 criteria:

- **The sum of residuals is zero**
- The sum of squared residuals is minimized

Prove it to yourself with some data:

```
dat = rbeta(100, 0.3, 0.5)
mean_dat = mean(dat)
resids = dat - mean_dat
resids_2 = dat - (mean_dat + 0.01)
resids_3 = dat - (mean_dat - 0.01)
```

```
>    sum(resids)
[1] 2.63678e-15
>    sum(resids_2)
[1] -1
>    sum(resids_3)
[1] 1
```

Simple Optimization: the Mean

The mean optimizes 2 criteria:

- The sum of residuals is zero
- **The sum of squared residuals is minimized**

Prove it to yourself with some data:

```
dat = rbeta(100, 0.3, 0.5)
mean_dat = mean(dat)
resids = dat - mean_dat
resids_2 = dat - (mean_dat + 0.01)
resids_3 = dat - (mean_dat - 0.01)
```

```
>    sum(resids^2)
[1] 13.66388
>    sum(resids_2^2)
[1] 13.67388
>    sum(resids_3^2)
[1] 13.67388
```

OLS: Optimizing Regression Parameters

Ordinary Least Squares is a great way to find *optimal* parameters for the simple linear regression models:

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_i$$

- The *parameters* are β_0 and β_1.
- x_i is a predictor variable
- y_i is the response variable
- e_i is the error term

OLS: Optimizing Regression Parameters

NOTE: McGarigal separates OLS and Likelihood

- They are not *always* equivalent, but if you are willing to assume the errors ϵ_i are *normally distributed*, then OLS and Maximum Likelihood are equivalent for linear regressions.
- OLS doesn't work for logistic regression, for example.

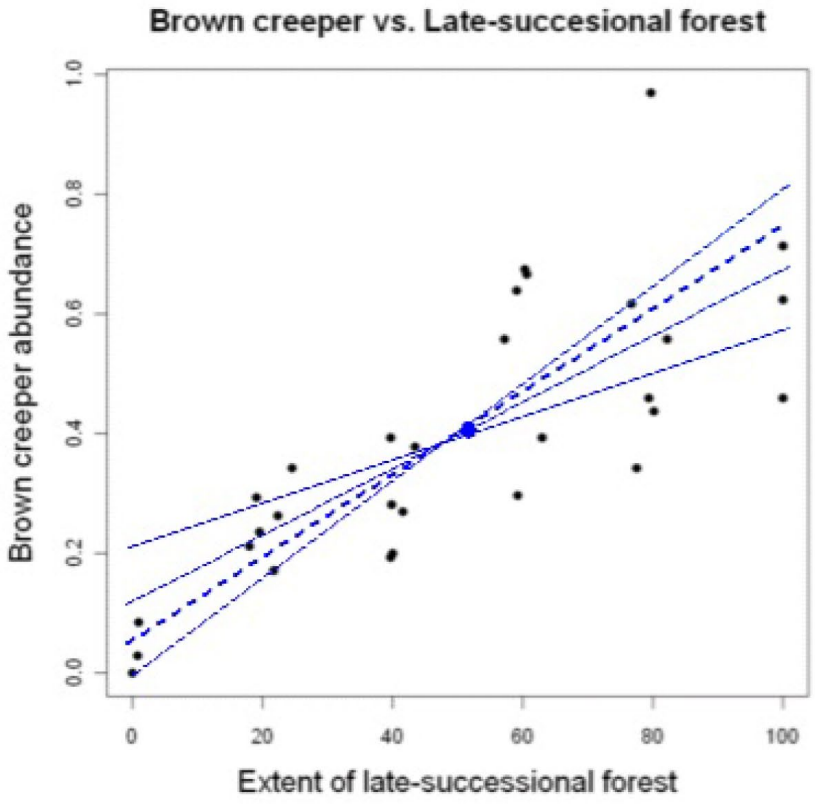
OLS: Optimizing Regression Parameters

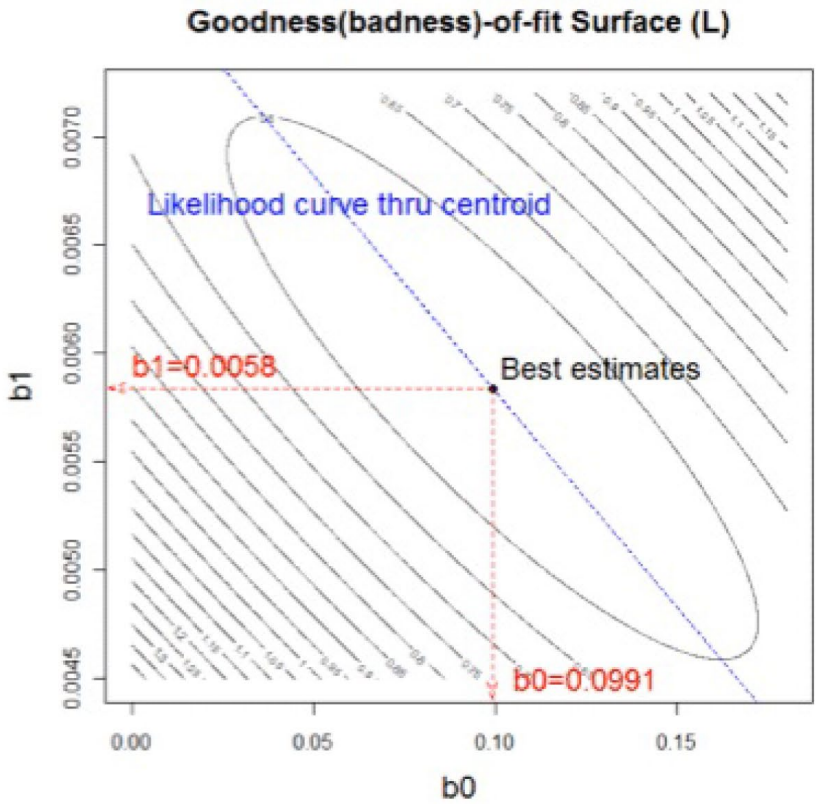
The optimization procedure:

1. Specify a model
2. Find model parameter values that minimize the sum of squared residuals via:
   - Numerical optimization
   - Analytical methods

Numerical Estimation

- Numerical solution





Numerical Estimation and Analytical Solutions

Numerical Methods: [educated] trial-and-error	**If we are lucky, a closed-form, analytical solution exists!**

Procedure for numerical methods:
• Specify starting values for parameters
• Search parameter space for optima
  • Algorithms, e.g. Newton's method
  • Simulation/resampling techniques, e.g. MCMC, gradient descent, machine learning techniques
	Exact solutions exist for OLS optimization for many regression techniques (like general linear models).
	• Solutions are based on techniques from Linear Algebra:
	  • Matrix multiplication, inversion, transpose, etc.
	  • Computers are really good at these operations.

Least Squares and Parametric Inference

We propose that a parametric distribution is appropriate for the stochastic model.

- Theoretical distributions are defined by density or mass functions.

- We want to optimize parameters in the context of the PDF/PMF.
  - We want parameters that make the observed data the most likely

For Normally-distributed errors: OLS methods also optimize for likelihood!

Recap Concepts

- Dual model paradigm
- Optimization criterion
- Parametric vs nonparametric stochastic models
- Least squares and likelihood

Likelihood

What is Likelihood?

- In a statistical inference context, likelihood is related to the probability of observing a specific event, x_i ,given a probability distribution and a set of parameters.
- We can use probability distribution functions to calculate the likelihood of specific events.
- The likelihood of an event is proportional to probability mass or density.

