Analysis of Environmental Data

Frequentist Concepts: Hypotheses and Confidence

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst
Michael France Nelson

Announcements: Oct 12

- Welcome to week 6, how are y'all doing?
- Can you believe the semester is nearly halfway done ☺☹?
- Make sure you've read the Library of Babel for Thursday's class.
  - Discussion
  - In-class activity

Frequentist Concepts

Frequentism Overview: Key Concepts

- What is the Frequentist paradigm?
- Assumptions
- Frequentist pros and cons
- Repeated sampling
- Infinitely large, unknowable population
- Confidence and significance

Nota Bene

Frequentism is not your enemy!

The following slides might sound critical of Frequentist statistics.

- Frequentism is a massively useful framework!
- There is lots of insight to be gained via Frequentist inference.

The Frequentist framework contains some subtle assumptions and concepts.

- Understanding the subtlety will allow you to make the most of Frequentist inference.

It's often fashionable to disparage Frequentist statistics: Disliking popular things does not make you interesting.

Frequentism

An inference *paradigm*	**Frequentist statistics is currently the dominant paradigm**
It's a framework that contains a set of tools and assumptions that we can use to do inference. All toolkits have assumptions and limitations – It's up to us to understand and work with them. There are lots of theoretical and practical tools for practicing Frequentist inference	It's the kind of inference you [usually] learn in a first course on statistics. • It's not the *only paradigm*! We'll have a chance to touch on the Bayesian world later on. You might already know the Frequentist versions of popular analyses like ANOVA, linear regression, etc. • Lots of other scientists know them too – It's easy to communicate in the language of Frequentism.

Frequentism

Some key features

- Makes no assumptions about prior knowledge of a system.
  - This is only approximately true – prior knowledge is incorporated implicitly within the kinds of questions we ask and how we implement our research.
- It's based on the idea of *hypothetical repeated sampling* of a population that is *unknowable*.
- The focus is on the *process of sampling and modeling* not on the parameters derived from a *specific data set*.
- Some describe the frequentist paradigm as less 'subjective' than Bayesian
- More widely known than Bayesian.
  - Usually more mathematically and computationally tractable than Bayesian.
  - Don't fall for the Bayesian vs. Frequentist hype – they're both useful frameworks, neither is perfect.

Frequentist Pros and Cons

Benefits of the Frequentist framework	Drawbacks of the Frequentist framework
• Widely known, lots of established methods, backed by a lot of theoretical work. • Many software tools • Assumptions are often reasonable and/or robust to violations • Frequentism is a very powerful inferential framework	• Repeated sampling assumption: initially non-intuitive interpretation of *confidence* and *significance*. • Focus on the process rather than the results of a single experiment. • Focus on hypothesis-testing: 'straw man null hypotheses' (Bolker)

Frequentism: Tools and Assumptions

Common Methods	Key assumptions
Most of the tools you have probably used in previous courses or research have seen frequentist versions of: • Linear regression, General Linear Models • T-tests • ANOVA • Generalized Linear Models	Some of the conceptual weirdness in Frequentism comes from the assumptions: • The population is large and unknowable • There is a set of true parameter values for a model of the population. • We can never know the true parameter values. • Our ability to do inference is based on hypothetical repeated sampling

Frequentist Populations and Models

Populations are large

In Frequentism, we propose:

- The population exists and is has true properties.
- We must use *samples* to make educated guesses about the properties of the population.

When we propose a model, we assume that there exist true parameter values to characterize the population.

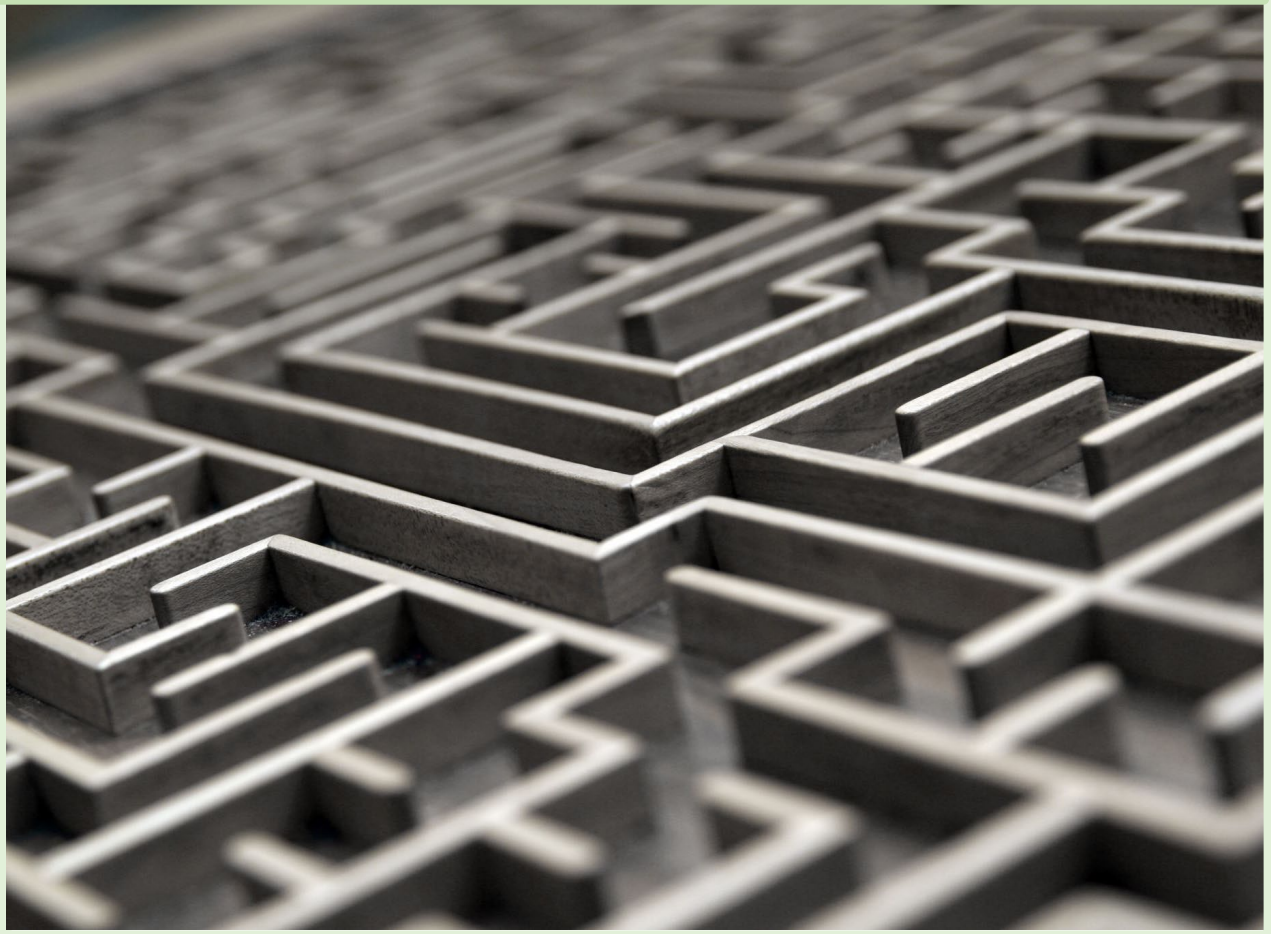
In general, we can't observe every *sampling unit* in a population.

Repeated Sampling

This concept is the source of much confusion

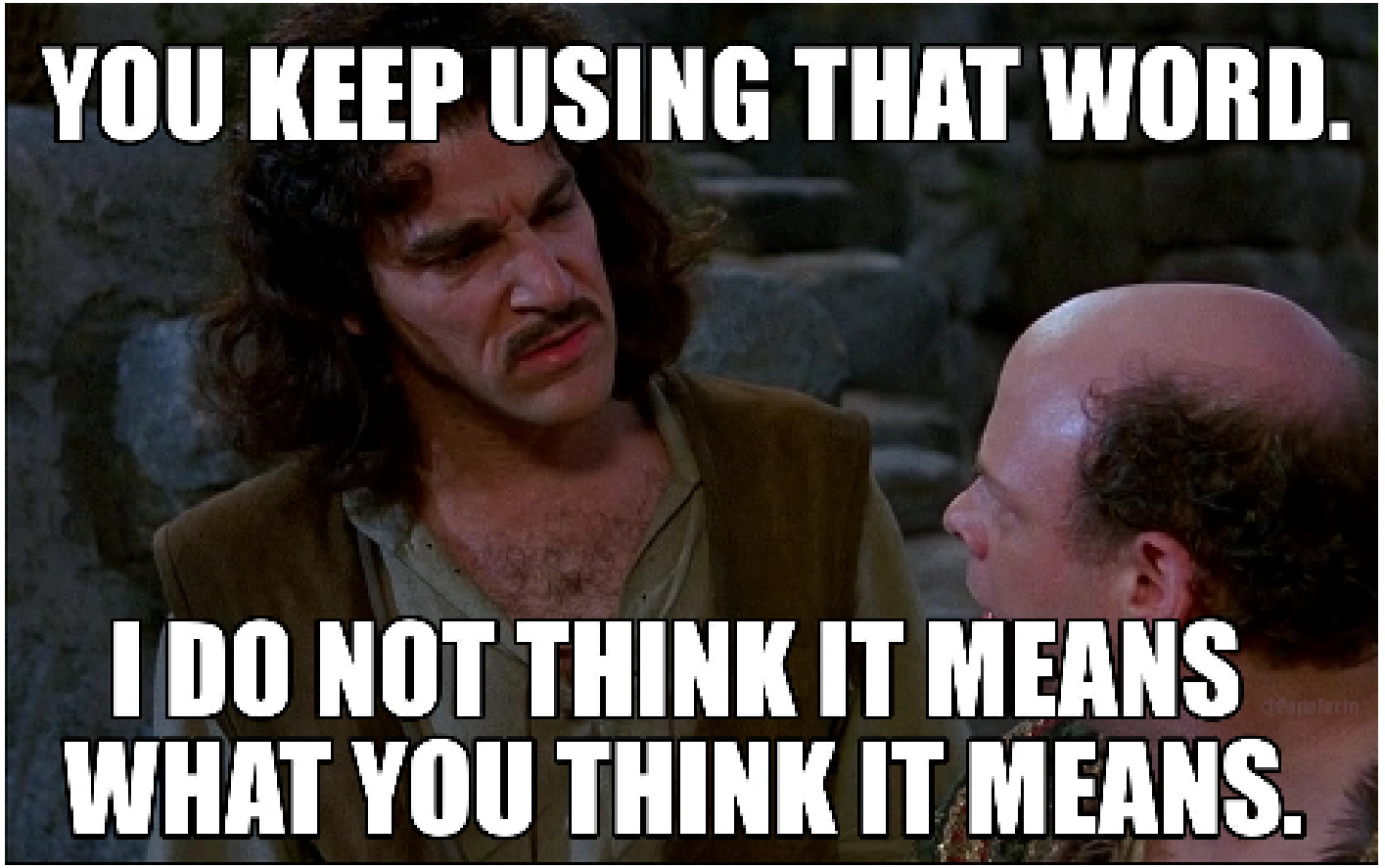
- The theoretical underpinnings of Frequentist inference rely on the concept of *repeated sampling*.
- Each sampling effort is a realization of a **stochastic process**.
- If we could repeat the sampling process an *infinite* number of times, our estimates would converge upon the true population parameter values.

We can simulate repeated sampling in R to help us build intuition.



Frequentist Confidence and Significance: Collision with Everyday Usage

- The meaning of these terms is in relation to hypothetical *repeated sampling*.



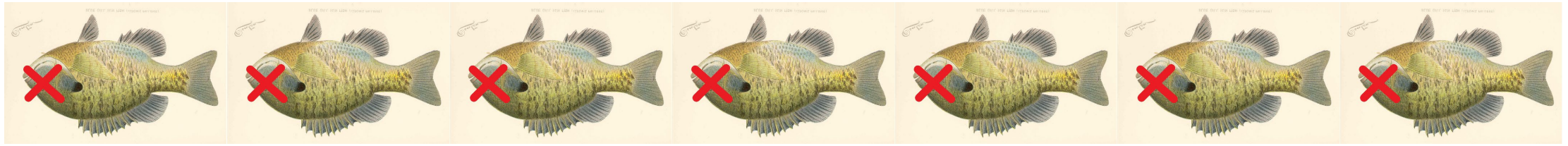
Frequentist Confidence and Significance

What we want a confidence interval to mean:

- Based on the results of my research, I am 95% sure that the true mean length of bluegill in Massachusetts is between 9 and 10cm.

What a confidence interval actually means:

- If I could repeat my research procedure many times, I expect that the true mean length of bluegill would fall within my 95% confidence intervals approximately 95% of the time.



Frequentist Confidence and Significance

Frequentism focuses on the process.

Frequentist statements of confidence and significance do not refer the outcome of a *specific* experiment.

- This isn't exactly what we want from our analysis.
- It's difficult to explain the "if I repeated…" idea to a nonscientific audience.

It's not that a Frequentist analysis doesn't let us make conclusions from an experiment, it's just that we have to be careful to acknowledge the subtly of focusing on the process in the context of repeated sampling, rather than a single outcome.

Watch for how popular media describes statistical results.

Recap of Key Concepts and Terms

- What is the Frequentist paradigm?
- Infinitely large, unknowable population
- Repeated sampling
- Assumptions
- Frequentist pros and cons
- Confidence and significance

Sources of Error

Key Concepts and Terms

- Types of ‘error’
  - Measurement
  - Process
  - Model
  - Sampling
- Stochastic processes and realizations

Sources of Error

McGarigal describes 3 sources:	**Error is such a negative-sounding term; I prefer 'noise'**
• Measurement • Process • Model	I'd like to highlight a fourth, underappreciated, source of noise: • **sampling error:** errors that arise from non-representative samples.

Measurement Error

How can we commit measurement errors?

We cannot measure with perfect precision or accuracy.

- Flawed instruments
- User error
- Failing to observe individuals or species that are actually present
- Over- or under-counting

Process Error

There is randomness inherent in any natural system, for example:

Demographic stochasticity	Environmental stochasticity

• Random fluctuations in births/deaths of individuals within a population. Each individual is unique, with a unique life trajectory.
• Small populations are especially subject to demographic stochasticity
	• Weather fluctuations
	• Climate fluctuations
	  • How is this different from weather
	• Chance extreme events
	  • Hurricanes, droughts, blowdowns

Model Error

A mis-specified model can arise from:

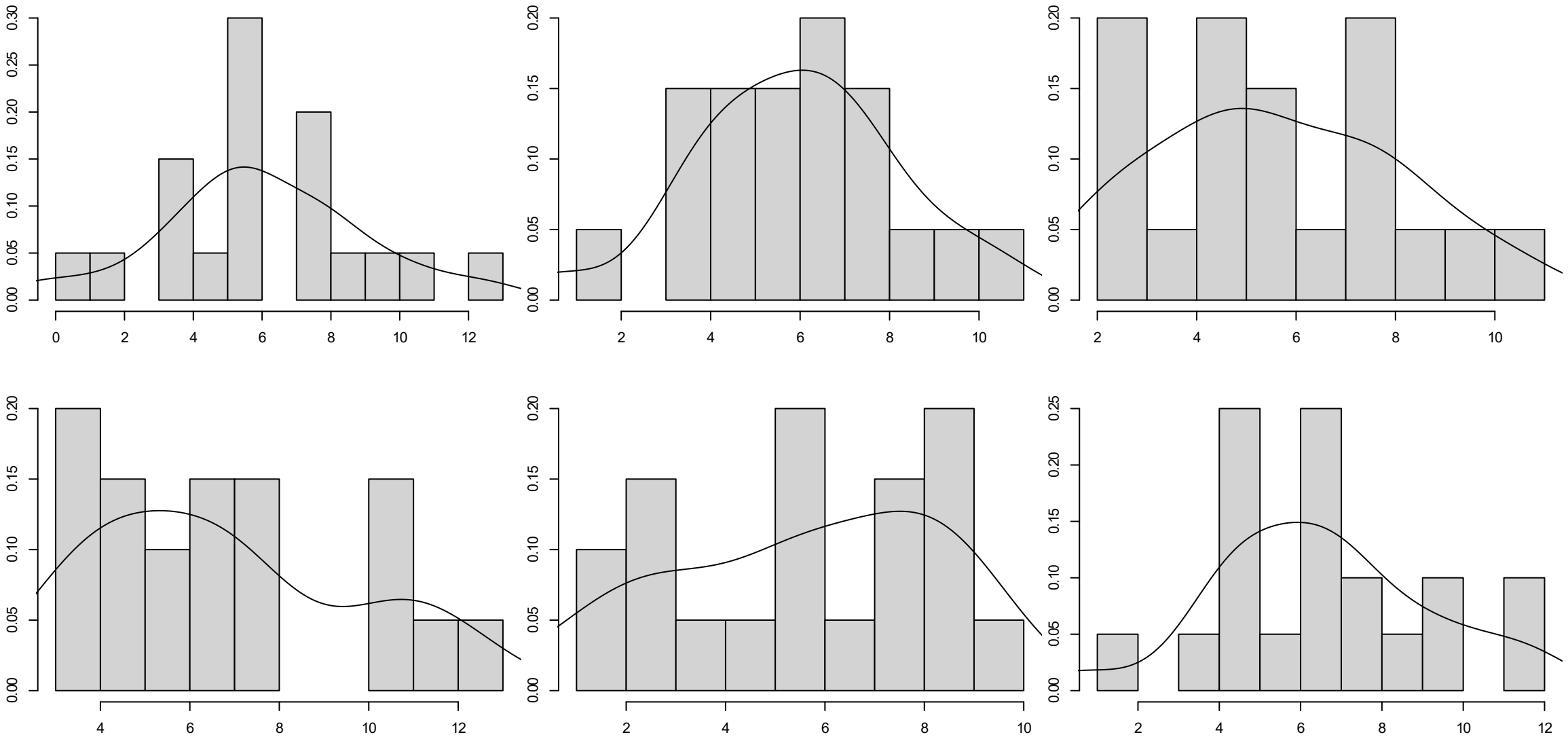
- omission of important predictors
- inclusion of unimportant predictors
- complicated dependencies in predictor variables: multicollinearity.
- mis-specified relationships: linear, exponential, logarathmic
- wrong deterministic model
- incorrect error structure: wrong stochastic model