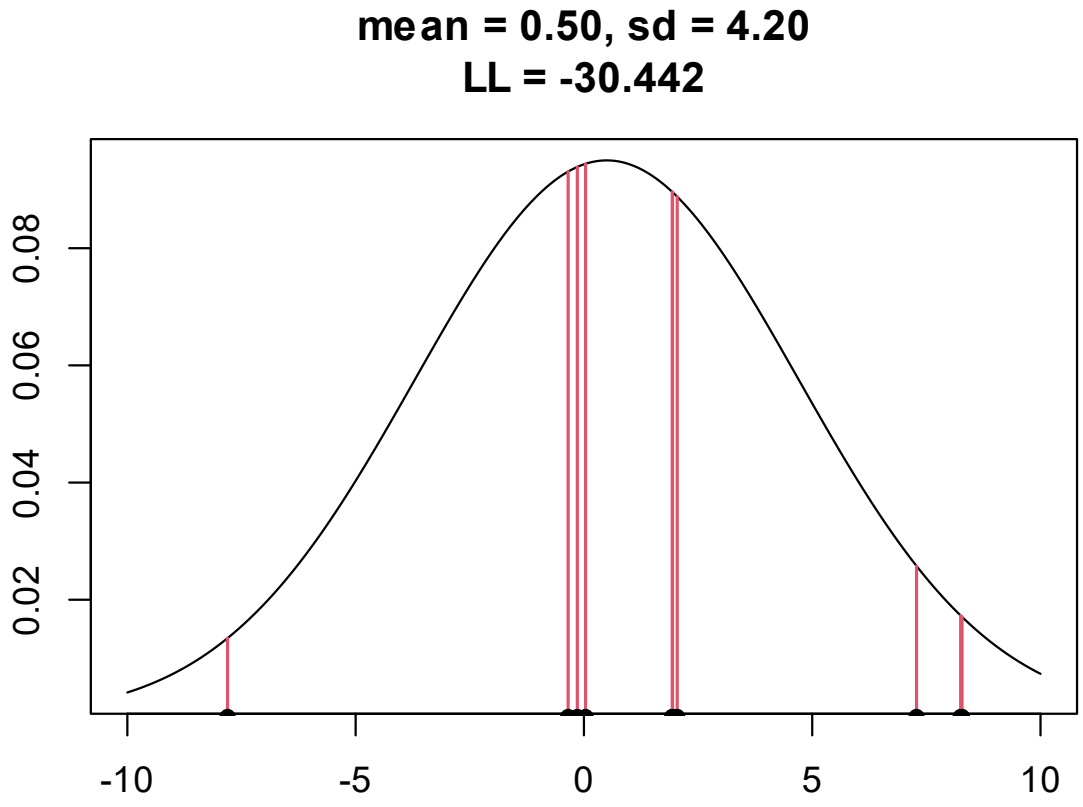
Likelihood of Data:

If we have more than one observation, the joint likelihood is the product of the probabilities of the individual events.

- Finally, we get to use the density function!

If the data were independently collected/observed.

- Remember independent events from probability theory?



Likelihood: The scenario

Main question: How likely am I to have observed the data I collected under my proposed model?

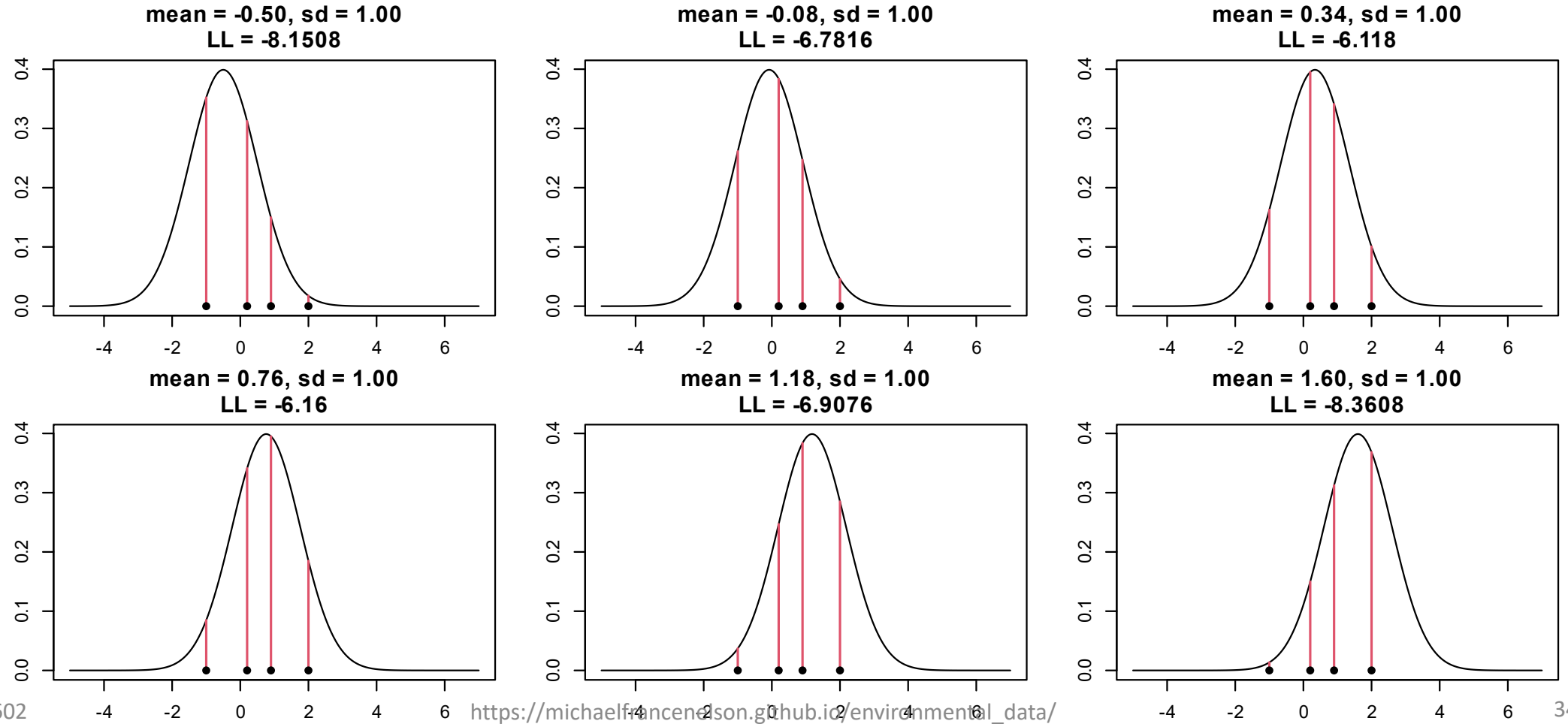
Likelihood can help if you have:

1. Data
2. A proposed a distribution or model of the data
3. A set of candidate distribution/model parameters

For inference: it might seem reasonable to try to find the parameter values that make our observed data most likely.

Maximum Likelihood

- Maximum likelihood method attempts to find the population parameters that make the observed data the most likely.



Likelihood: independent samples

Since you are a whiz at designing experiments, you know that all of your samples are independent!

- What do we already know about the joint probability of multiple, independent events?
- The joint probability of observing multiple independent events is the product of the probabilities of the individual events.
- Likelihood is an estimate of how probable your particular data are given a model and a set of model parameters.

The overall likelihood is proportional to the product of the likelihoods of each observation given your model/parameters!

Likelihood: data and model

How do we calculate the likelihood for **a specific event** or observation if we have a theoretical distribution?

- Use the height of a density/mass function.

How do we calculate the likelihood for **an entire sample** if we have a theoretical distribution?

- Multiply the density/mass of each observation.

Likelihood: procedure

1. Collect data
2. Propose model and candidate parameter values
3. Calculate the probability density of each observation given your model and parameter values:
   - From a theoretical distribution.
   - From an empirical/resampled/simulated distribution
4. Multiply the densities.
   - In practice we calculate the logarithm of the densities and add them together.
   - Why might this be better than multiplying probabilities?
5. Voilà: your likelihood value for your data Y given your proposed model and parameter values: Φ_m
   - In symbols $L(Y|\Phi_m)$

Likelihood calculations

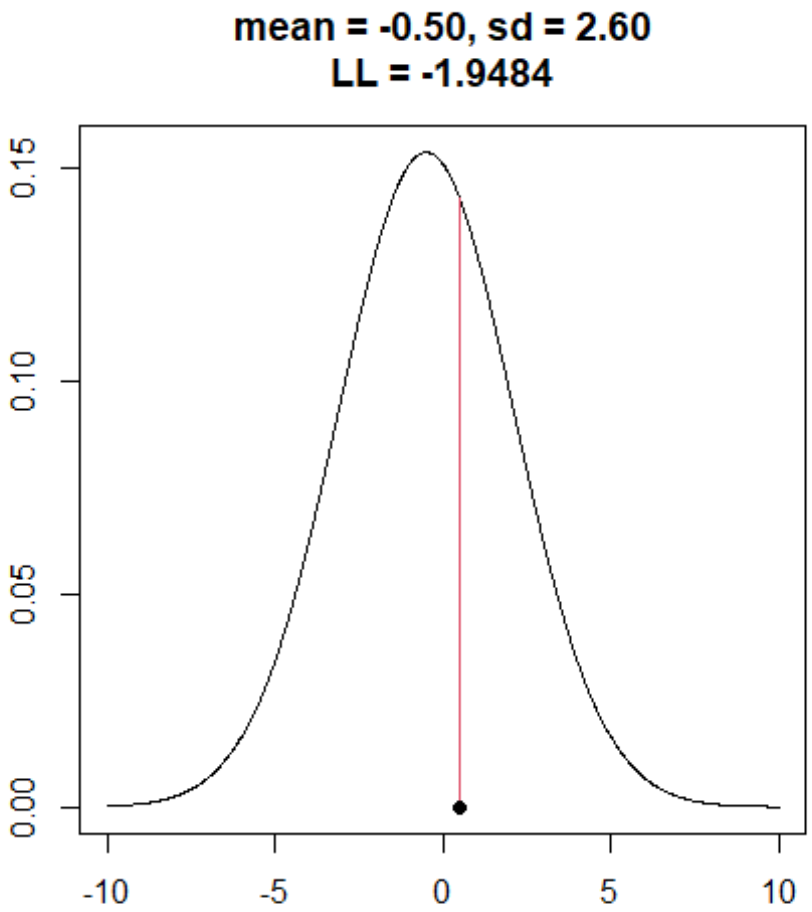
In Maximum Likelihood inference we want to maximize the [log] likelihood of the parameters.

Wouldn't it be nice if we had a simple formula?

- Sometimes we can find a formula and then find it's minima/maxima via calculus.
- Frequently such formulas don't exist.

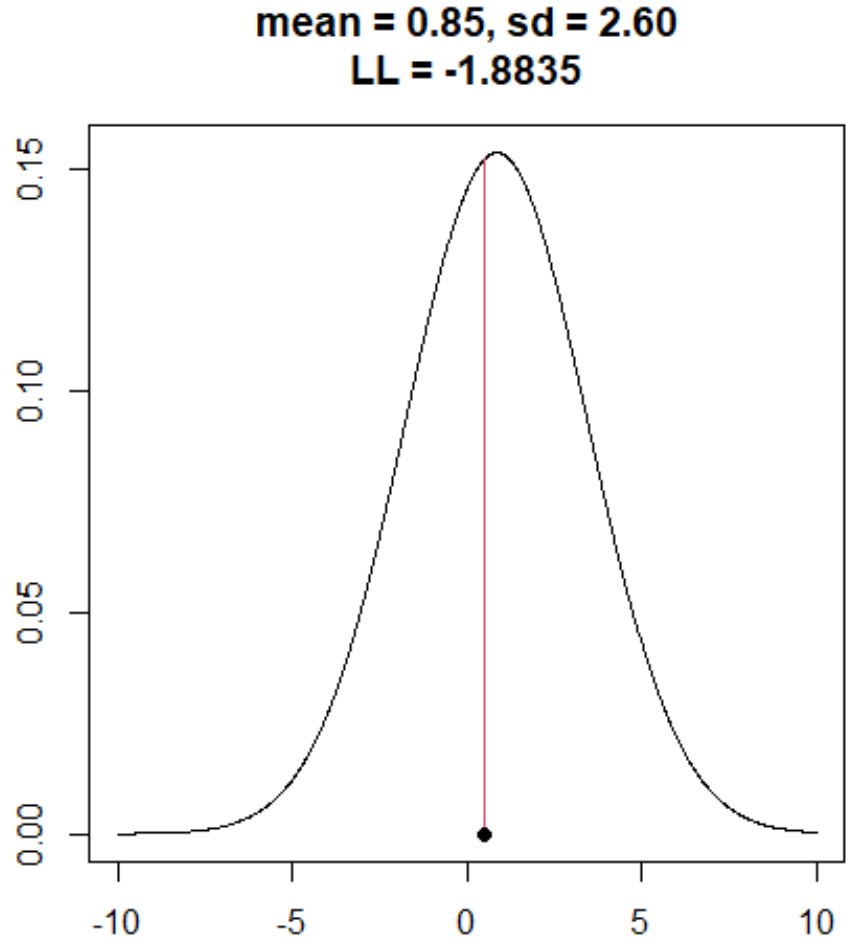
Likelihood Example Calculations

- one point: $x = 0.5$
- Normal distribution, test values:
  - $\mu = -0.5$
  - $\sigma = 2.6$



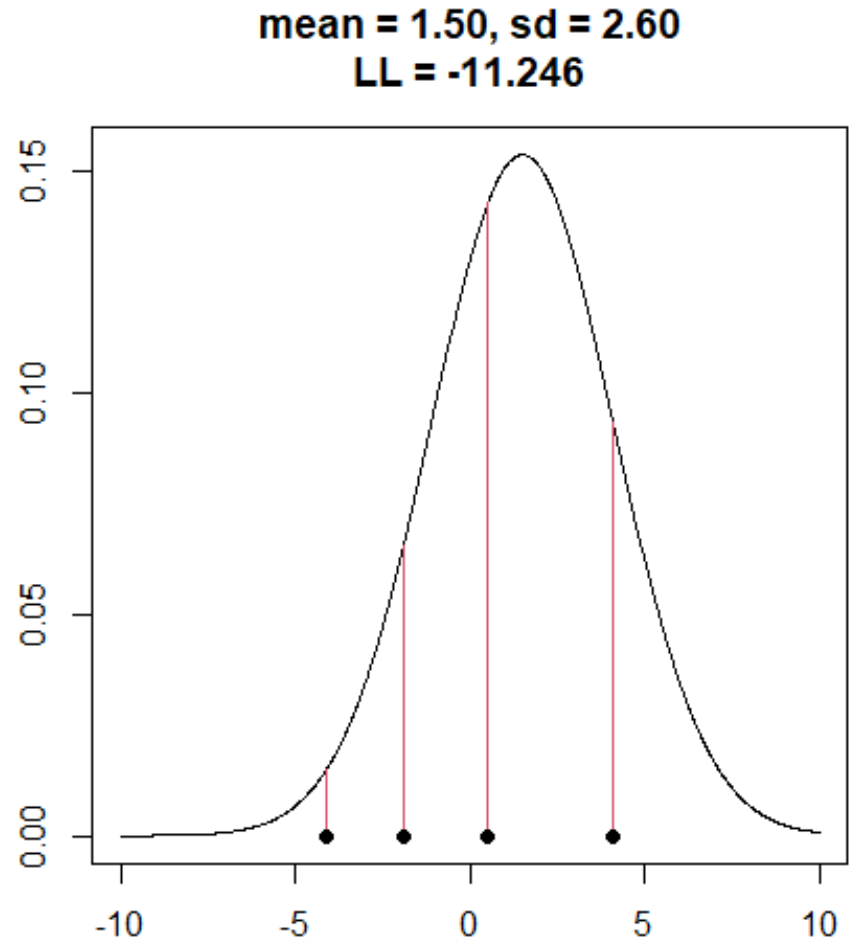
Likelihood Example Calculations

- one point: $x = 0.5$
- Normal distribution, test values:
  - $\mu = .85$
  - $\sigma = 2.6$