Sampling Error

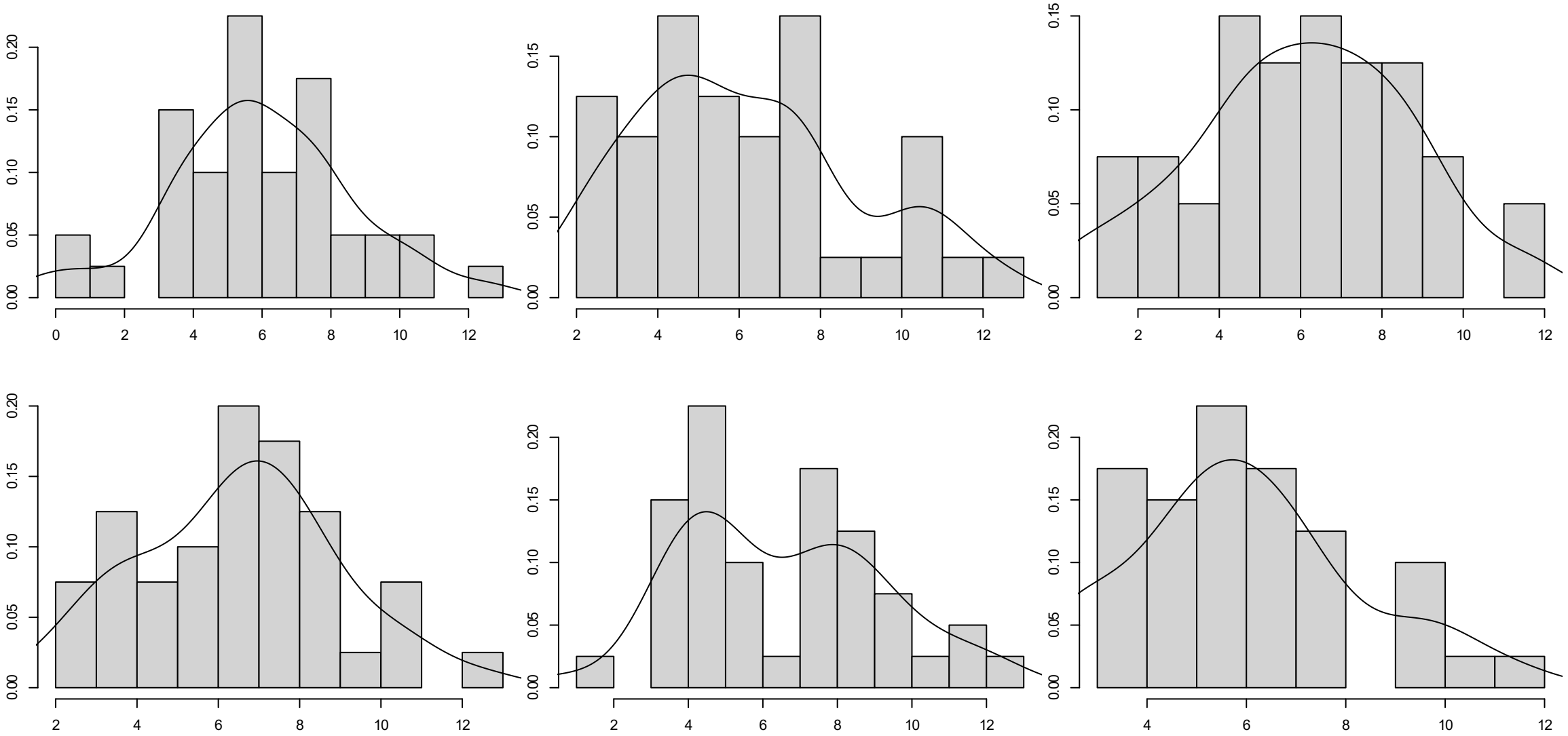
Sampling error is due to *nonrepresentative* sampling

Sampling is, or should be, a realization of a *stochastic process*.	**Small samples are especially subject to sampling error**
• Different realizations produce different samples. • Some randomly chosen samples are not representative! • What if you get 10 heads in a row? • You can reduce, but never eliminate, sampling error by increasing sample size	• Genetic drift is a form of sampling error • With a small number of observations, it's hard to tell which distribution a population might follow. • Can you guess the distribution of the population in the following slides?

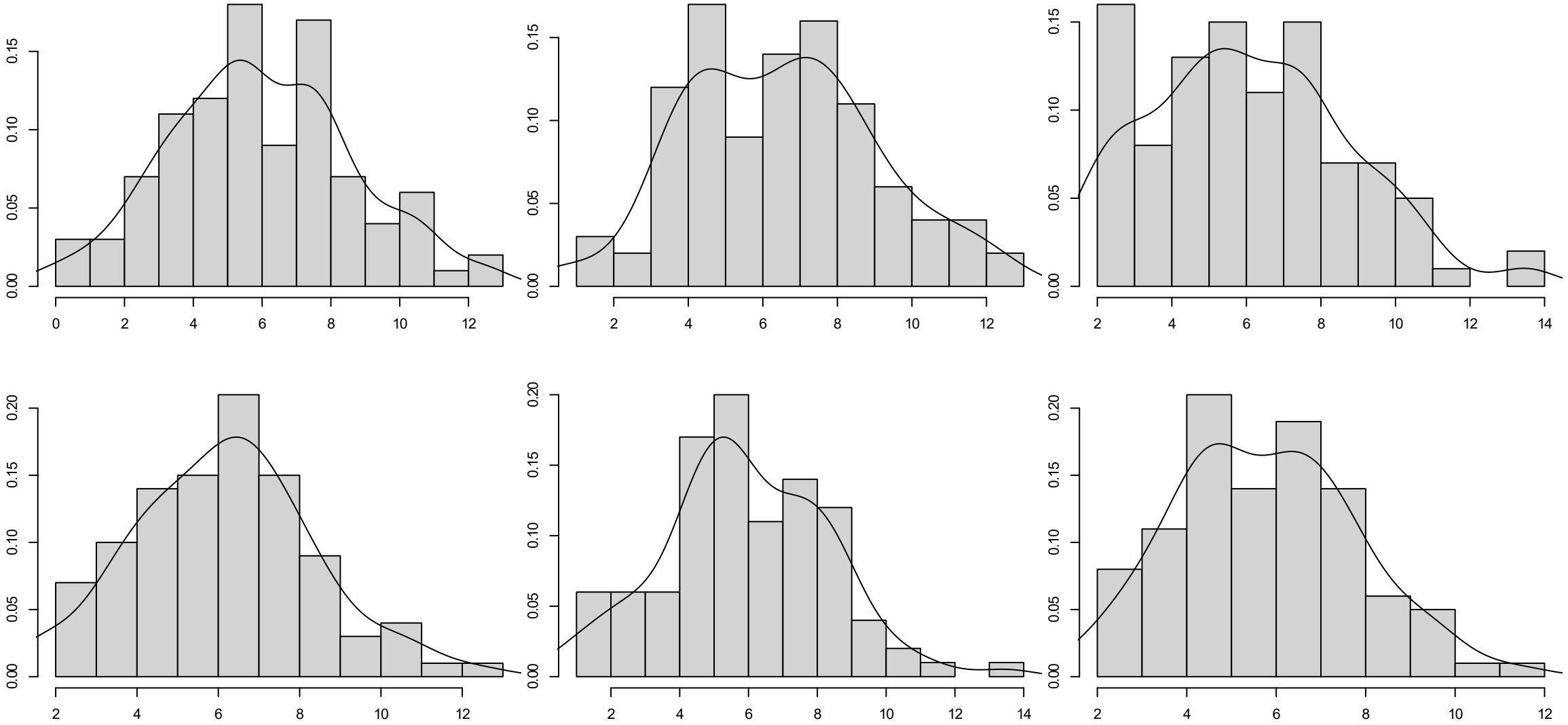
Sampling Error: Graphical Intuition n = 20



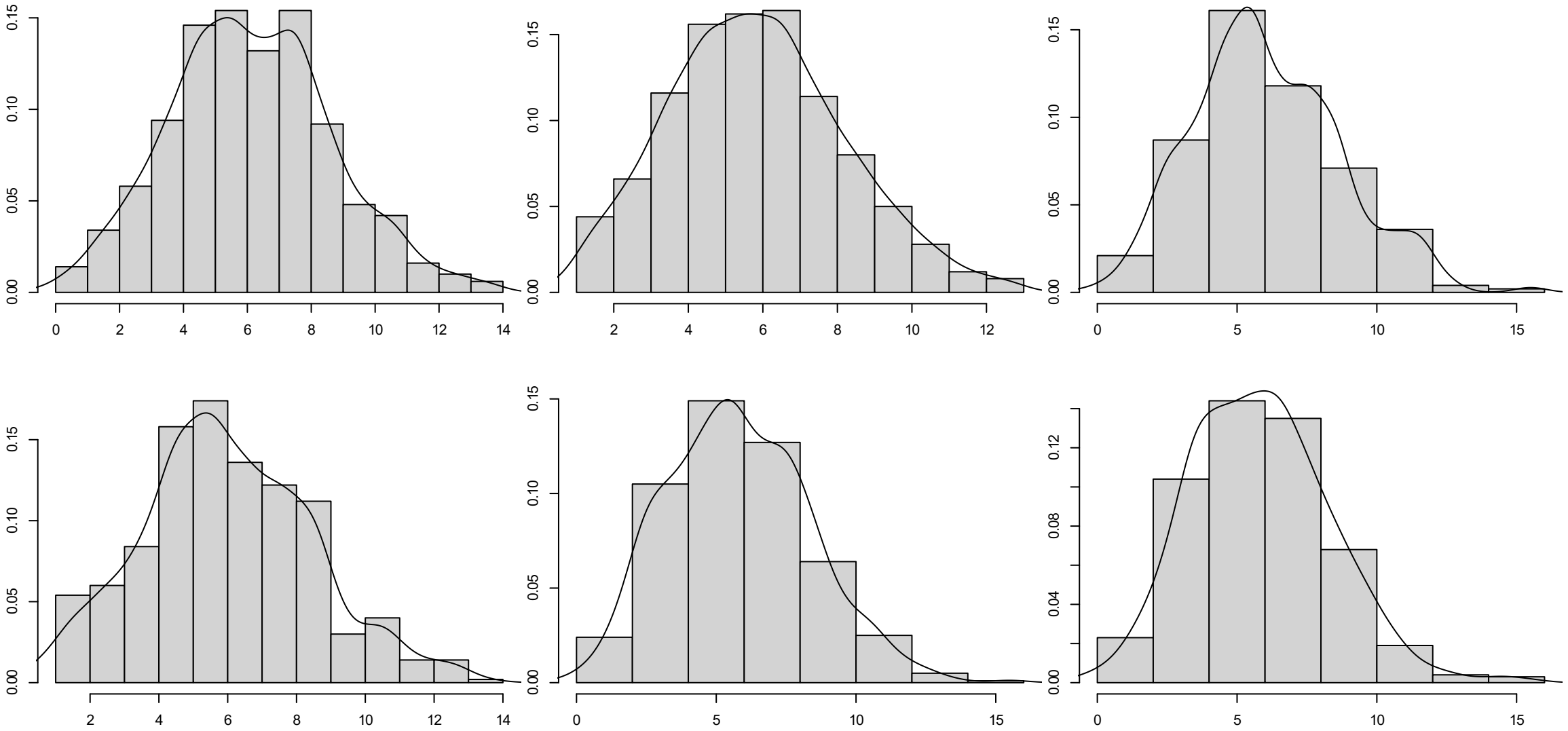
Sampling Error: Graphical Intuition n = 40



Sampling Error: Graphical Intuition n = 100



Sampling Error: Graphical Intuition n = 500



- They were sampled from a Poisson distribution

Stochastic Process

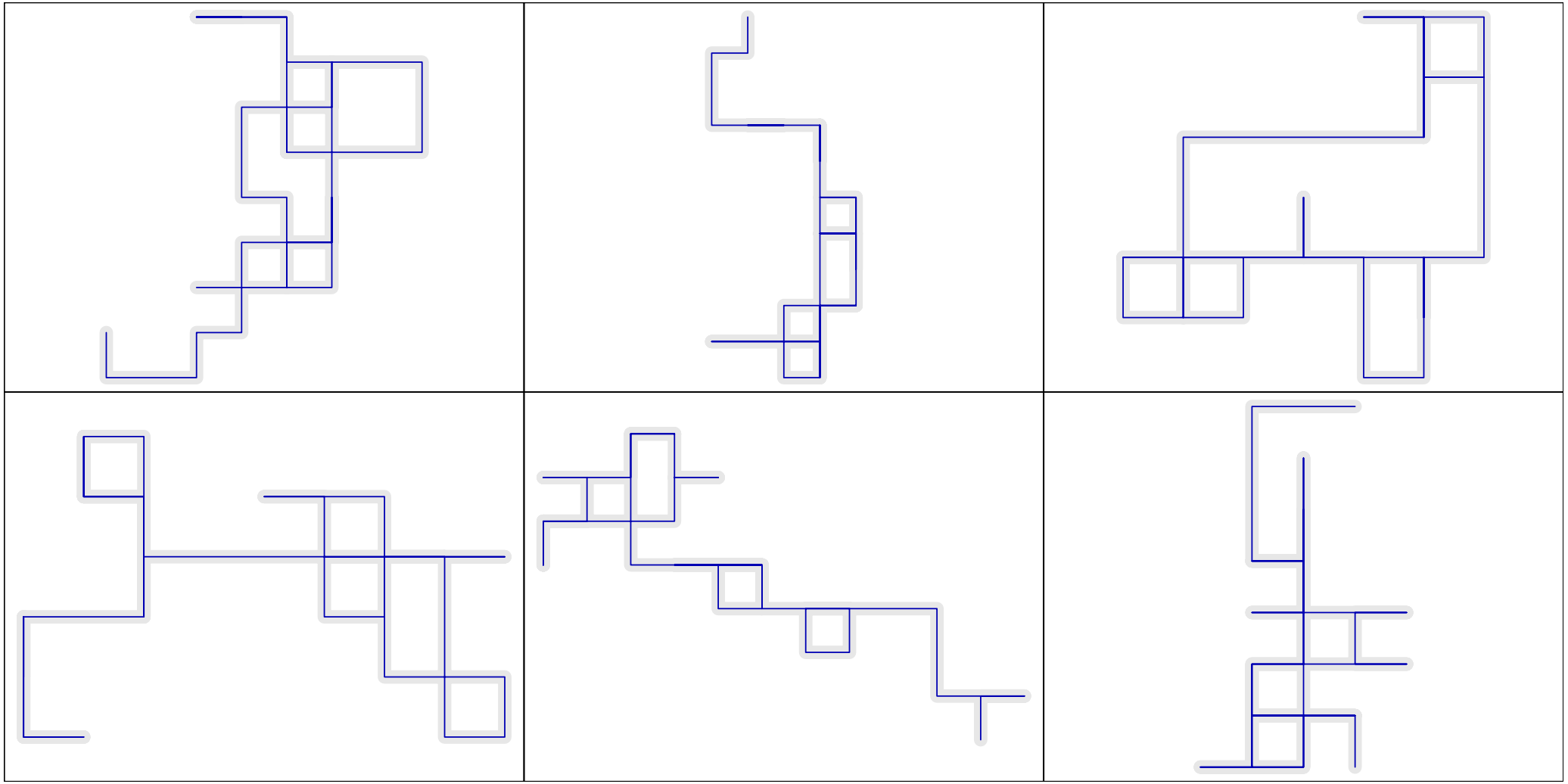
"A stochastic process is any process describing the evolution in time of a random phenomenon. From a mathematical point of view, the theory of stochastic processes was settled around 1950. Since then, stochastic processes have become a common tool for mathematicians, physicists, engineers, and the field of application of this theory ranges from the modeling of stock pricing, to a rational option pricing theory, to differential geometry."