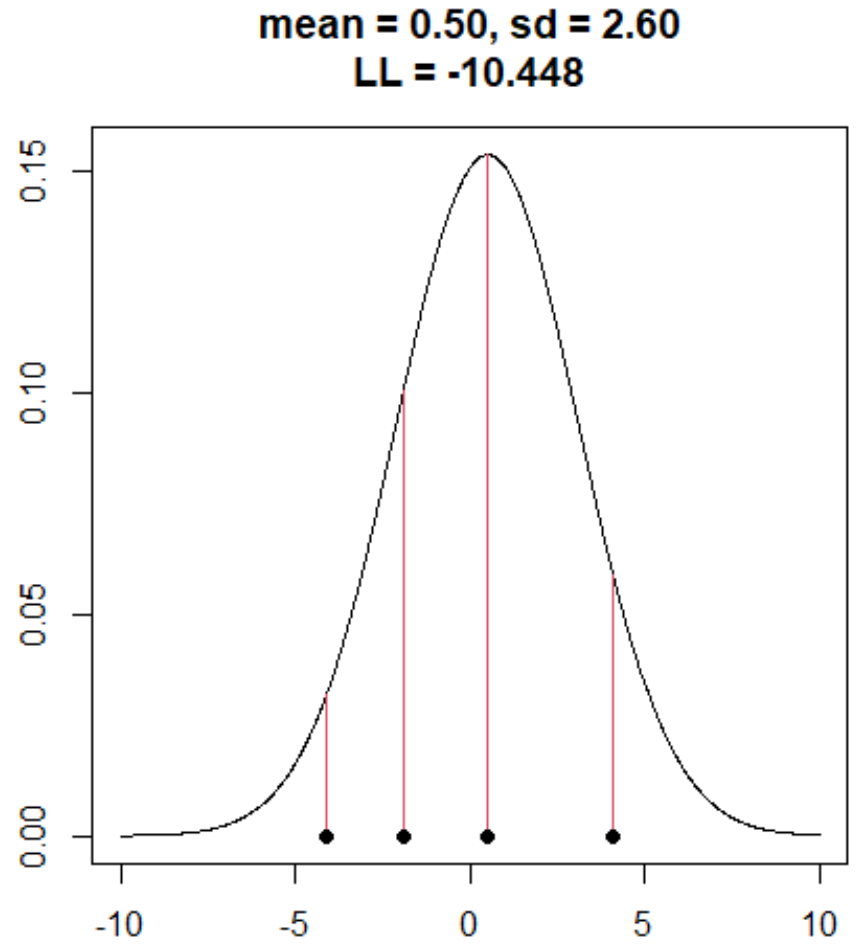
Likelihood Example Calculations

- Multiple points
- Normal distribution: $\mu = 1.5$, $\sigma = 2.6$



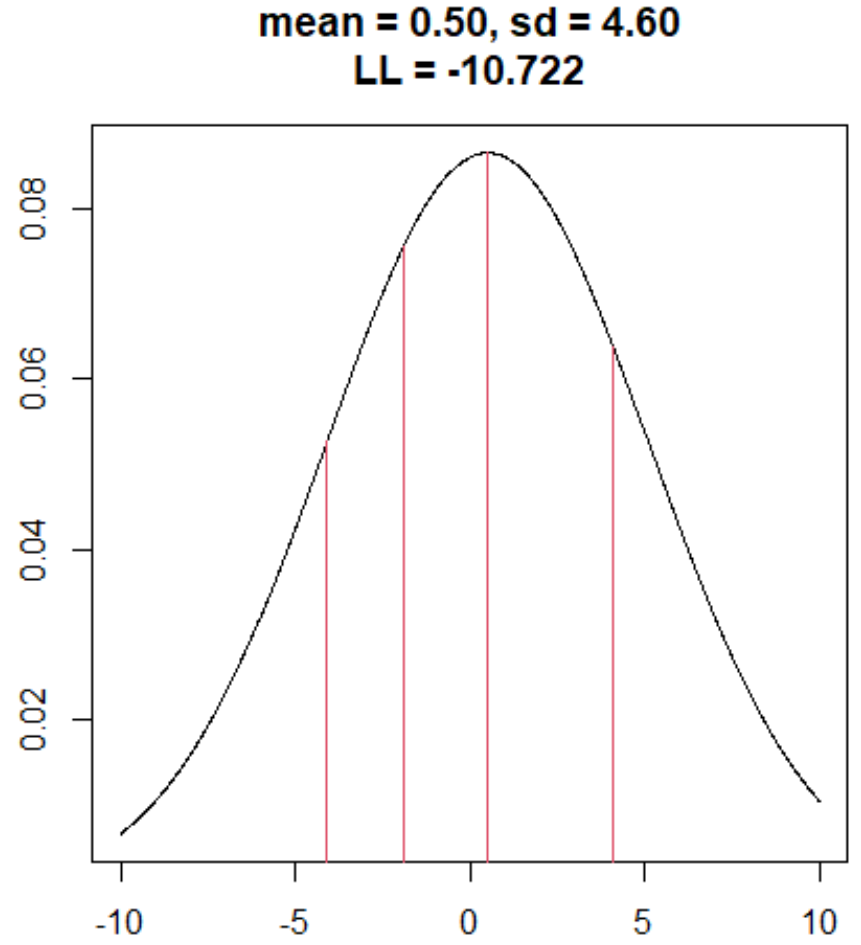
Likelihood Example Calculations

- Multiple points
- Normal distribution: $\mu = 0.5$, $\sigma = 2.6$



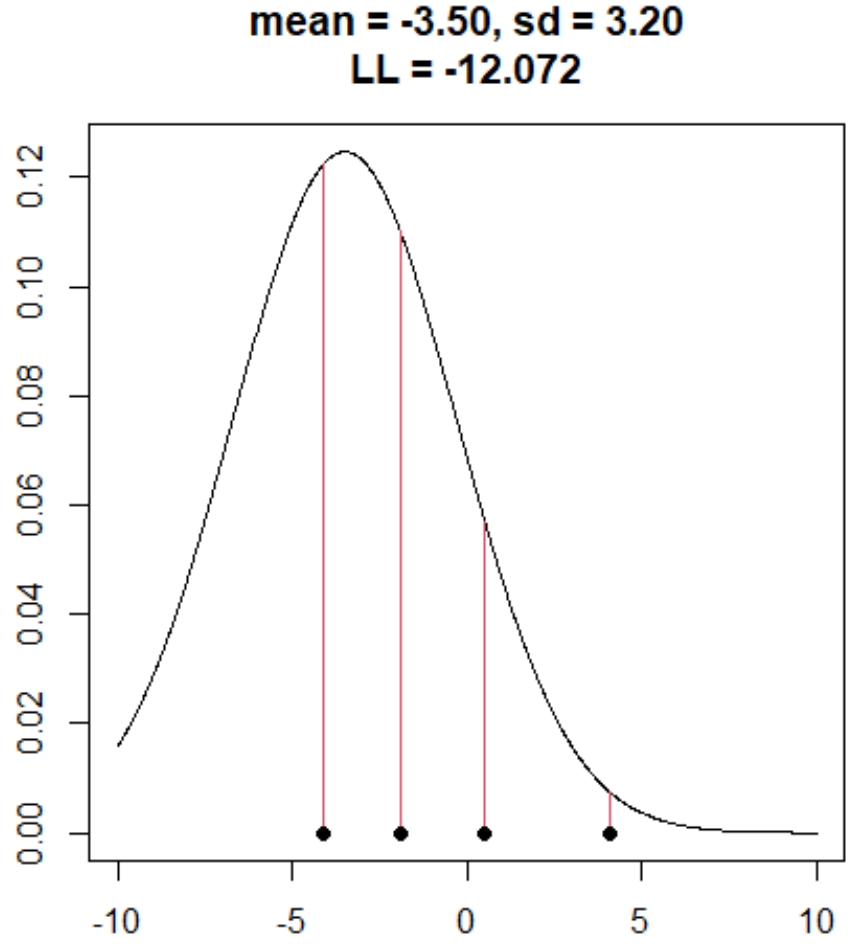
Likelihood Example Calculations

- Multiple points
- Normal distribution: $\mu = 0.5$, $\sigma = 4.6$



Likelihood Example Calculations

- Multiple points
- Normal distribution: $\mu = -3.5$, $\sigma = 3.2$



Key Concepts

- What is likelihood?
- How do we calculate it?
- What is maximum likelihood?
- Numerical and Analytical methods

Resampling

Bootstrap and Monte Carlo Methods

What's in This Section?

Slides	Take-Home Concepts
• What is resampling? • Sampling with replacement. • Bootstrap and Monte Carlo randomization • Resampling the null hypothesis • Resampling the alternative hypothesis	• Resampling for the null (Monte Carlo) and alternative (Bootstrapping) hypotheses. • Why sample with replacement? • When is resampling useful? • Breaking vs. retaining associations in the data. • Labeled data.

What is resampling?

Resampling methods create new samples from our existing sample.

1. Resampling *with replacement* allows us to create many "new" data sets from our original that we can analyze.
2. It sounds like cheating, but….
    1. Remember our random sampling scheme?
    2. Nonparametric inference can't help us if we use a poor sampling design.

What is resampling good for?

1. Nonparametric inference
2. Null and alternative hypotheses
    1. Monte Carlo randomization helps us characterize the null hypothesis.
    2. Bootstrapping is like the alternative hypothesis.
3. Confidence intervals
    1. Especially helpful when we don't want to claim the population follows a theoretical distribution

Resampling x and y

1. Bootstrapping: samples entire rows of the data, preserves structure
   - Keeps x_i and y_i together.
   - Keeps all of the attributes of a *sampling unit* together
   - Preserves associations among data columns (if they exist!)
2. Monte Carlo resampling: Sample predictor/response variables separately.
   - Samples each column of the data separately.
   - Can pair different x, and y values: e.g. x_2 and y_{53}.
   - Jumbles attributes among *sampling units*.
   - Destroys the associations among columns, removes the structure from data

Bootstrap Resampling

Simple concept: randomly select **entire rows** of data *with replacement* from the original data set.

The new data sets may have some repeated observations, and some may be left out.

- There are many possible resamplings of the data.
  - Many will resemble the original data.
  - Due to sampling error, some resamplings will be very different than the original.
    - Imagine rolling a 6 on a die 20 times out of 25.

Bootstrapping uses