- F. Baudoin, in International Encyclopedia of Education (Third Edition), 2010

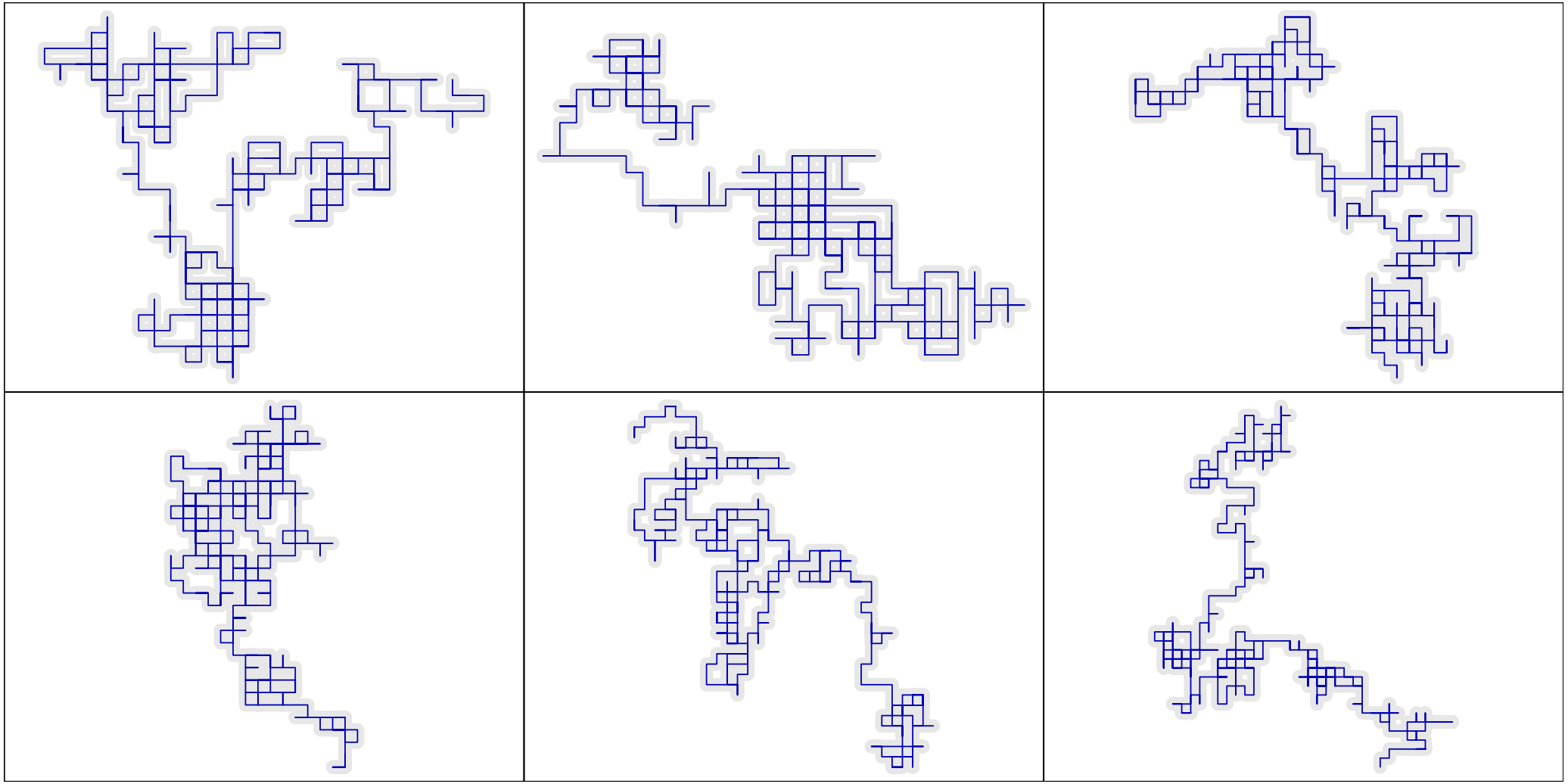
Some Stochastic Process Examples

stochastic processes	stochastic process realizations
• Flipping a coin • Rolling dice • Drawing a hand of cards • The lottery • A random walk • Brownian motion	• A book or phrase in the Library of Babel: Oh tiempo, tus piráides (oh time, your pyramids) • What other phrases are possible? • A royal flush • Heads, tails, tails, heads • Observing 7 brown creepers in a 1-ha plot • The ending point after a random walk

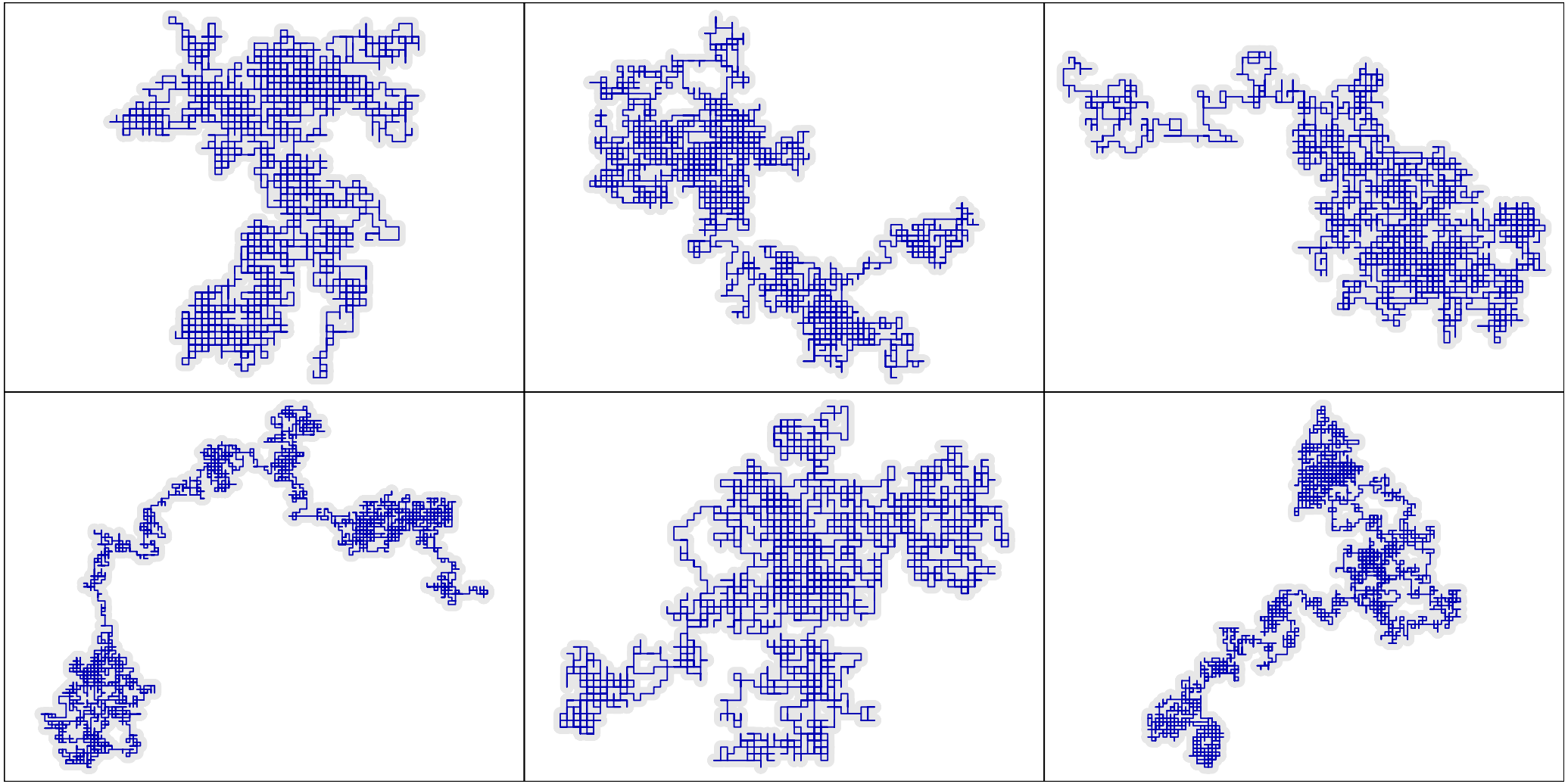
Random Walks: 50 steps



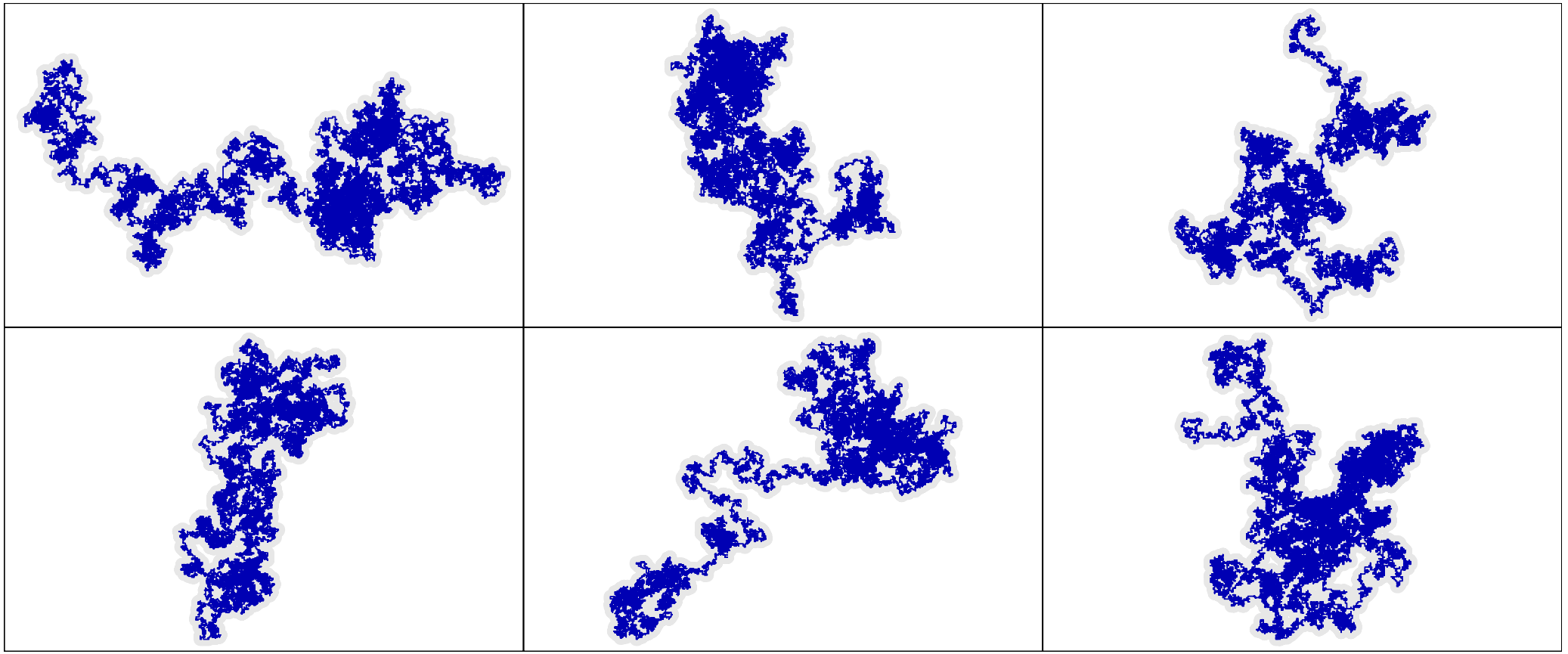
Random Walks: 500 steps



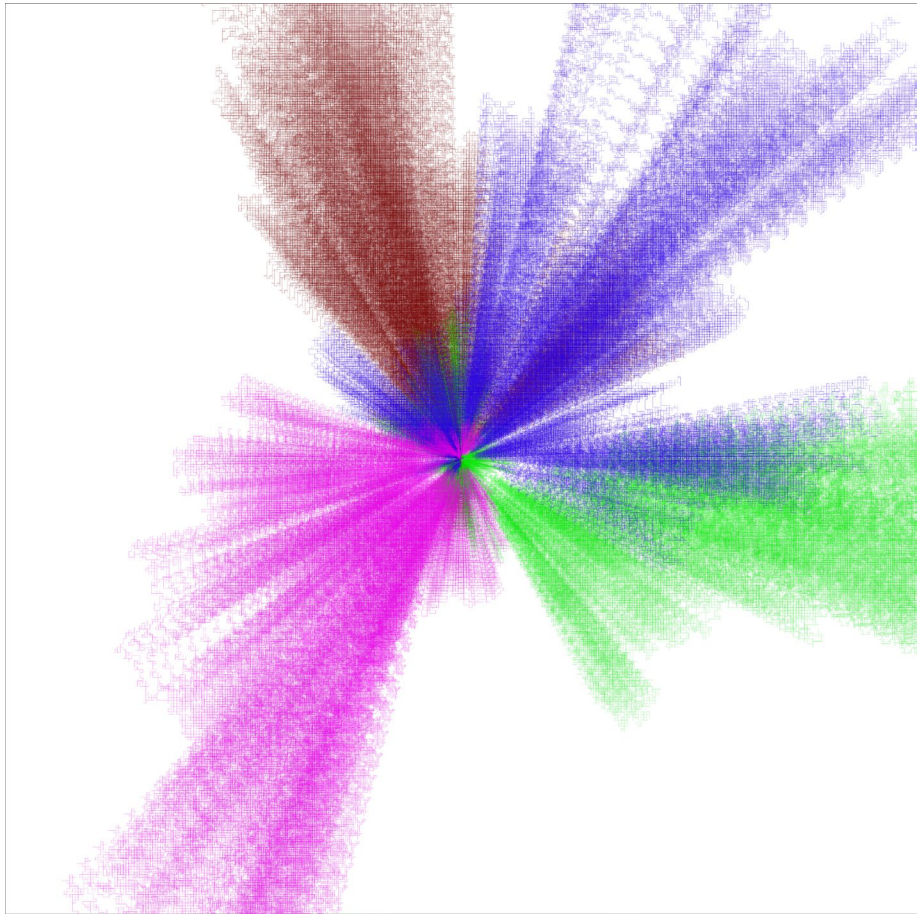
Random Walks: 5000 steps

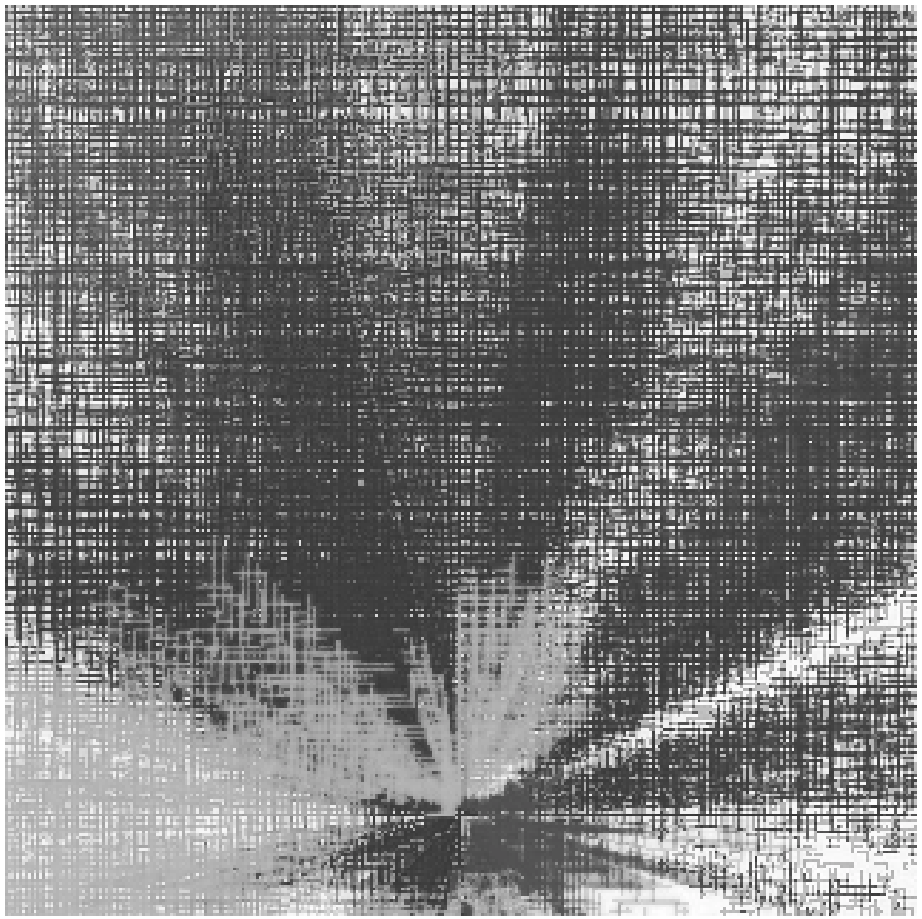


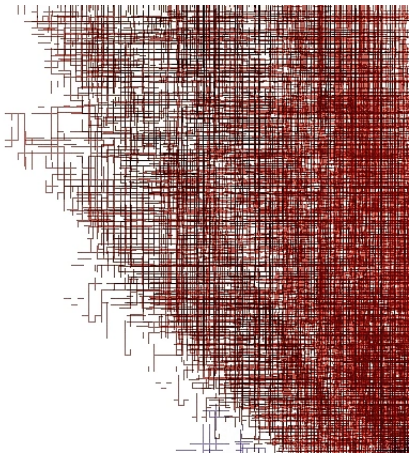
Random Walks: 50k steps

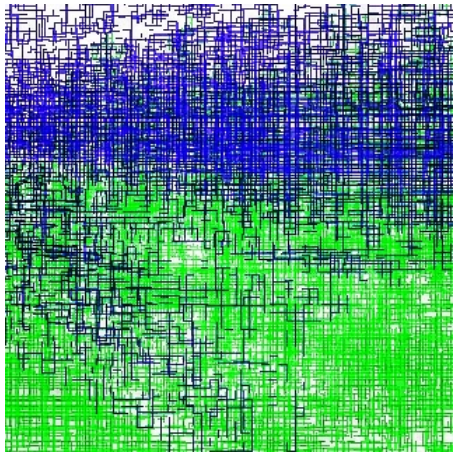


Random Walks: 5000 steps, cool R art









Key Concepts and Terms

- Types of ‘error’
  - Measurement
  - Process
  - Model
  - Sampling
- Stochastic processes and realizations

Hypothesis Testing

A Parametric Frequentist Approach

What's in This Section?

Slides	Take-Home Concepts
• Null Hypotheses • Alternative Hypotheses • Decision Criterion • 1- and 2-tailed hypotheses • T-test: applying the decision criterion	• Understanding Frequentist hypotheses • Interpreting a p-value • Don't do data-snooping! • What is a t-test?

A Classical Approach

Hypothesis testing is central to Frequentist statistics.

- Frequentist hypothesis testing often uses *parametric* distributions, and their accompanying assumptions to do inference.
- This lecture uses a t-test example. T-tests are easy to understand in a Frequentist parametric paradigm!

Other approaches to hypothesis testing

Note: there are other possible approaches that use Frequentist concepts, such as simulation and resampling techniques.

Hypotheses

What is hypothesis testing?

- A way to quantify the strength of evidence against a null hypothesis.
- We propose a *null hypothesis* that captures what we expect to see if associations are random.
- Perhaps a *straw man hypothesis* as Bolker mentions.
- We propose an *alternative hypothesis* stating what we think is *actually* happening.
- We propose a decision criterion, often a p-value, by which we can reject or fail to reject the null hypothesis in favor of the alternative.

Null Hypotheses: Model Thinking Perspective

A null hypothesis is…

- When we build a model, we like to consider *associations* among model components.
- We usually have some idea about the nature of the association:
- More water is associated with greater plant biomass, etc…

If the association were random…

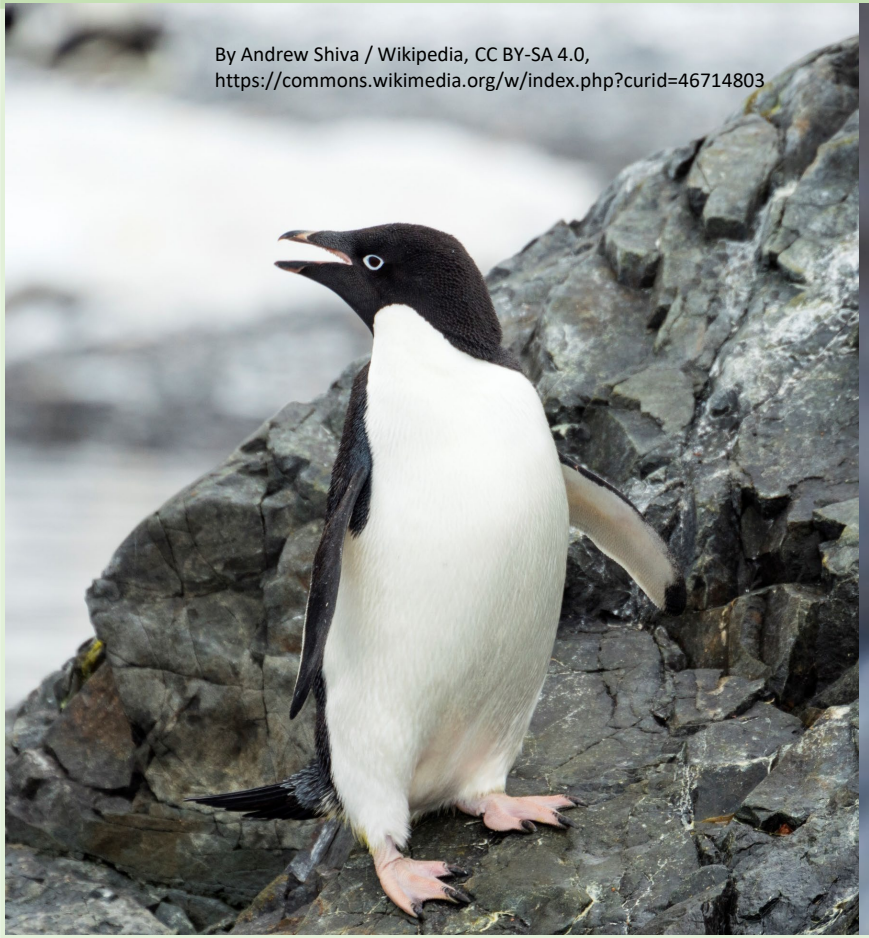
- In a null hypothesis, we usually that associations are totally random, i.e. quantities do not vary in a coordinated way.

Alternative Hypotheses

2-tailed hypotheses: non-directional	**1-tailed hypotheses: directional**
General: We don't propose a particular direction of the association. • Increased light might be associated with *greater* or *lower* biomass in an invasive plant. • We have no prior knowledge of its light response • We think that there will be *some* difference, but we don't' know the direction	More specific than 2-tailed hypotheses. • We have some previous knowledge that allows us to propose a *direction* to the association. • From our experience in the field and previous research we think that the plant will have *more* biomass in *higher* light. • We propose a *positive* association.

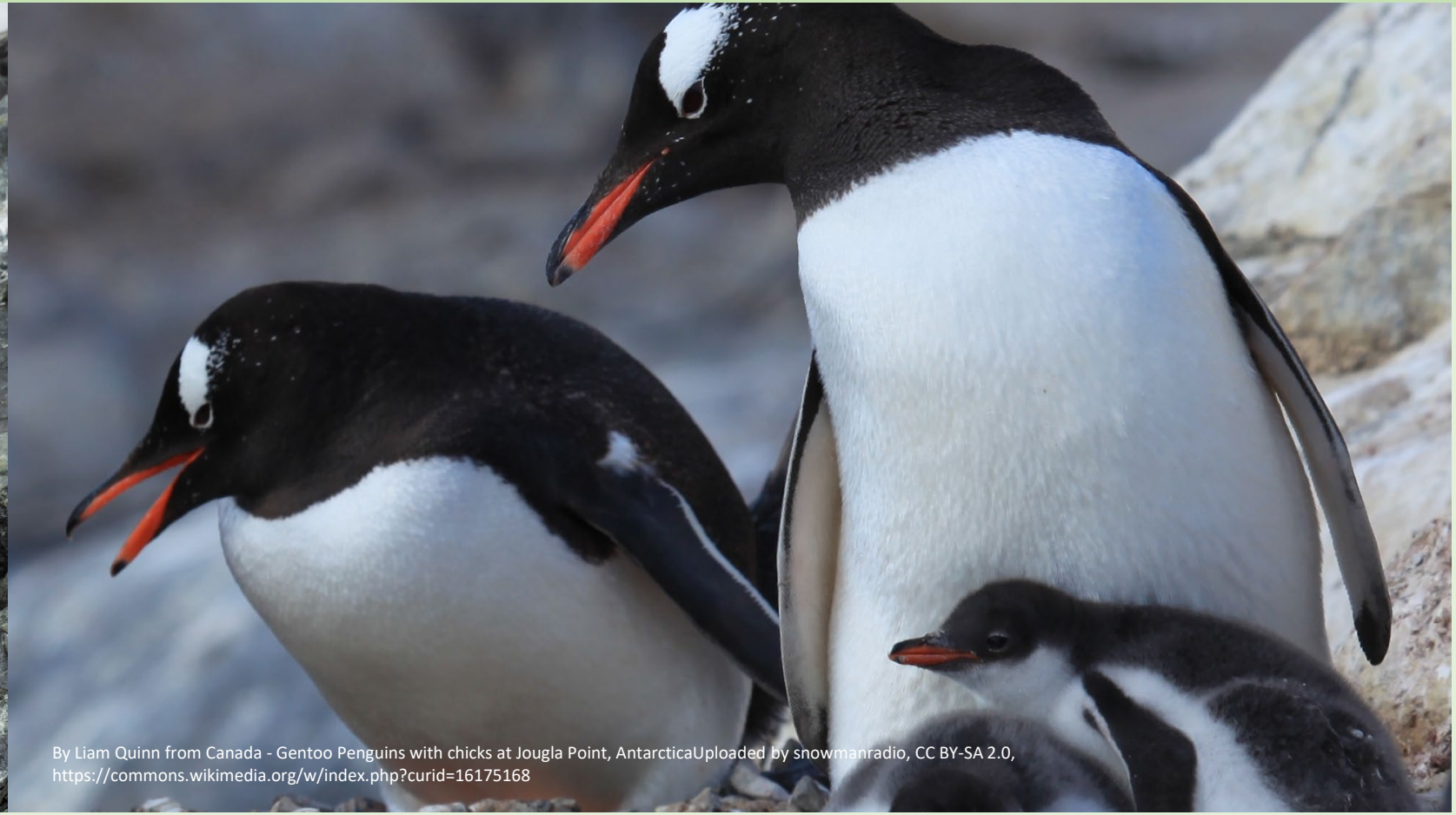
A Tale of Two Penguins

Adelie Penguins



By Andrew Shiva / Wikipedia, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=46714803

Gentoo Penguins



By Liam Quinn from Canada - Gentoo Penguins with chicks at Jougla Point, AntarcticaUploaded by snowmanradio, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=16175168

A Null Hypothesis

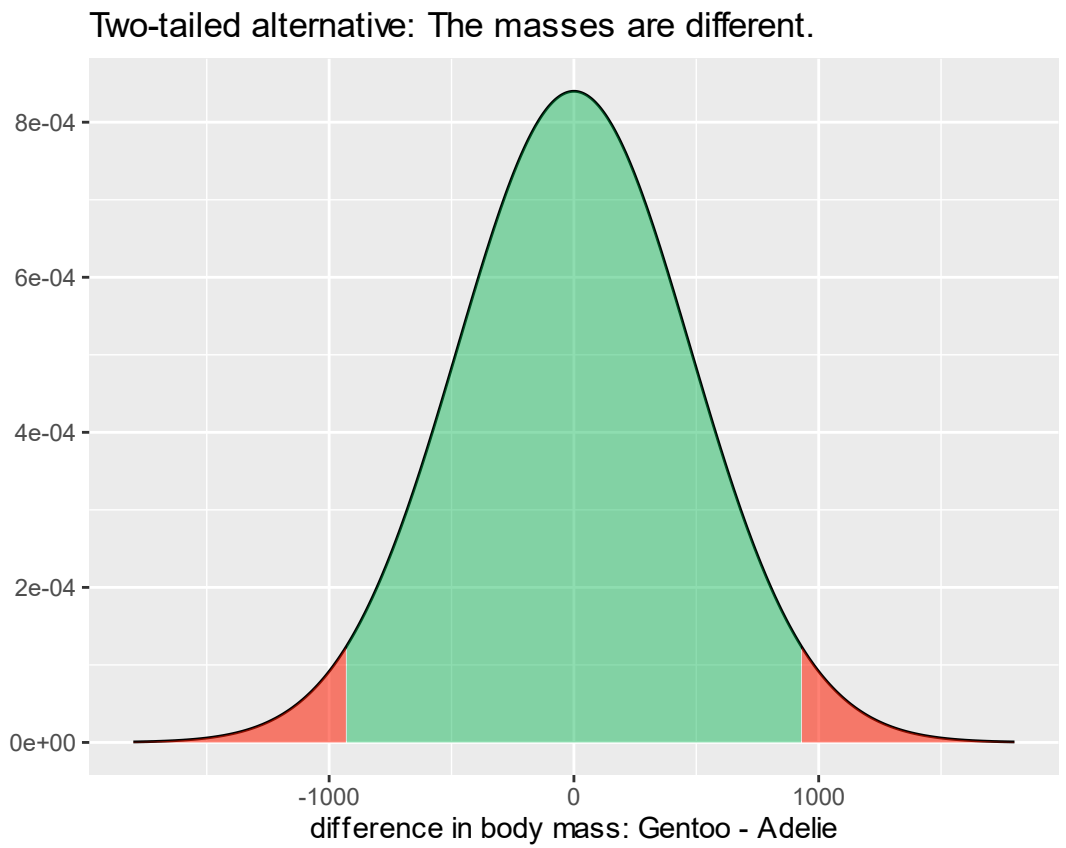
Question	Simple Null Hypothesis
• We want to know about differences between the two species.	Gentoo and Adelie penguins do not differ in body mass. A more technical, Frequentist formulation: • "The body masses of Adelie and Gentoo penguins are drawn from the same population of possible penguin body masses."

Alternative Hypotheses

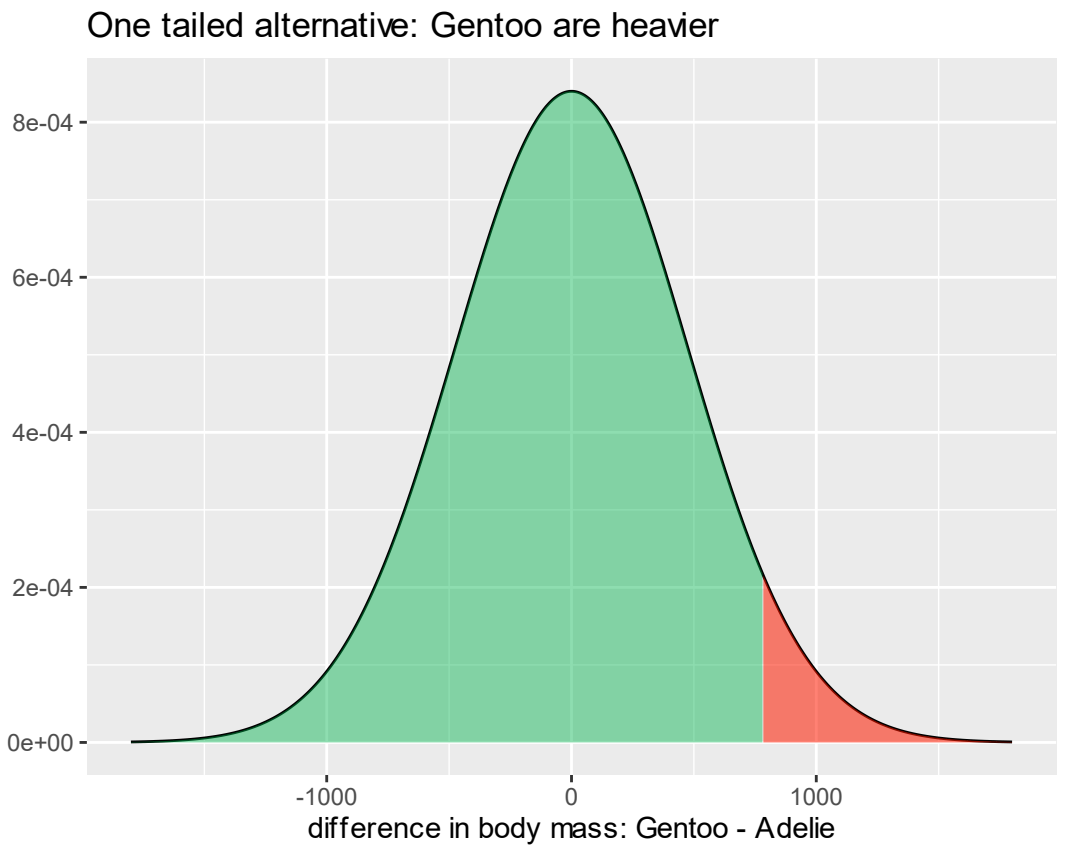
2-tailed hypotheses: non-directional	**1-tailed hypotheses: directional**
General: We don't propose a particular direction of the association. • We have no prior knowledge of the two species • We think that there will be *some* difference, but we don't know which one is heavier • **"Gentoo and Adelie penguins have different body masses."**	More specific than 2-tailed hypotheses. • We have some previous knowledge that allows us to propose a *direction* to the association. • From our experience in the field and previous research we think that Gentoo penguins are heavier • **"Gentoo penguins are heavier than Adelie penguins."**

Alternative Hypotheses

2-tailed hypotheses: non-directional



1-tailed hypotheses: directional

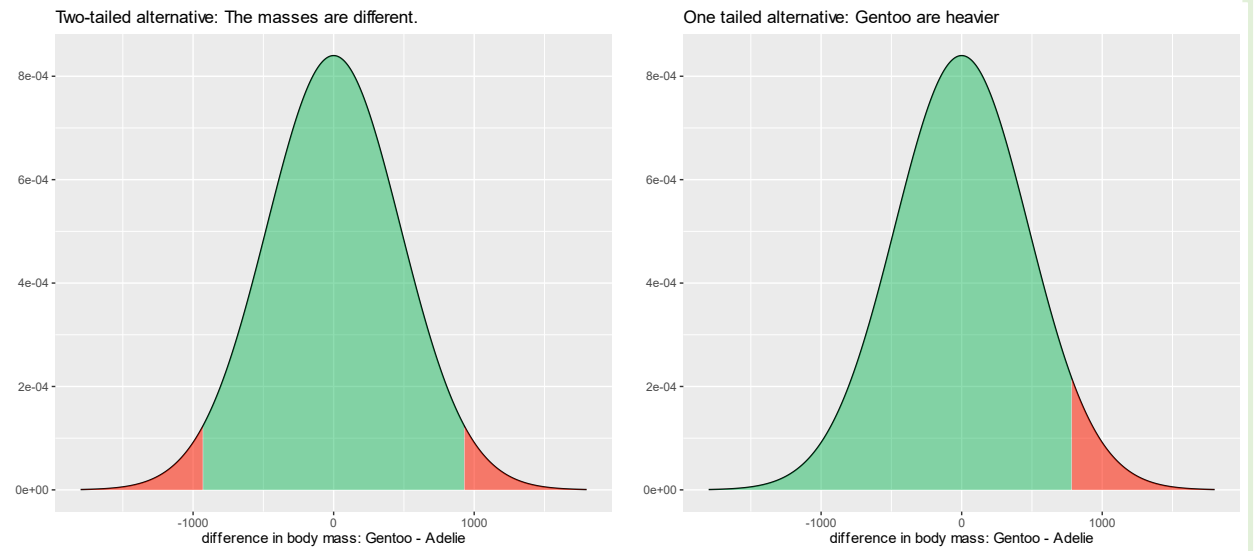


Decision criterion: tails and rejection regions

The area in red is α our Type I error rate, in this case 0.05.