Estimate standard errors and sampling distributions.

- Remember SE is a parameter of the sampling distribution of a statistic, not the population distribution.

Estimate confidence intervals.

- Helpful with small samples when we don’t want to, or can’t, assume a theoretical distribution of the population and don’t want to rely on the Central Limit Theorem.
- Recall the Central Limit Theorem doesn’t always apply with less than 30 observations.

Bootstrap pitfalls

Why don’t we always use bootstrapping (or other resampling techniques), instead of estimating parameters ?

- Bootstrap sampling distributions tend to be too narrow: narrowness bias.
- Bootstrap distributions won’t fix nonrepresentative or too-small samples.
- Bootstrap estimates of the median can be problematic.
- May be computationally intensive.

Bootstrap Advantages

Conceptually simple, easy to implement, may be more intuitive than formulas for calculating standard errors:

- SE of the mean calculation is simple, but SEs of other statistics are much more complicated.
- Formulas for more than one predictor or response can be very complicated!

Can be used to illustrate concrete examples of theoretical principles:

- Bootstrapping is a good way to show how an empirical distribution compares to the theoretical.

Confidence Intervals

Parametric slope/intercept CIs

- If we use parametric inference, we can often* find a closed form solution for parameter estimates and standard errors.
- However,… often cannot find analytical solutions for the models we actually want to use!
  - For example, nonlinear models or other models with complicated structures.
- Bootstrapping can simulate a standard error for the alternative hypothesis.
- MC resampling can simulate a standard error for the null.

Monte Carlo Randomization

Resampling each column of the data separately.

- Creates new observations: combinations of x and y that weren't in the observed data.
  - Bill width of penguin #1 paired with bill depth of penguin #37.
- Breaks associations within sampling units:
  - The flipper length of a large penguin might be paired with the body mass of a small penguin.

Monte Carlo Randomization and the Null Hypothesis

The null hypothesis is that predictors and responses vary independently.

- There is no *coordinated* variation between x and y.
  - Large values of x are equally likely to be paired with large or small values of y.

Monte Carlo randomization destroys within-row associations thereby simulating the null.

- MC resampled data are what we could observe if the null hypothesis were true.
  - But remember that sampling error can result in unrepresentative samples in which strong associations are estimated by chance alone.

Resampling Examples: The Penguin Data - Categorical

Consider penguin species, a categorical predictor, and flipper length, a continuous response:

What are null and alternative hypotheses?