1-tailed hypotheses can give us better *statistical power*.

- The tails define the *rejection region*, the region in which we say we have strong evidence to *reject the null hypothesis*.
- If the observed differences in masses falls in the rejection region, we can reject the null!



In-Class Probability 3

Calculating probabilities with R

Announcements

- Strategies for sample space questions:
  - All probability questions are hard. They feel like trick questions.
  - Draw a picture! What are the possible events in the sample space?
  - Consider whether events are combinations or permutations:
    - Permutation: 2-coin flips – heads followed by tails
    - Combination: exactly one tail
  - Think carefully about independent events: does knowing the outcome of one event give you any information about the other event?

A procedure for hypothesis testing

We'll walk through a simple example using the Palmer Penguin data

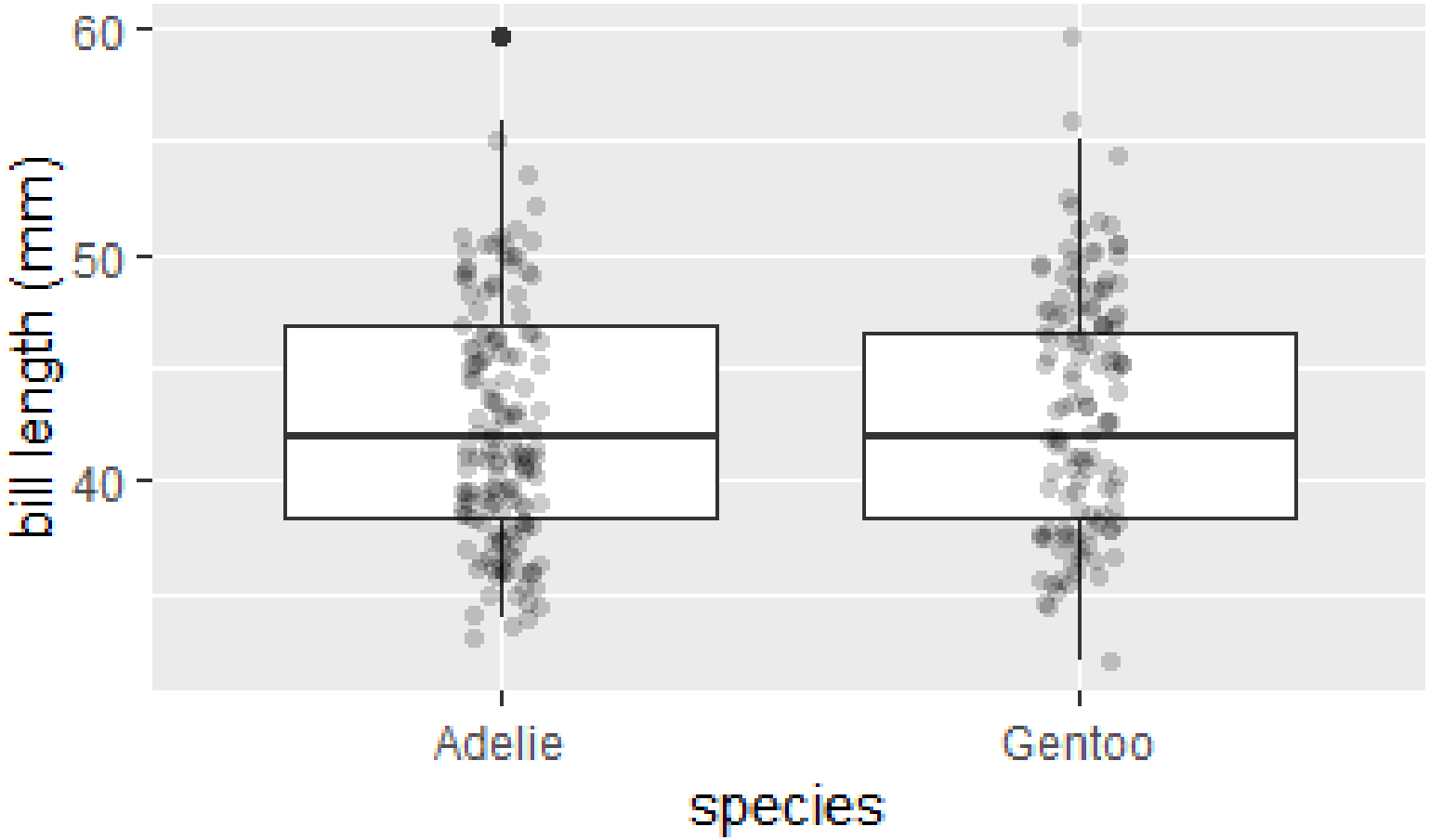
1. Propose a null hypothesis
2. Propose an alternative hypothesis
3. Select a decision criterion
4. Data curation
5. Analyze data
6. Evaluate model structure, assumptions, etc.
7. Reject or fail to reject the null hypothesis

- Iterate the last three steps until satisfied

Null Hypotheses: 2-Level Categorical Variable, 2 Tails

A simple scenario: Adélie and Gentoo Penguins' bill lengths

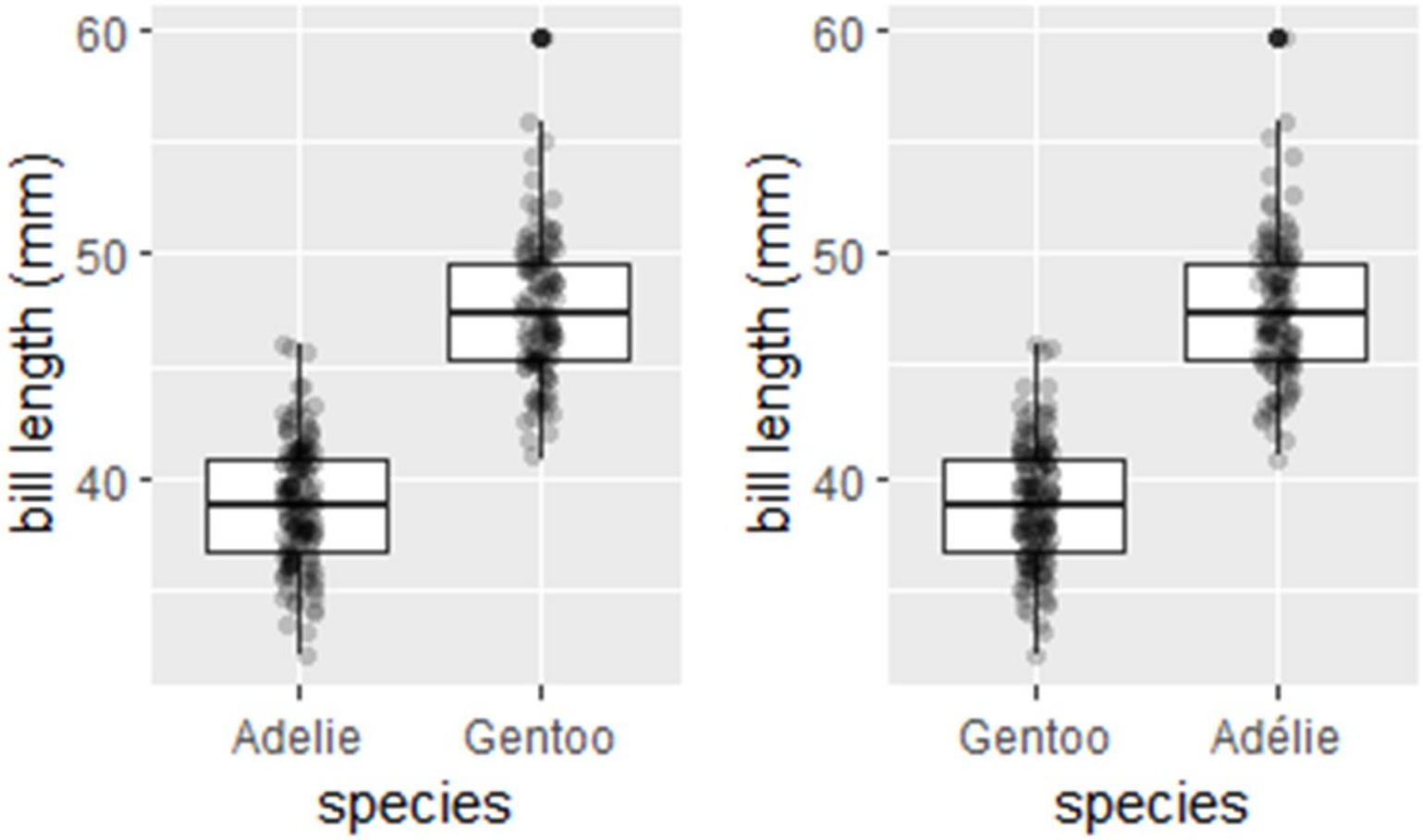
A simple null hypothesis: the bill length does not vary between the two penguin species.



Alternative Hypotheses: 2-Level Categorical Variable, 2 Tails

A simple scenario: Adélie and Gentoo Penguins' bill lengths

A simple alternative hypothesis: we propose that the bill lengths are different, but we don't know which one should be larger.

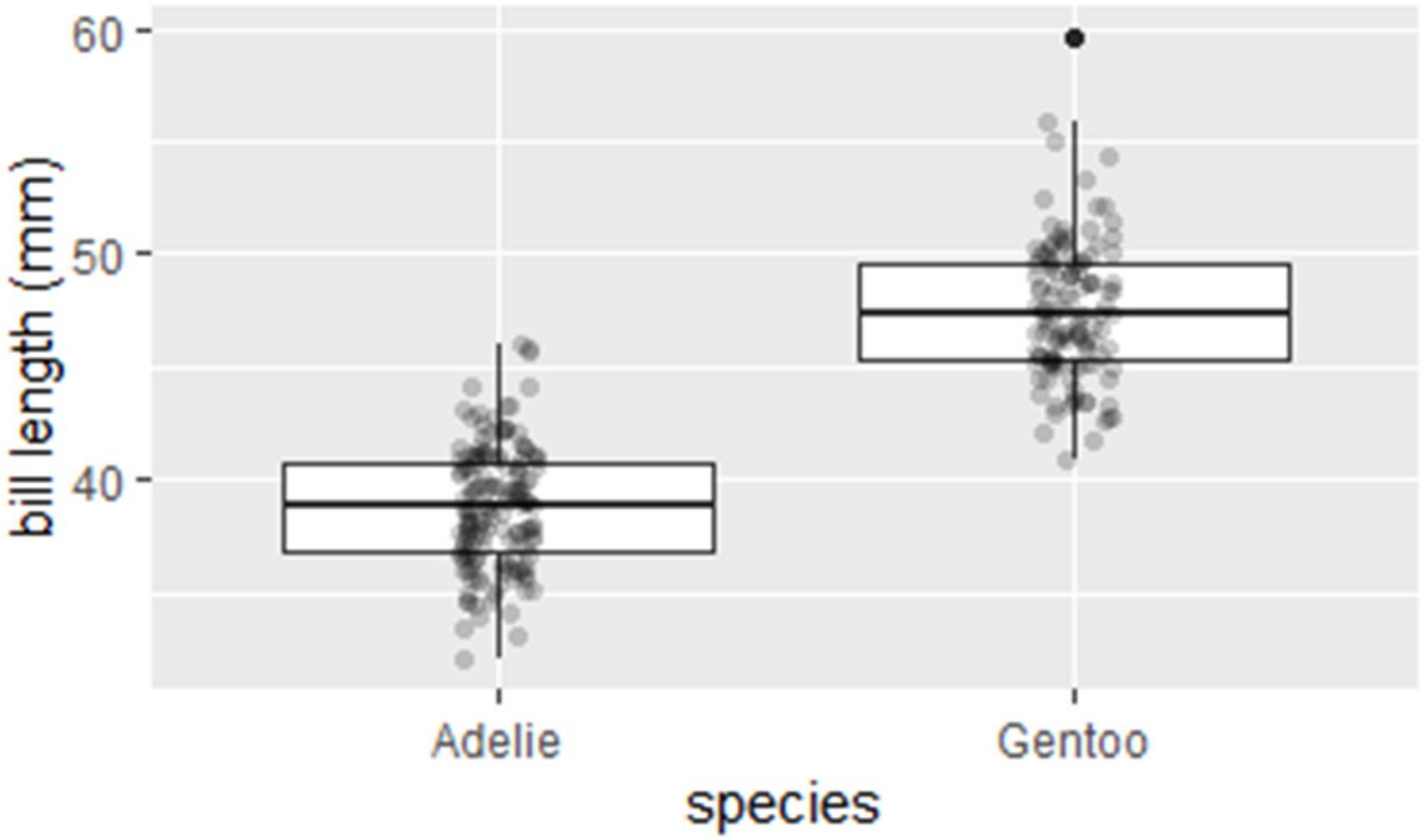


Alternative Hypotheses: 2-Level Categorical Variable, 1 Tail

A simple scenario: Adélie and Gentoo Penguins' bill lengths

A directional alternative hypothesis: we think Gentoo penguins have larger bills.

NOTE: I can't decide to do a directional hypothesis after I look at my data. It must be based upon prior knowledge. This is very important!



Decision Criteria and Statistical Errors

Decision criterion

How should we choose a decision criterion?

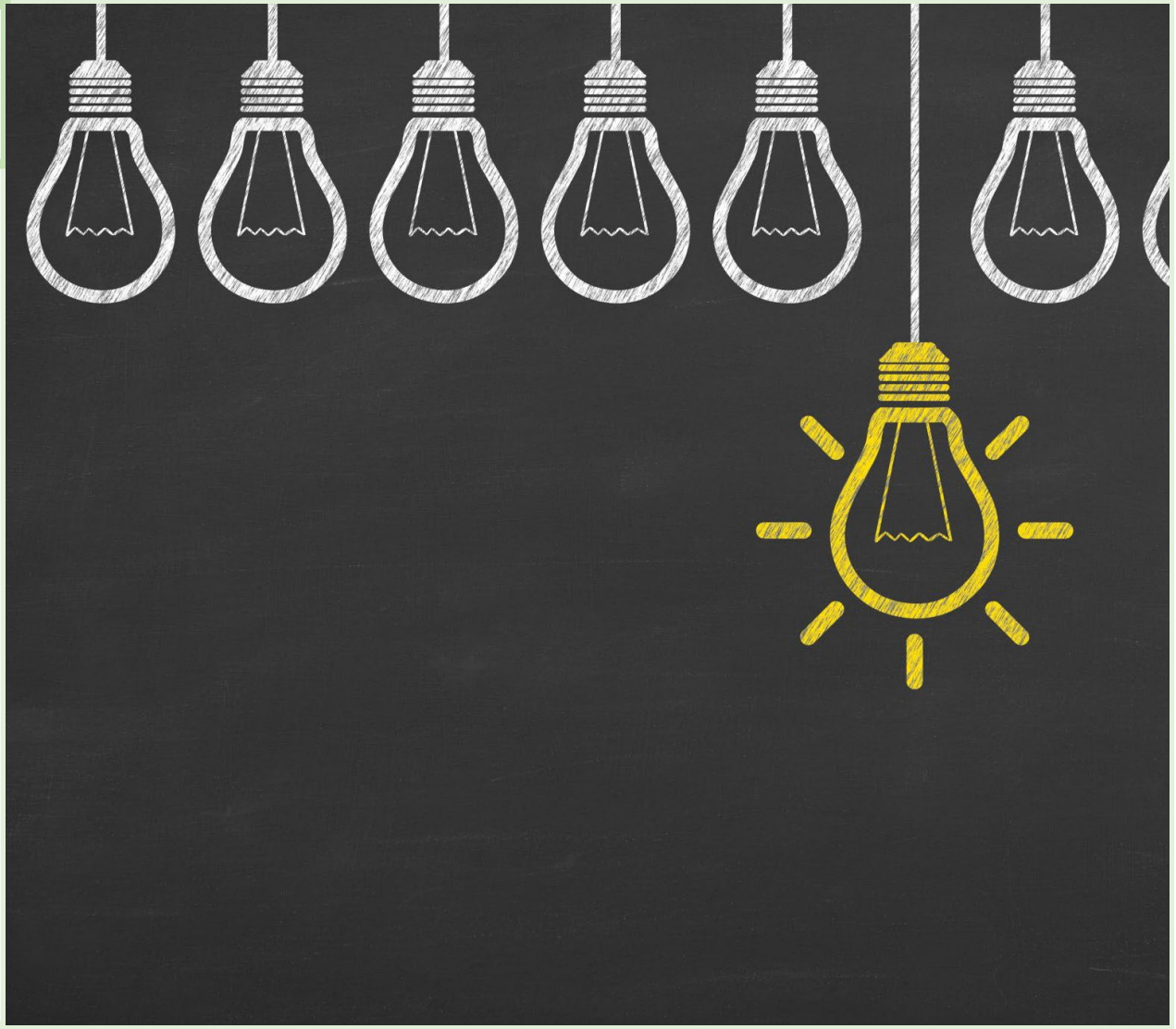
We've already done the following:

1. Propose a null hypothesis
2. Propose an alternative hypothesis

We would like to quantify how likely any patterns data could have occurred *by chance alone*.

Parametric distributions are very powerful tools, if we can approximately meet the required assumptions.

We need to decide upon a criterion *before* we carry out the experiment! This is usually a 5% false-positive rate.

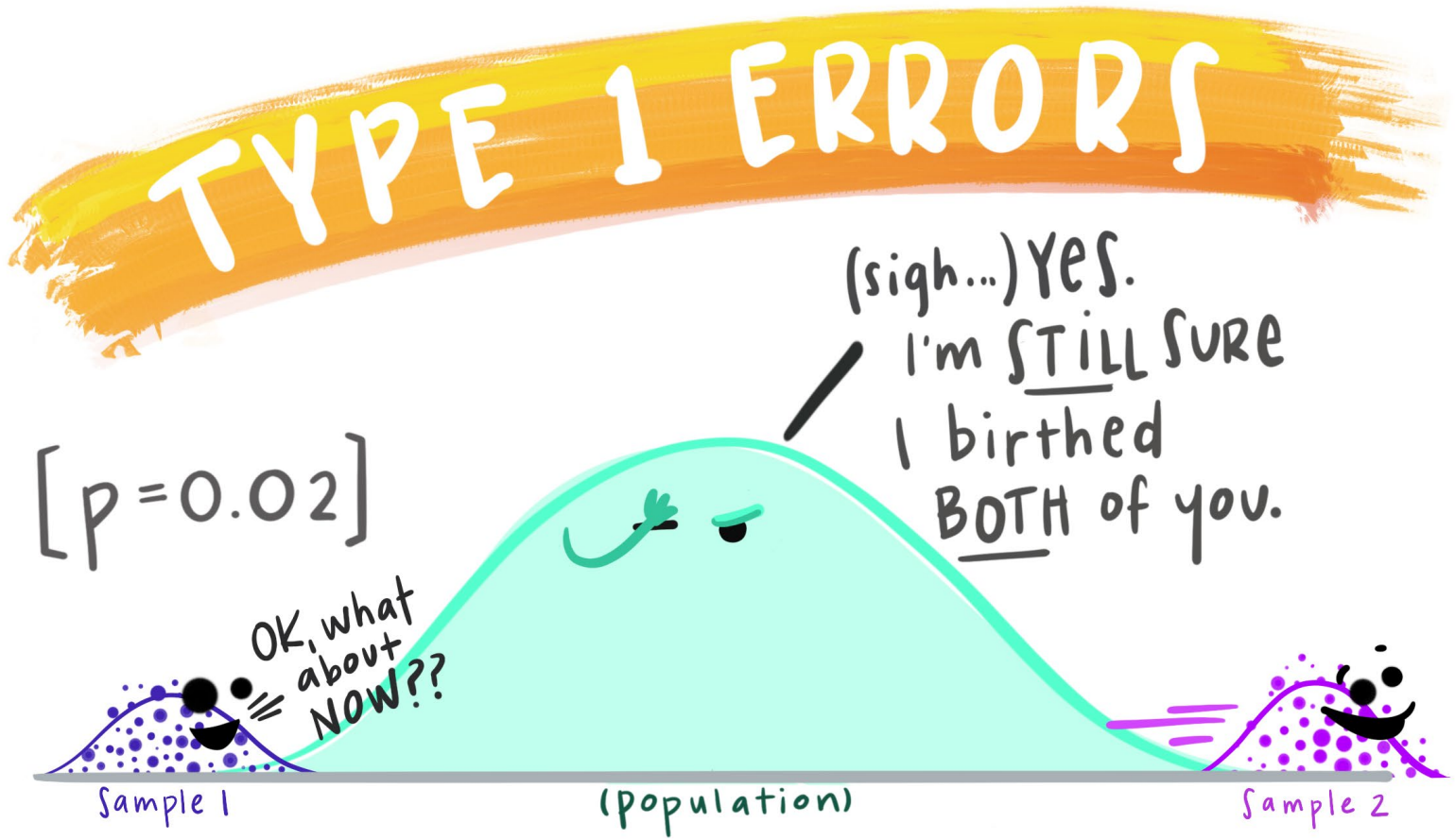


Decision criterion: false positive rate

In Frequentist inference, we often use the Type I Error Rate as a decision criterion.

You can think of it as the false positive rate.

We observed something *weird*, simply by chance (sampling error).



Artwork by @allison_horst

Type I errors

I like to think of Type I errors as *false positives*.

They happen when we falsely conclude something interesting is happening.

- This can happen from *sampling error*: when our random sampling efforts happen to collect a very non-representative sample.
- We decide ahead of time what level of *false positives* we are willing to accept.
- This is our *significance level*, also known as the Type I error rate.

Decision criterion: alpha

In a Frequentist analysis, we decide on a false positive rate, α, that we can live with. Often 5%.

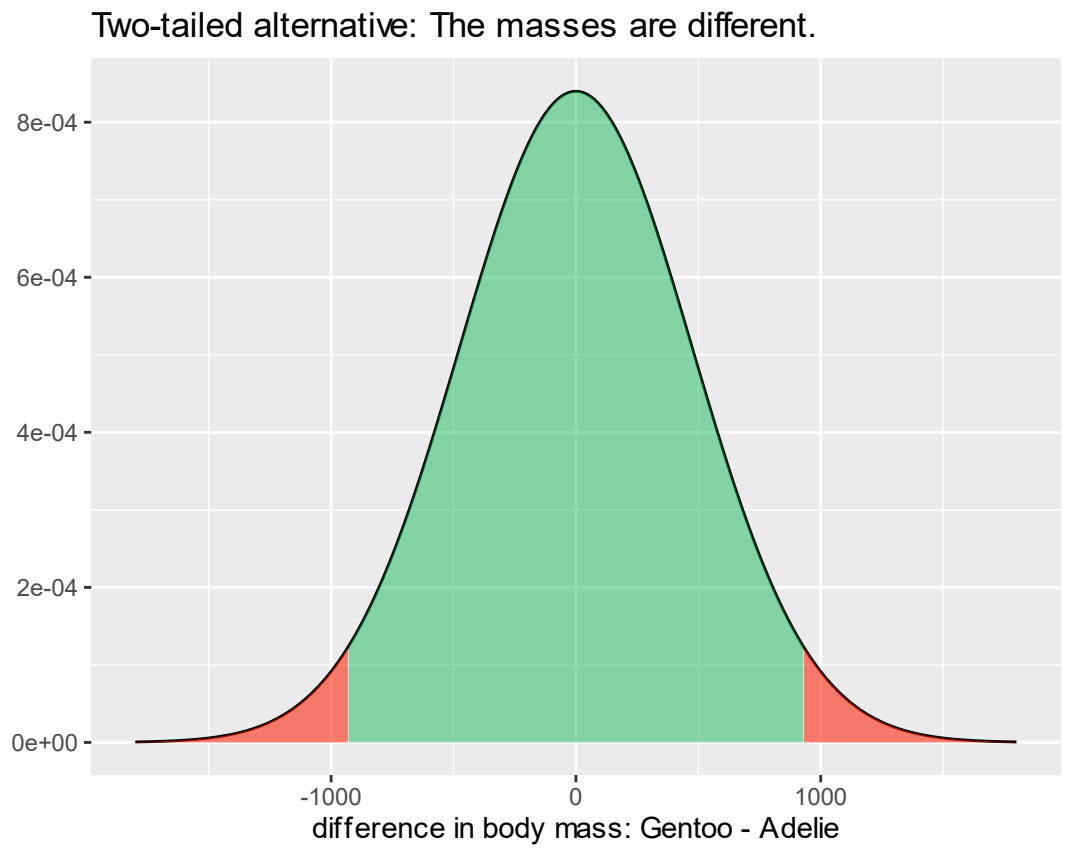
Test statistic

- When we perform our analysis, we will calculate the value a *test statistic* that we use to compare against our chosen parametric distribution.
- The test statistic reflects how likely our data were to have occurred by chance alone.
- If the test statistic resides in a *tail* of the distribution, we have strong evidence to reject the null hypothesis.
- We can calculate a p-value from the test statistic and the distribution.

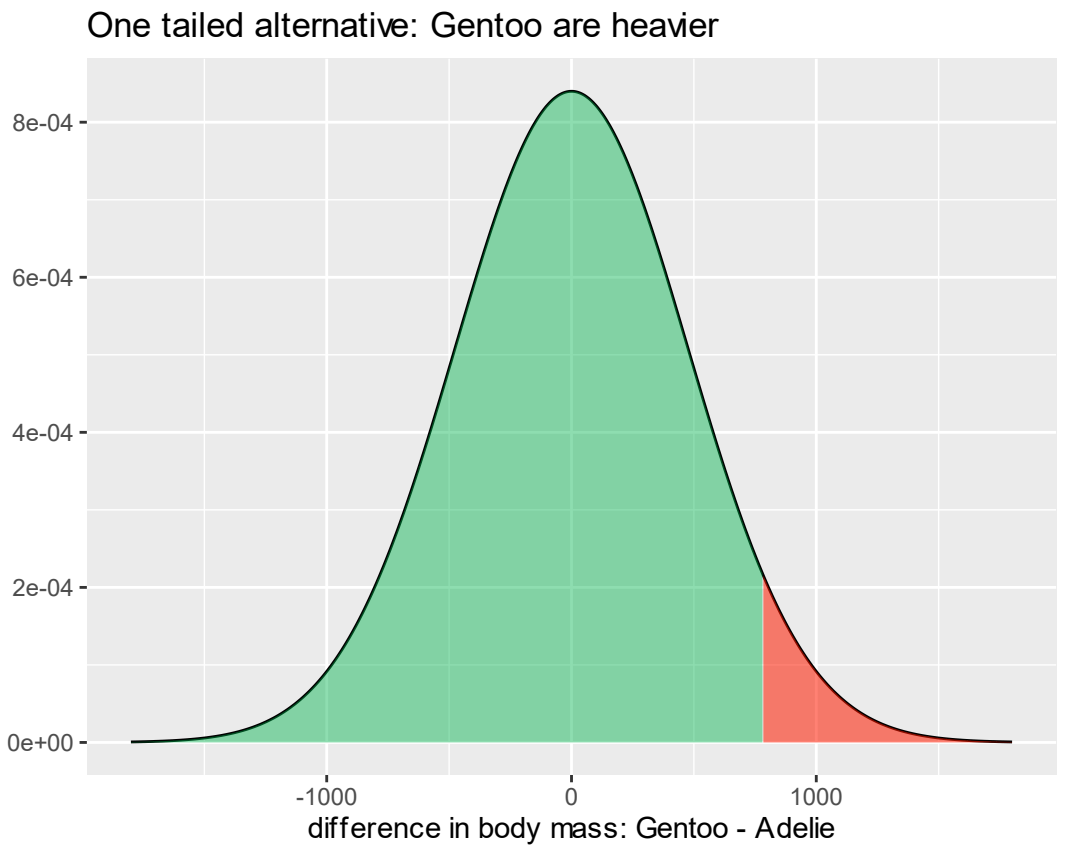
Let's go with a 5% significance level for our example test.

Distribution Tails and the Decision Criterion

2-tailed hypotheses: non-directional



1-tailed hypotheses: directional



Penguin T-test

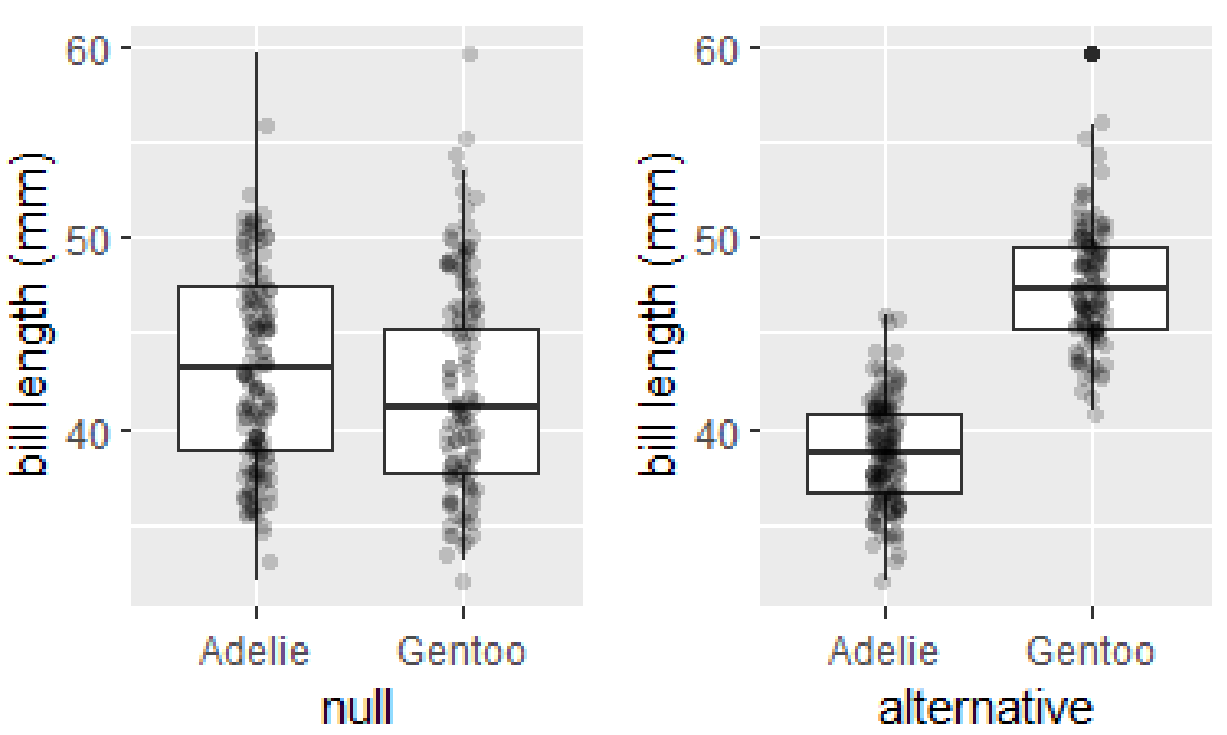
Penguin Bill Lengths	Hypothesis Boxplots (alternative plot shows the real data)

The alternative boxplot shows the real data.

Let's calculate the means:

	species	mean bill length
1:	Adelie	38.79
2:	Gentoo	47.50

- It looks like a difference of about 8 mm.
- Based on the boxplots and our mean calculations, do you think there's a real difference in bill length?



Penguins: Analyze data

A 2-sample t-test is a good match for our question.

- It's easy to conduct the test in R:

```
t.test(
    bill_length_mm ~ species,
    data = subset(penguins, species != "Chinstrap"),
    alternative = "less")
```

R's t-test output

```
        Welch Two Sample t-test

data:  bill_length_mm by species
t = -24.725, df = 242.58, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Adelie
 group Gentoo is less than 0
95 percent confidence interval:
      -Inf -8.131593
sample estimates:
mean in group Adelie mean in group Gentoo
            38.79139             47.50488
```

Welch Two Sample t-test: `bill_length_mm` by `species`

R's t-test output

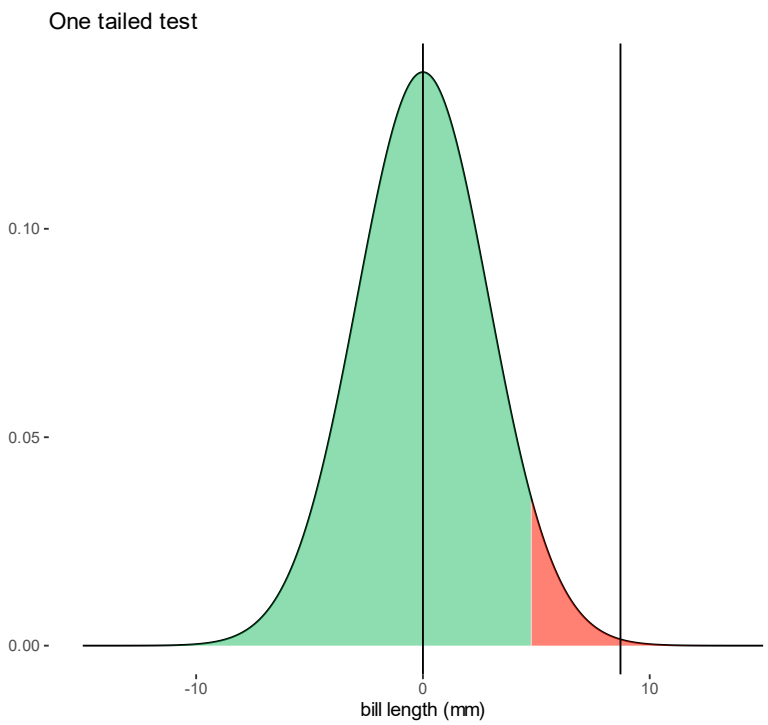
T-test output key points	Assess evidence against the null
• The value of the t-statistic: -24.73 • Degrees of freedom: 242.6 - This is related to sample size and model complexity • We'll talk lots more about degrees of freedom in the next few weeks! • P-value is less than our chosen α	• Based on our graphical, numerical, and statistical evidence, do you think Gentoo penguin have larger bills than Adélie penguins?