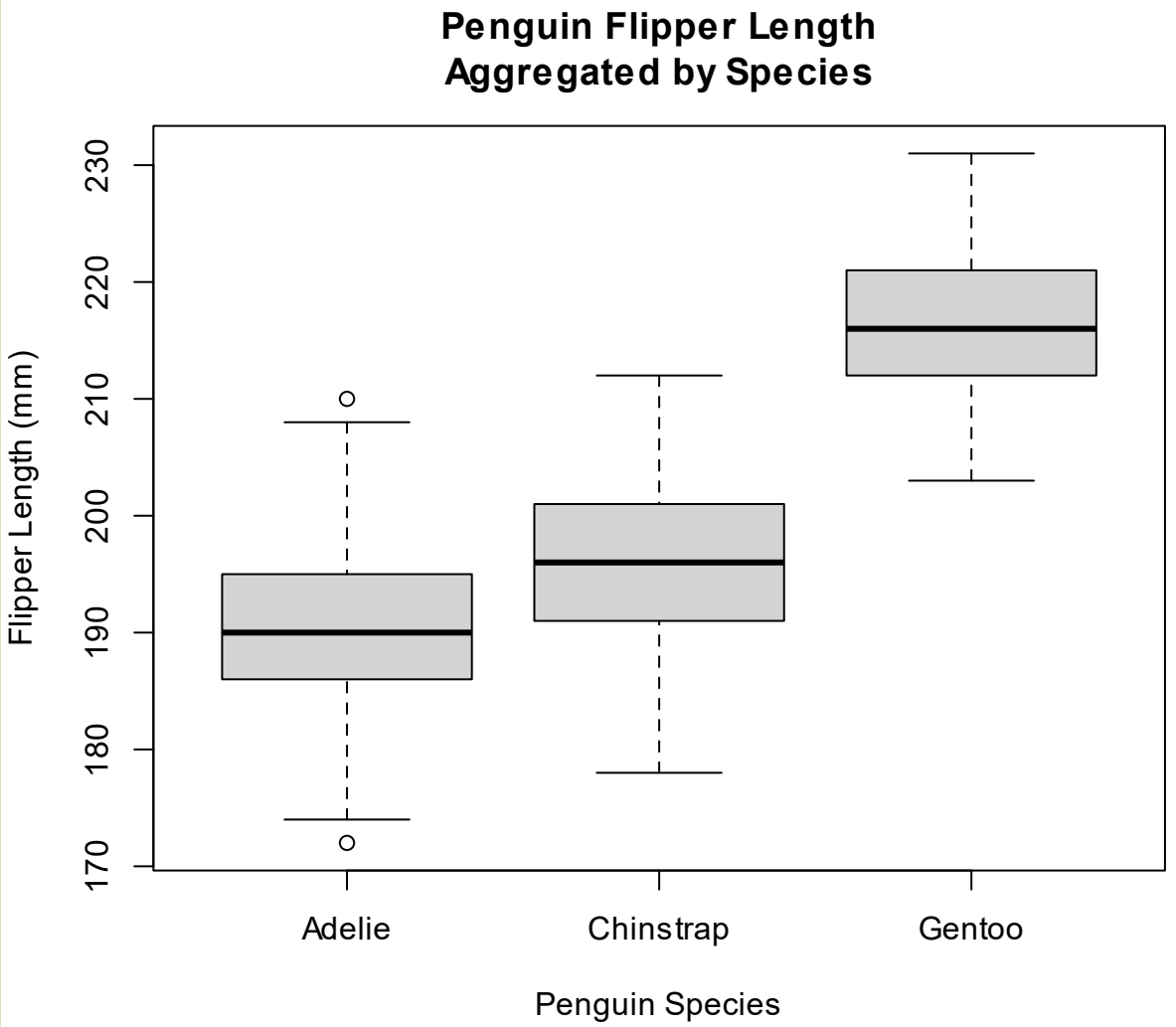
- Null: there is no association between penguin species and flipper size.
  - Flipper size does not vary among species.
- Alternative: Flipper size differs between *at least* one pair of penguin species.
  - If we have more than 2 species, we might not expect all species to be different.

Resampling Examples: The Penguin Data - Categorical

The flipper lengths, segregated by species.

- Think of the species ID as a label as an associated measurement.

Groups look well-separated.



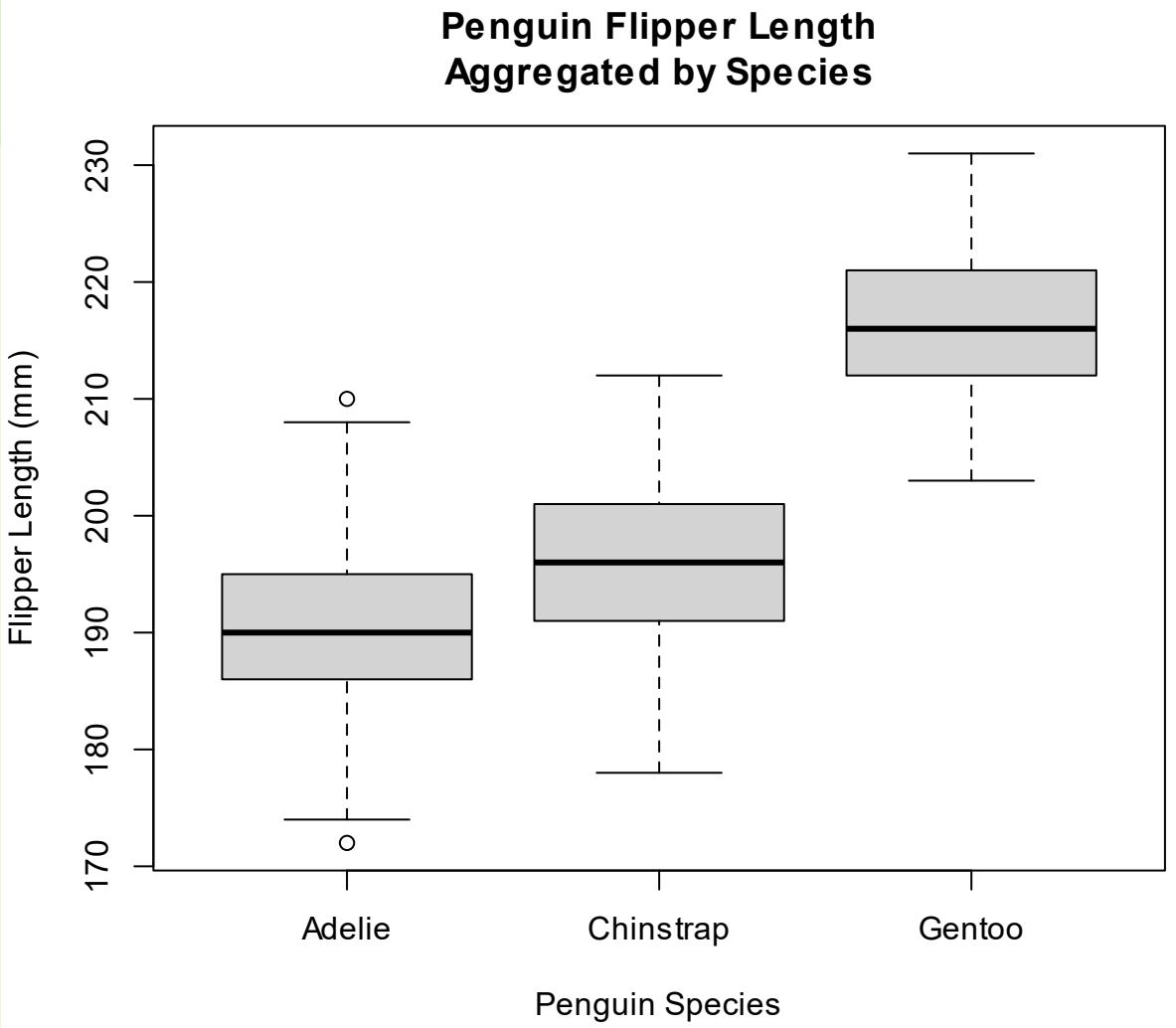
Resampling Examples: The Penguin Data - Categorical

The Original Data

It looks like Gentoo penguins have long flippers, while Adelie and Chinstrap penguins have shorter flippers.

- Could the apparent differences be due to sampling error?

We could perform some MC resampling to see how often we observe a pattern like what we see in the original data.