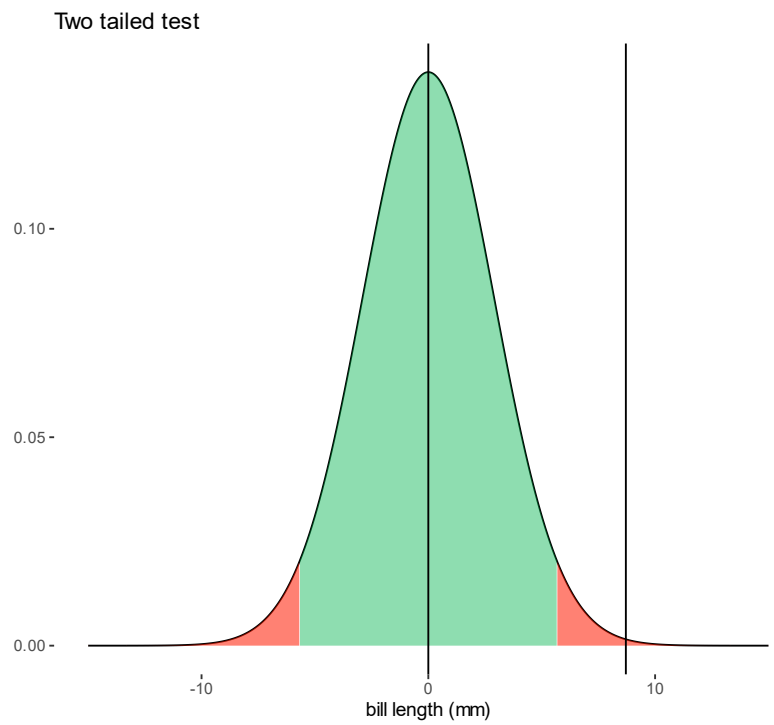
T-tests and tails

Interpreting the Test

- Vertical lines: mean bill length for the 2 penguin species
- Red fill: rejection region
- Difference in means does not have to be as large for 2-tailed alternatives!

We could have rejected the null with either the 1- and 2-tailed alternatives





Key Concepts and Terms

- Null Hypotheses
- Alternative Hypotheses
- Decision Criterion
- 1- and 2-tailed hypotheses
- T-test: applying the decision criterion

- T-Test: General procedure
- Visual intuition of null and alternative hypotheses

Recap: procedure for hypothesis testing

1. Propose a null hypothesis
2. Propose an alternative hypothesis
3. Select a decision criterion
4. Data curation
5. Analyze data
6. Evaluate model structure, assumptions, etc.
7. Reject or fail to reject the null hypothesis

- Iterate the last three steps until satisfied

Preparing for Confidence Intervals

Sample Standard Deviation and The Sampling Distribution

What's in This Section?

Slides	Take-Home Concepts
• Distribution of the sample • Sampling distribution • Sample standard deviation • Standard error	• What is a sampling distribution? • Understanding the difference between the distribution of observations in a sample and the sampling distribution. • Factors that affect the sampling distribution.

Sampling Distributions

What is a Sampling Distribution?

- It describes the distribution of a sample statistic (like the mean) if we could make repeated samples. It's a probability distribution (It approaches a t- or z- distribution in fact).
- Sampling distributions are **very** important, but often misunderstood.
- Sampling distributions are **not** the same as the distribution of a population variable.

Sampling Distributions

Sampling distributions depend on:

- The sample size
- The population standard deviation (and therefore the sample standard deviation)

Each sample statistic has a sampling distribution, and a standard error.

- We usually work with the sampling distribution and standard error of the mean.

Parameterizing the sampling distribution

We already know:

- The sampling distribution is a probability distribution of a *sample statistic*.
- For sample sizes > 30, the sampling distribution approaches a normal distribution.
- This is *very* useful for inference.

We can treat the sampling distribution like a Normal distribution

- It's a 2-parameter distribution: mean and standard deviation
- The mean is the population mean (or our estimate of it)
- The standard deviation is the *standard error*.
  - Standard error is a sample size-adjusted version of the sample standard deviation

Standard error of the mean: intuition

We would like to know the population mean and standard deviation, but we know that in the frequentist paradigm we assume these are *unknowable*.

Some intuition questions:

- What could we do to improve our estimates of the population mean?
- Do you think there would be greater variability in the means in repeated samples of 5 or 50?
- What do you think happens to the sample standard deviation as you increase the sample size?

Standard error of the mean: intuition

Intuition from sample size

As sample size grows, our estimates of the population parameters get better.

- If we took repeated samples of 5 observations, the sample means would bounce around due to *sampling error*.
- If we took repeated samples of 500 observations, the sampling error would be smaller and the sample means would be closer together.
- In other words, with increasing sample size our sample means **stabilize** around the true population mean.
- With increasing sample size, the sample standard deviation **stabilizes** around the population standard deviation.

Sample size and the standard error

Recall the sample variance: $var(X) = s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

- The sample standard deviation is just the square root: $s_x = SSD = \sqrt{(var(X))}$

The standard error is the SSD adjusted for sample size:

$$SSE = \frac{s_x}{\sqrt{n}}$$

The following are key findings:

- The standard error **gets smaller** as the sample size increases!
  - It's adjusted 'twice' for the sample size.
- The sample standard deviation **stabilizes** as the sample size increases!

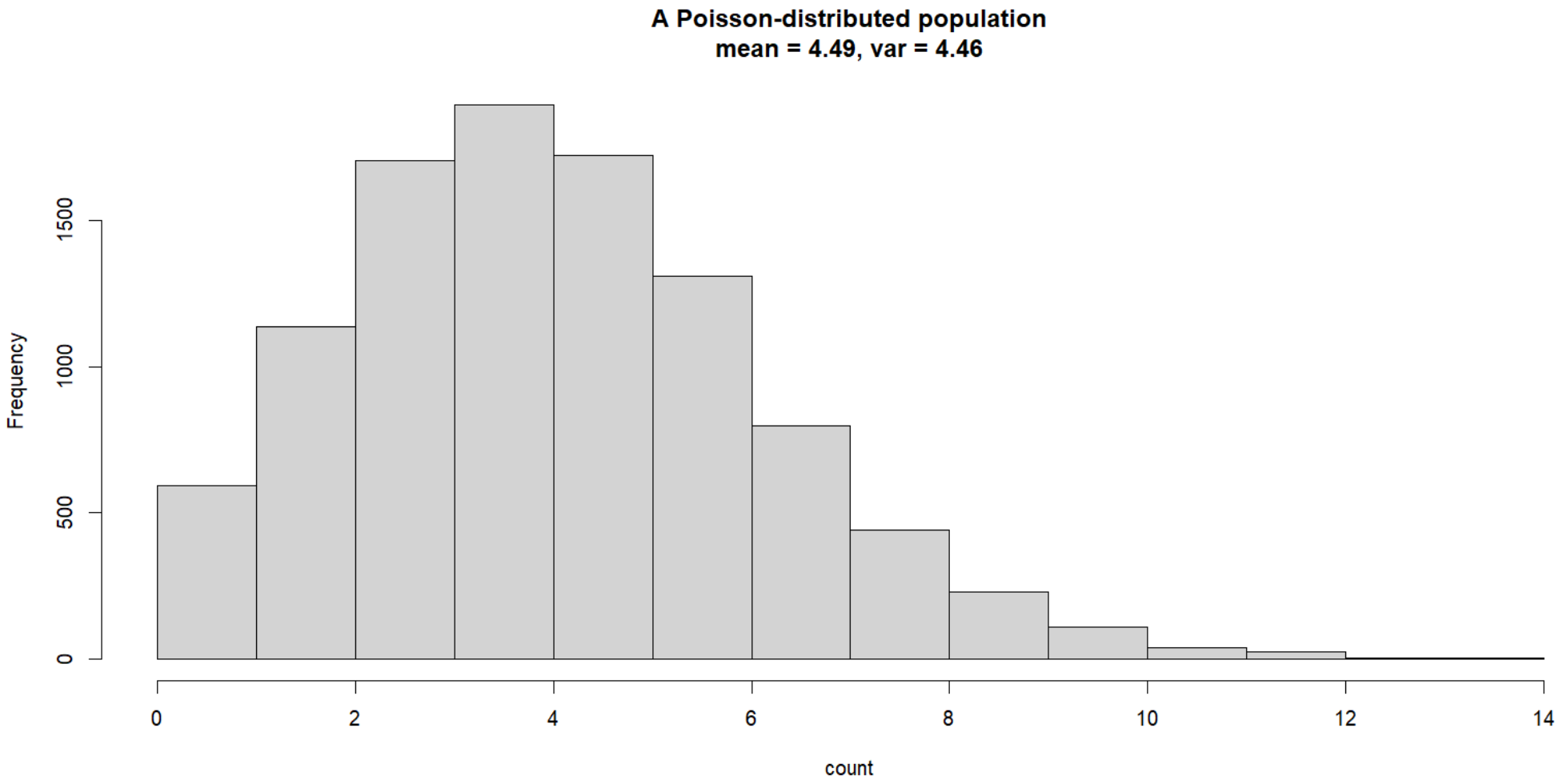
Intuition: simulation experiments

This is technical material. Let's try to build some intuition with graphical examples.

A non-normal distribution: a Poisson-distributed population

- Create a population of 10,000 individuals.
- Do repeated sampling and calculate the mean.
- Examine the *sampling distribution* of the mean.

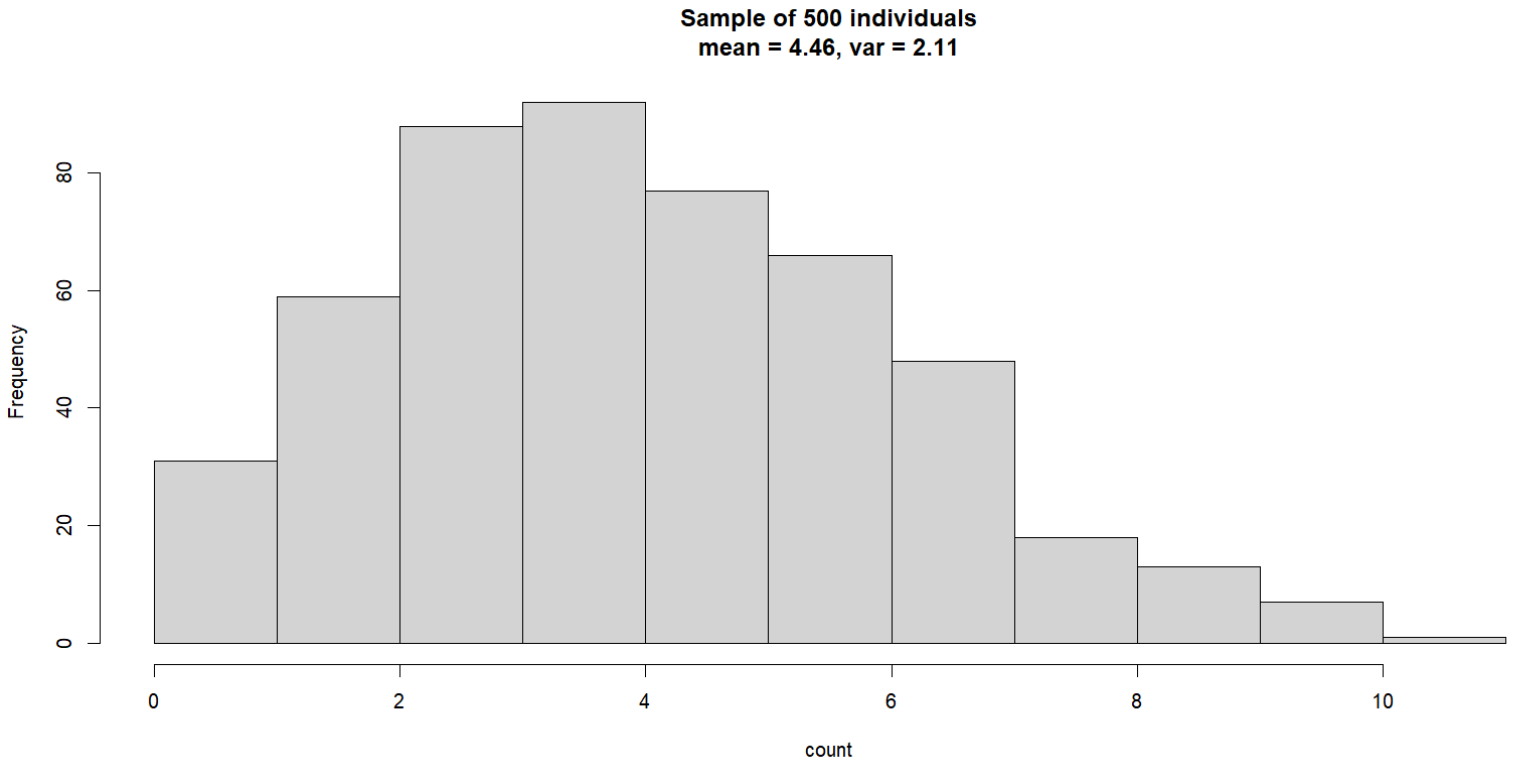
10000 individuals, lambda = 4.5



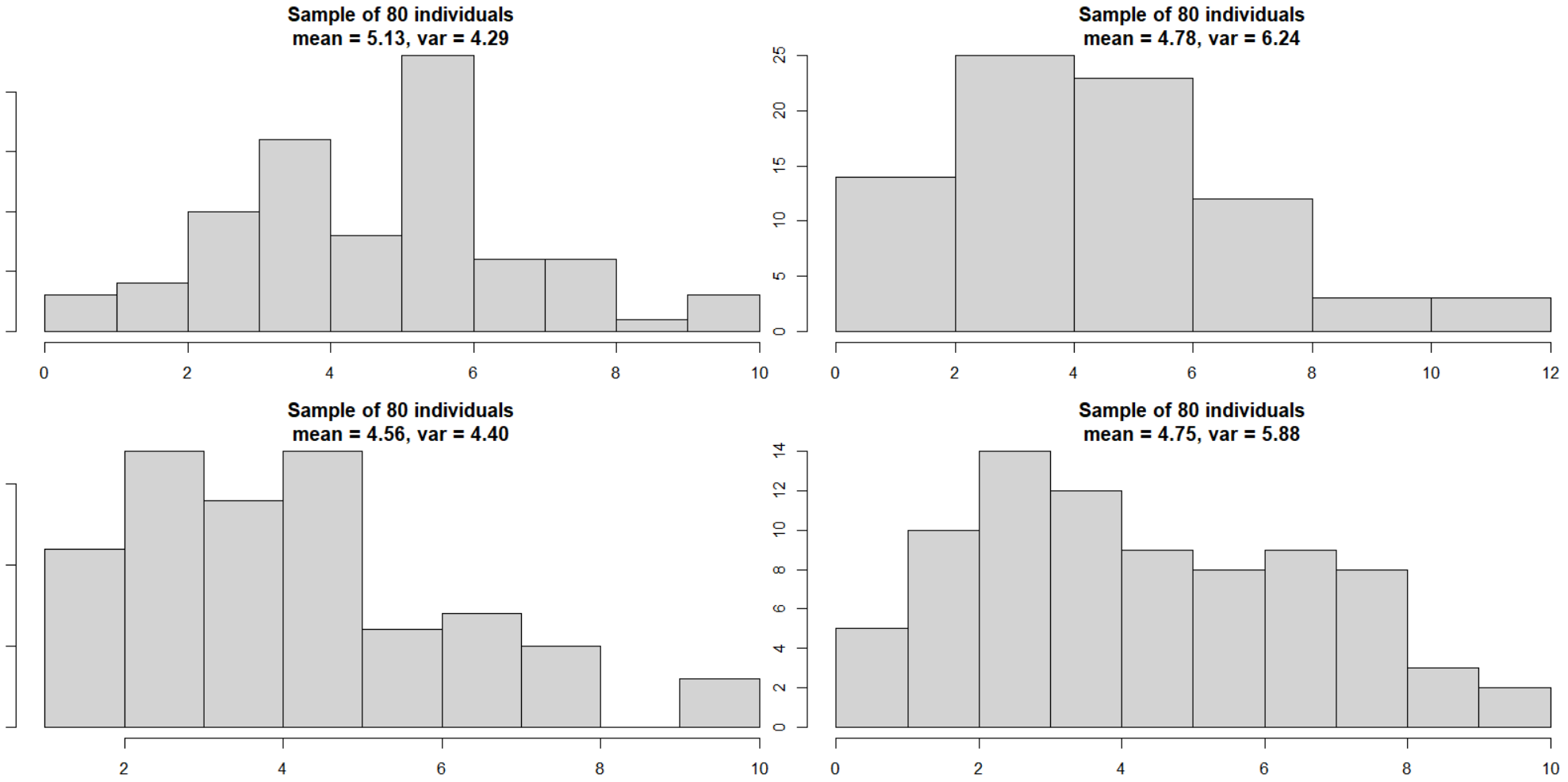
Sample from the population

Intuitively, we might think a sample would have a similar distribution to the population.