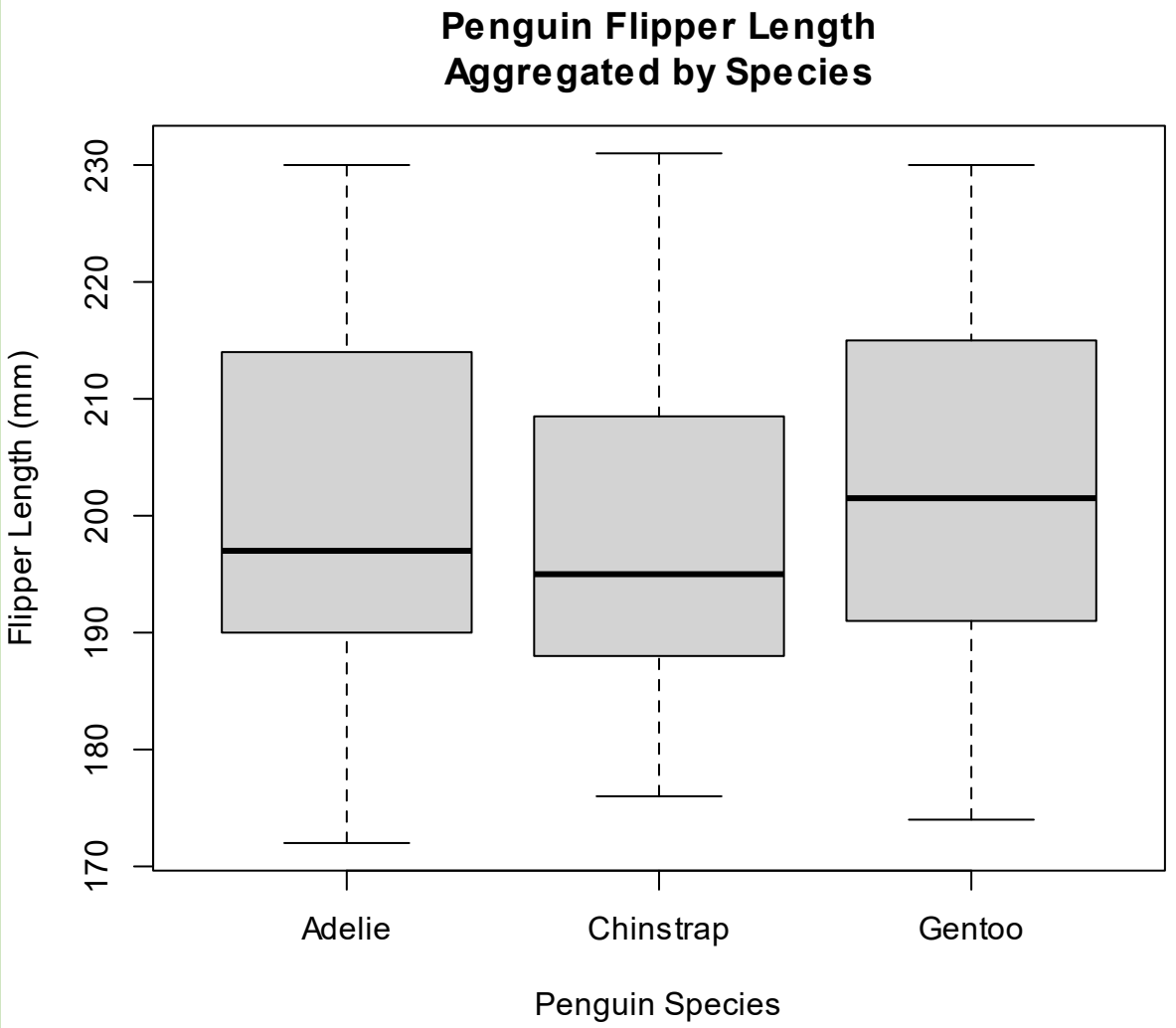
Penguin Flippers: Monte Carlo Randomization

Let's randomly assign species to flipper lengths:

- The medians are still slightly different, but there's more overlap. Now it looks like Chinstrap penguins are the smallest.

Why would I want to do this?

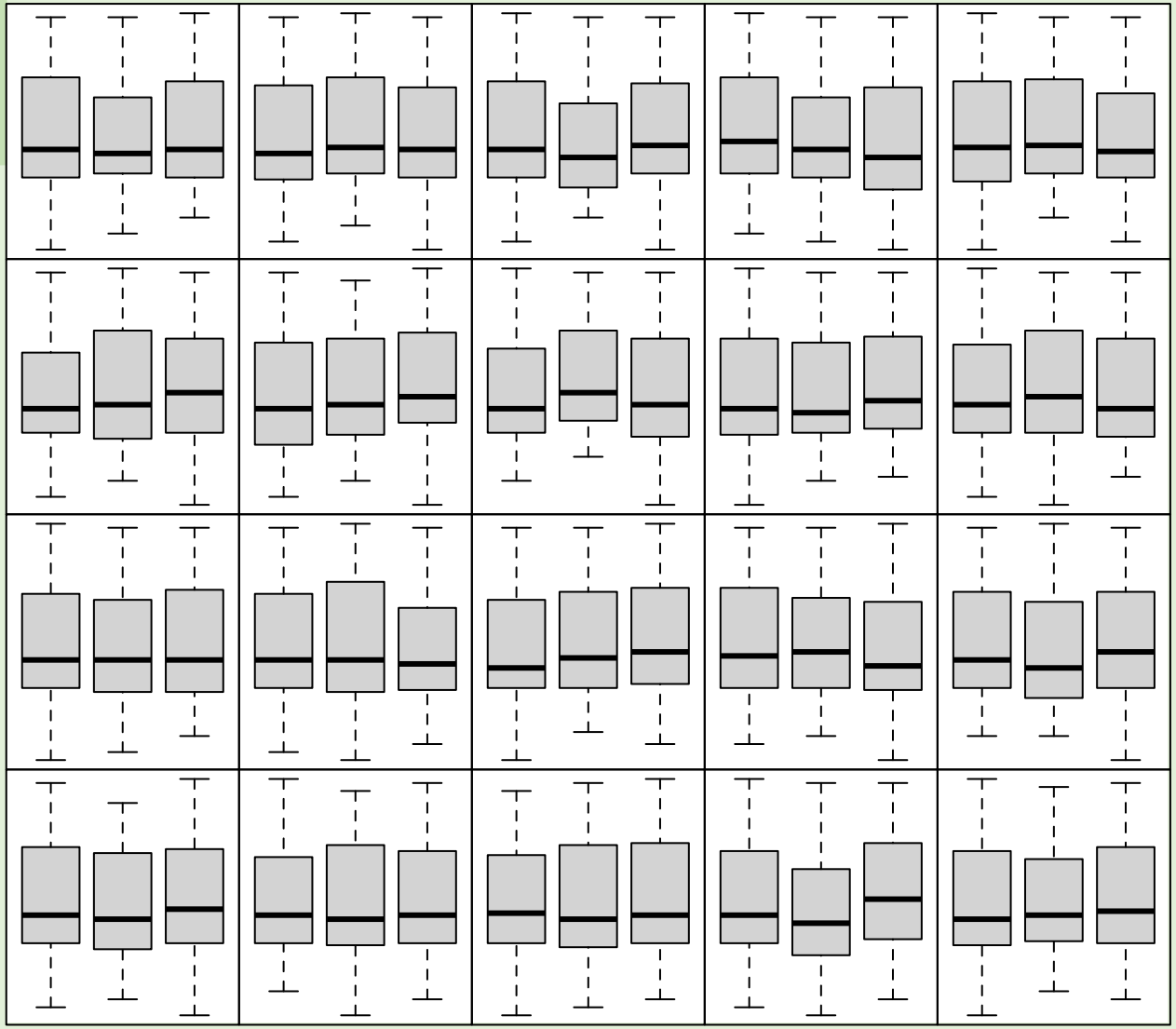
- It's all about comparing observed and expected.
- MC randomized data is the expectation under the null



Penguin Flippers: Monte Carlo Randomization

Let's look at some more MCMC randomizations:

- There is some noise, but they all look pretty same-ish.
- It's like collecting many samples if the null hypothesis were true!
- The behavior of many MC resamplings represents a **null distribution.**



Bootstrap Resampling: Resampling the Alternative Hypothesis

Null hypothesis: we've seen what the data look like when we randomly relabel the flipper lengths with species.

- This is a good representation of a null hypothesis.
- Labels are untethered from measurements.

Alternative hypothesis: flipper lengths for each species are drawn from *different* distributions.

- Bootstrapping resamples entire rows: it's analogous to taking multiple samples from the population.
- Labels are kept with their measurements
- If the null hypothesis were true, bootstrapping results should be indistinguishable from the MCMC results.

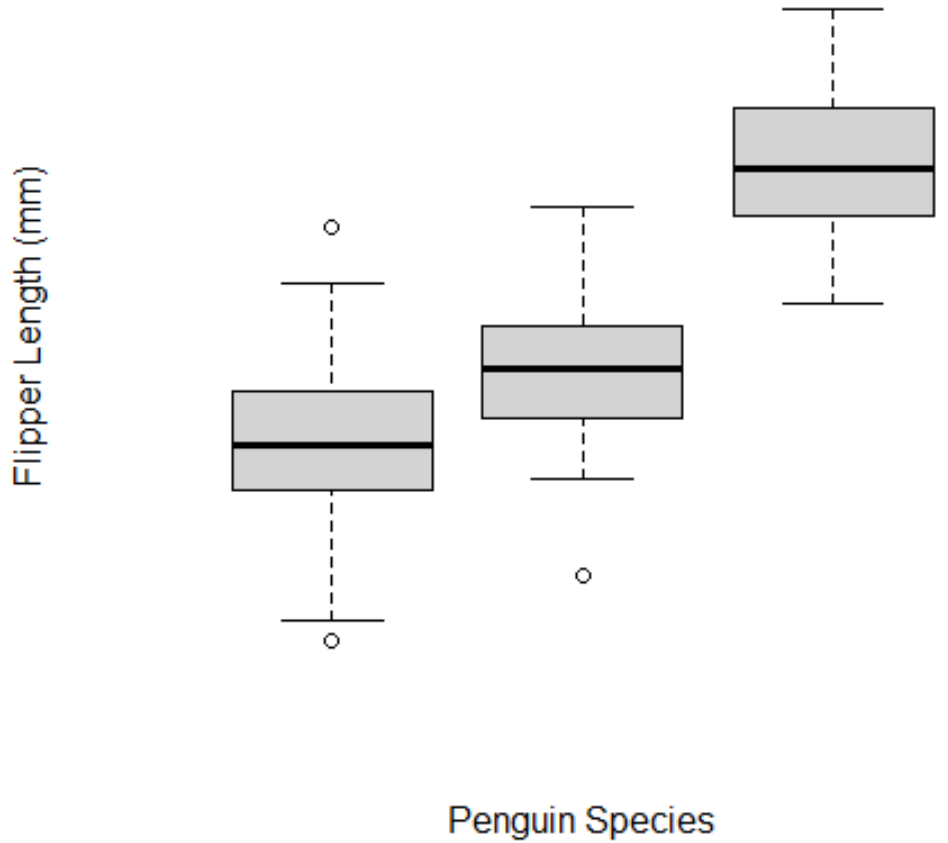
Penguin Flippers: Bootstrap Randomization

Let's randomly sample some rows of data

- It looks pretty similar to the original data.
- This approximates what we might have observed if we could do another experiment

Why would I want to do this?

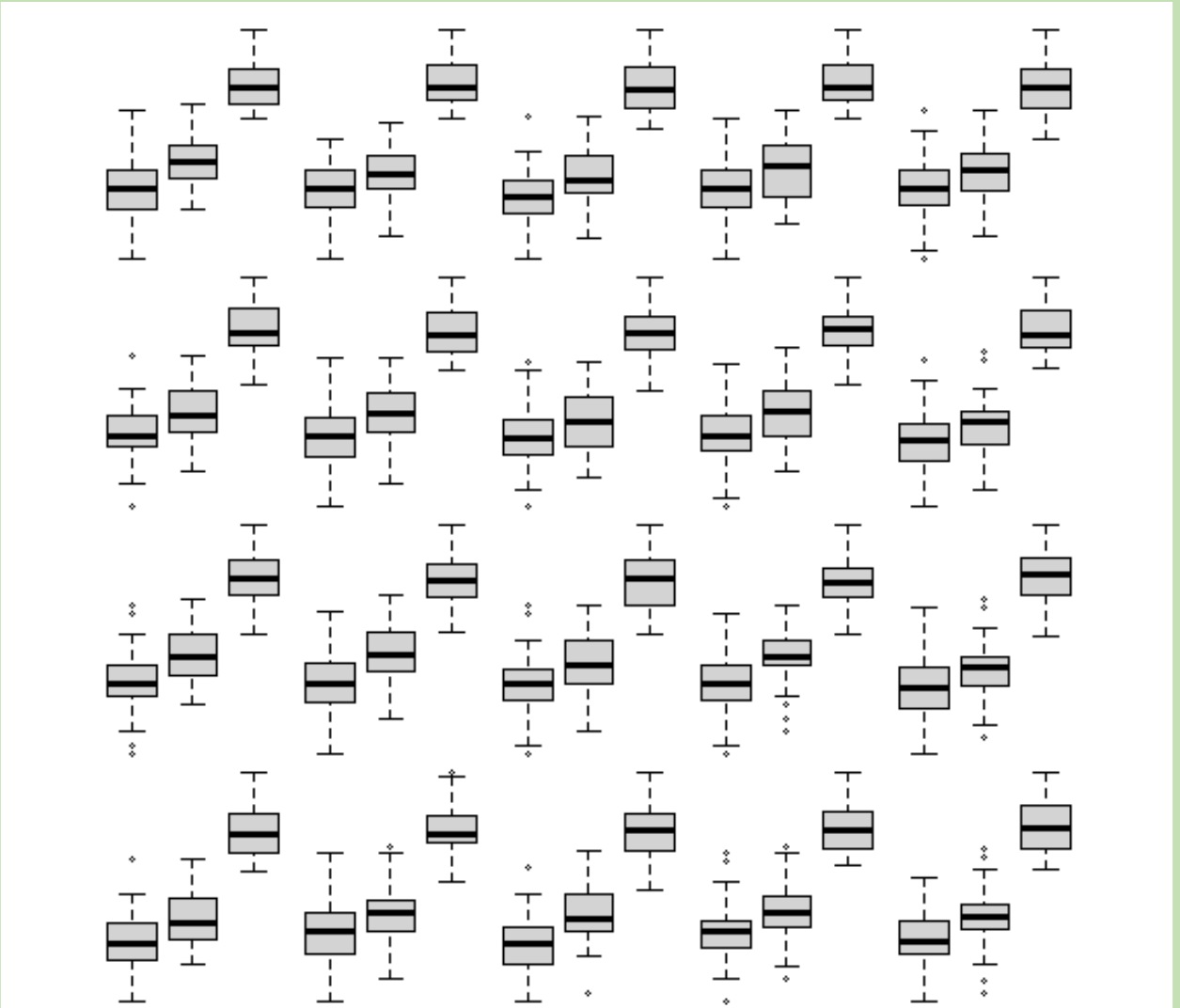
- Generates an alternative distribution
- Can help detect outliers/influential observations.
  - These can be real data, or they might indicate a data quality issue.



Penguin Flippers: Bootstrap Randomization

Let's look at some more bootstrap randomizations:

- There is some noise, but the all look pretty same-ish. It's like collecting many samples if there were the null hypothesis were false!
- This is the essence of the Frequentist ideal of repeated sampling.
- Since most realizaitons look the same, there are probably not many outliers or extremely influential observations.



What's in This Section?

Slides	Take-Home Concepts
• What is resampling? • Sampling with replacement. • Bootstrap and Monte Carlo randomization • Resampling the null hypothesis • Resampling the alternative hypothesis	• Resampling for the null (Monte Carlo) and alternative (Bootstrapping) hypotheses. • Why sample with replacement? • When is resampling useful? • Breaking vs. retaining associations in the data. • Labeled data.

Deck 6 Recap

Important concepts

- Optimization criteria
  - What are the criteria for least squares and maximum likelihood methods?
- What is likelihood?
  - How do we measure/optimize for likelihood?
- What are least squares methods?
  - How do we measure/optimize least squares?
- When are likelihood and least squares equivalent?
- What are two major classes of resampling, and what do the tell us?

In-Class Likelihood and general Q+A

Any questions?