Let's take a sample of 500:



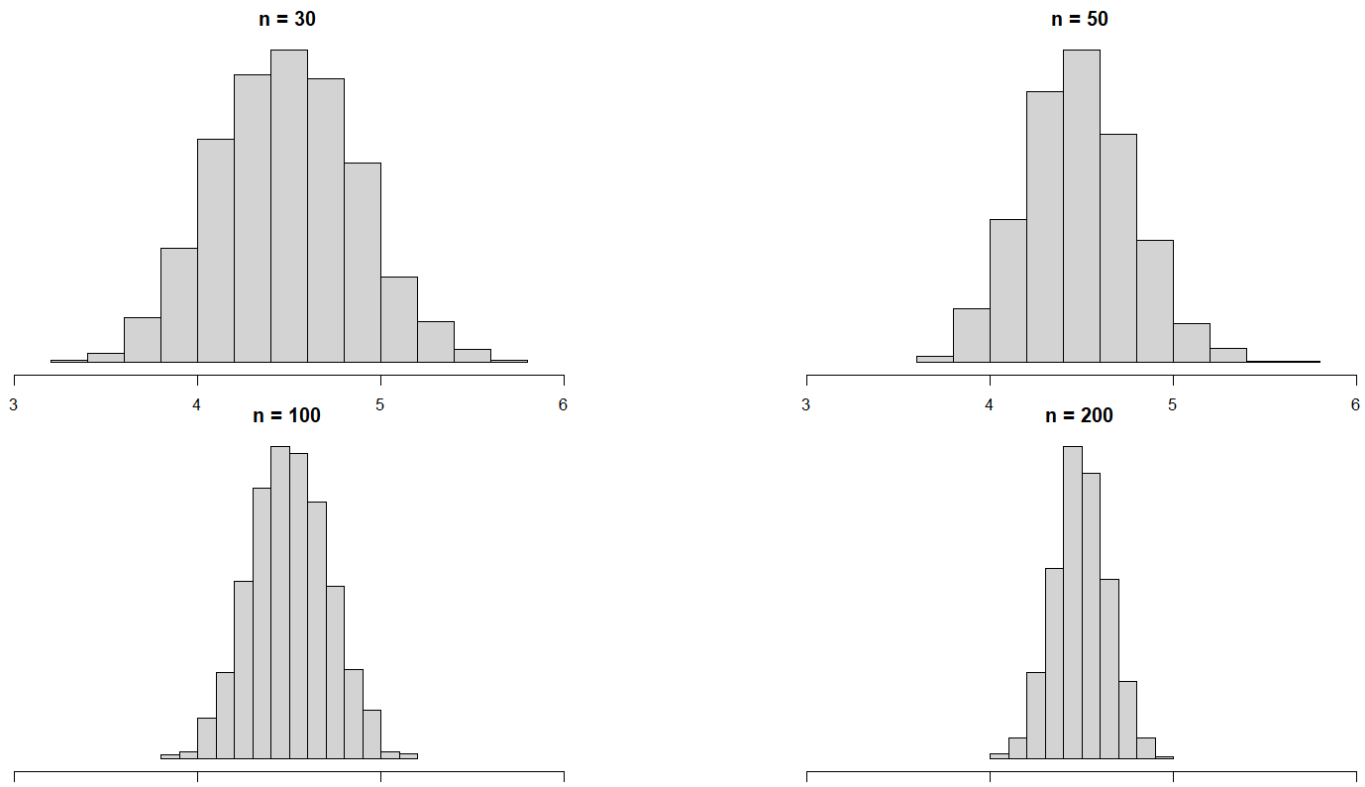
Sample from the population: 80 samples



Sampling Distribution of the Mean

What should the distribution of the *means* in repeated sampling look like?

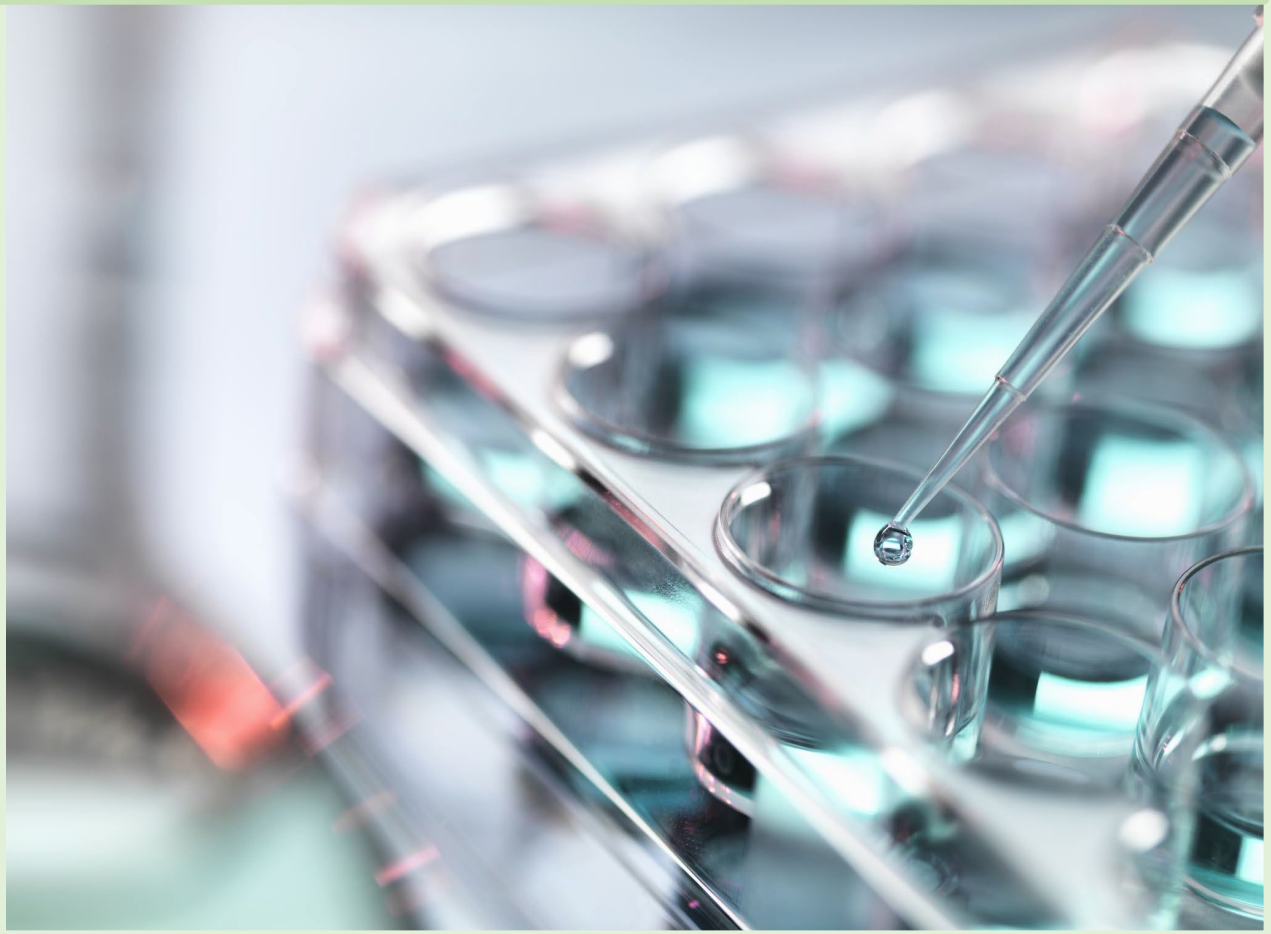
- The standard error gets smaller with increasing sample size.
- The sample standard deviation (and other sample statistics) stabilizes with increasing sample size.
- Confidence intervals are calculated from standard errors: CIs get narrower with larger samples!



Key Concepts and Terms

Sampling distribution and distribution of the sample

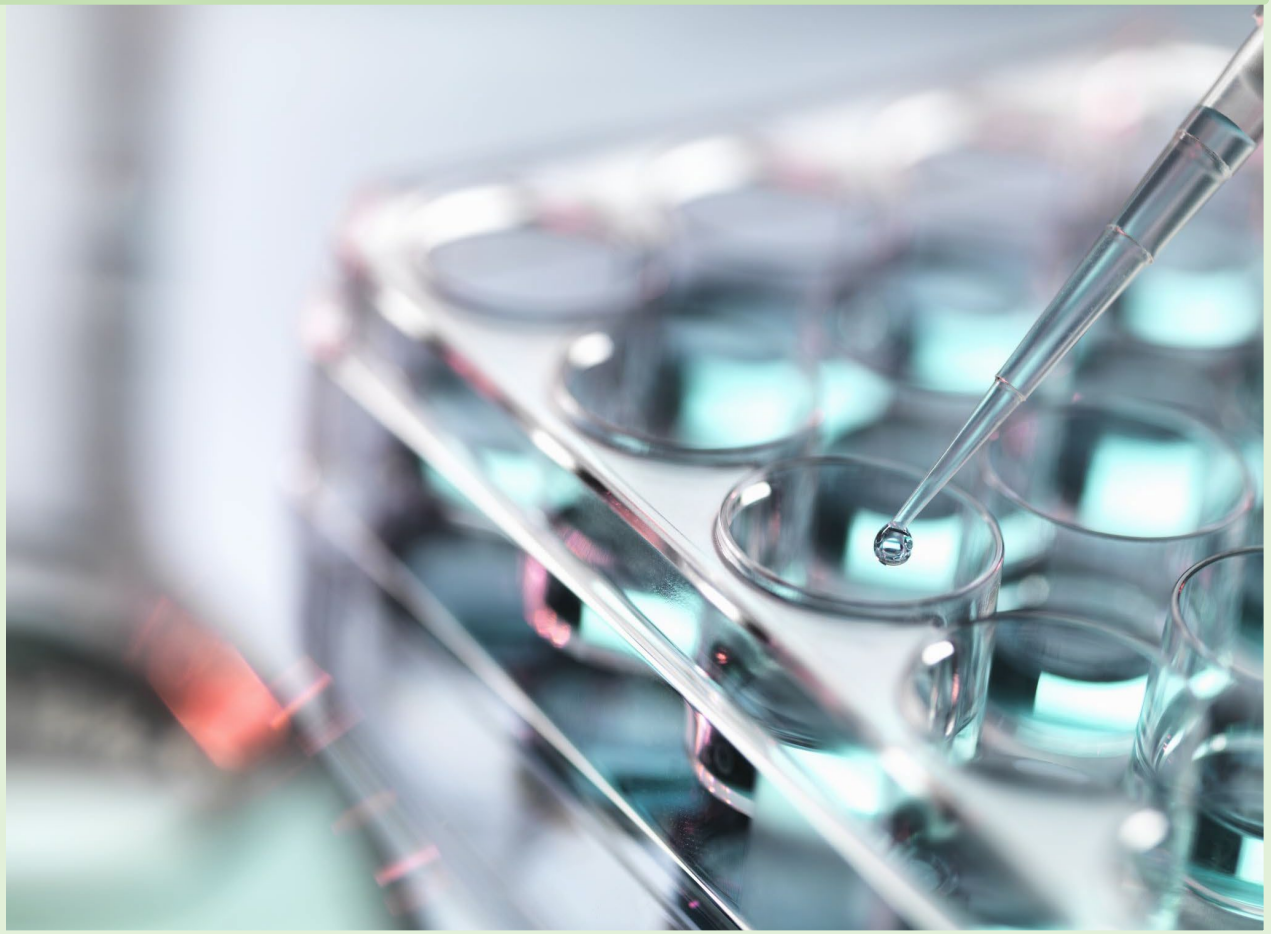
- Distribution of the sample
- Sampling distribution
- Sample standard deviation
- Standard error



Key Concepts and Terms

Sampling distribution and distribution of the sample

- What affects the width of the sampling distribution?
- What happens to the sample standard deviation as sample size increases?
- What happens to the standard error as the sample size increases?



In-Class Stuff: Probabilities 3 + Babel

Announcements

- Dreary weather today, but fall has been lovely.
- Updated deck 5.
- Questions for me?
- What did you think of the Library of Babel?

Confidence Intervals 1

Background

What's in This Section?

Slides	Take-Home Concepts
• Confidence Intervals • Calculating Confidence Intervals	• Population distribution • Sample distribution • Sampling distribution • Standard error • [Hypothetical] repeated sampling • Why the sampling distribution is the real key player.

Outline

Confidence intervals aren't that interesting…

- This slide deck focuses on building up to CIs from the underlying *sampling distribution*
- If you understand the sampling distribution, you've got the tools to understand much of frequentist inference (not just the confidence interval part).

Reminder: sample size and the standard error

Recall the sample variance: $var(X) = s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

- The sample standard deviation is just the square root: $s_x = SSD = \sqrt{(var(X))}$

The standard error is the SSD adjusted for sample size:

$$SSE = \frac{s_x}{\sqrt{n}}$$

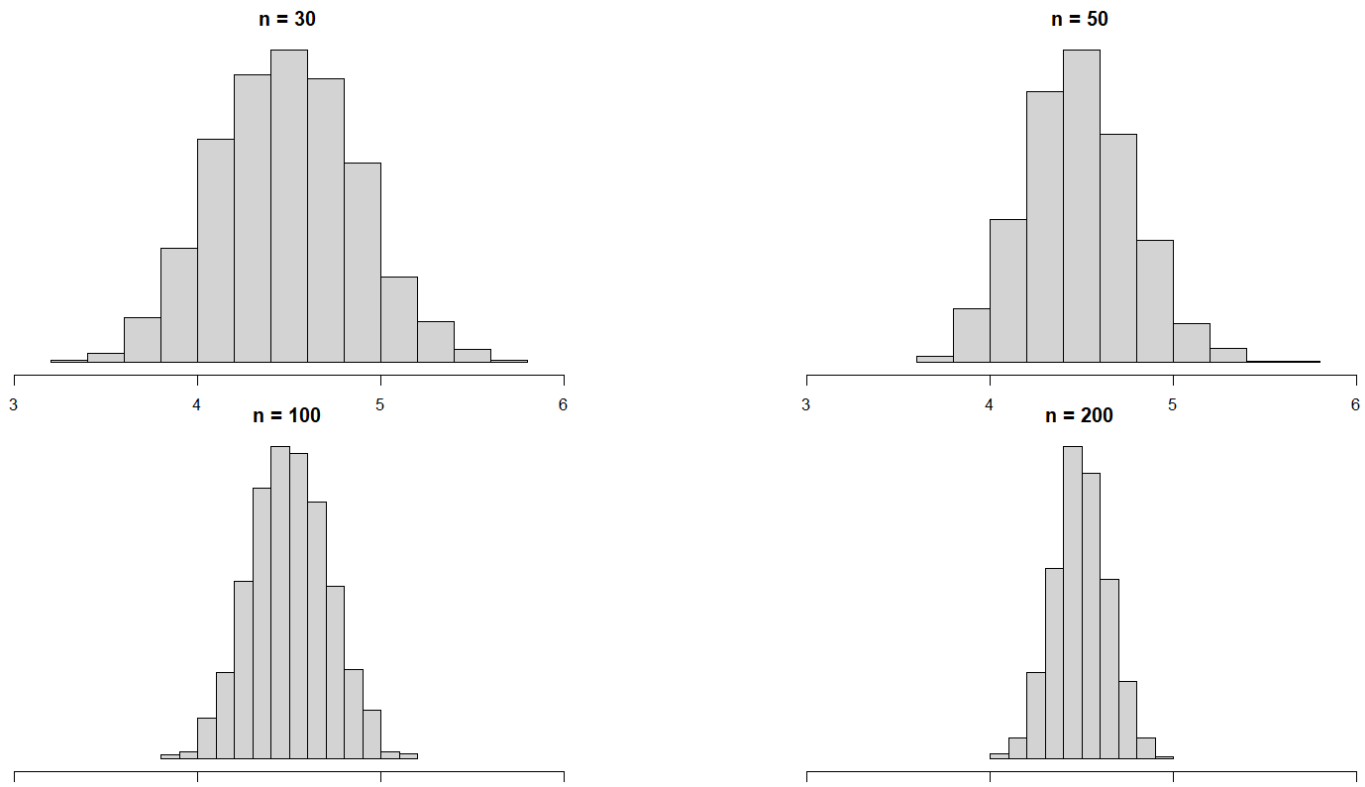
The following are key findings:

- The standard error **gets smaller** as the sample size increases!
  - It's adjusted 'twice' for the sample size.
- The sample standard deviation **stabilizes** as the sample size increases!

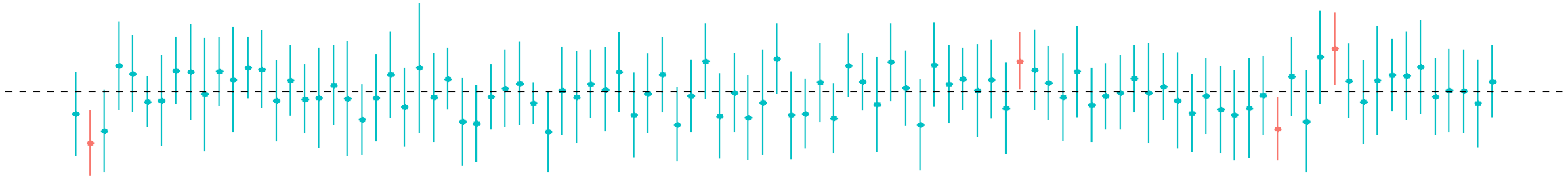
Sampling Distribution of the Mean

What should the distribution of the *means* in repeated sampling look like?

- The standard error gets smaller with increasing sample size.
- The sample standard deviation (and other sample statistics) stabilizes with increasing sample size.
- Confidence intervals are calculated from standard errors: CIs get narrower with larger samples!



Note on terminology: frequentist *confidence* and *significance*



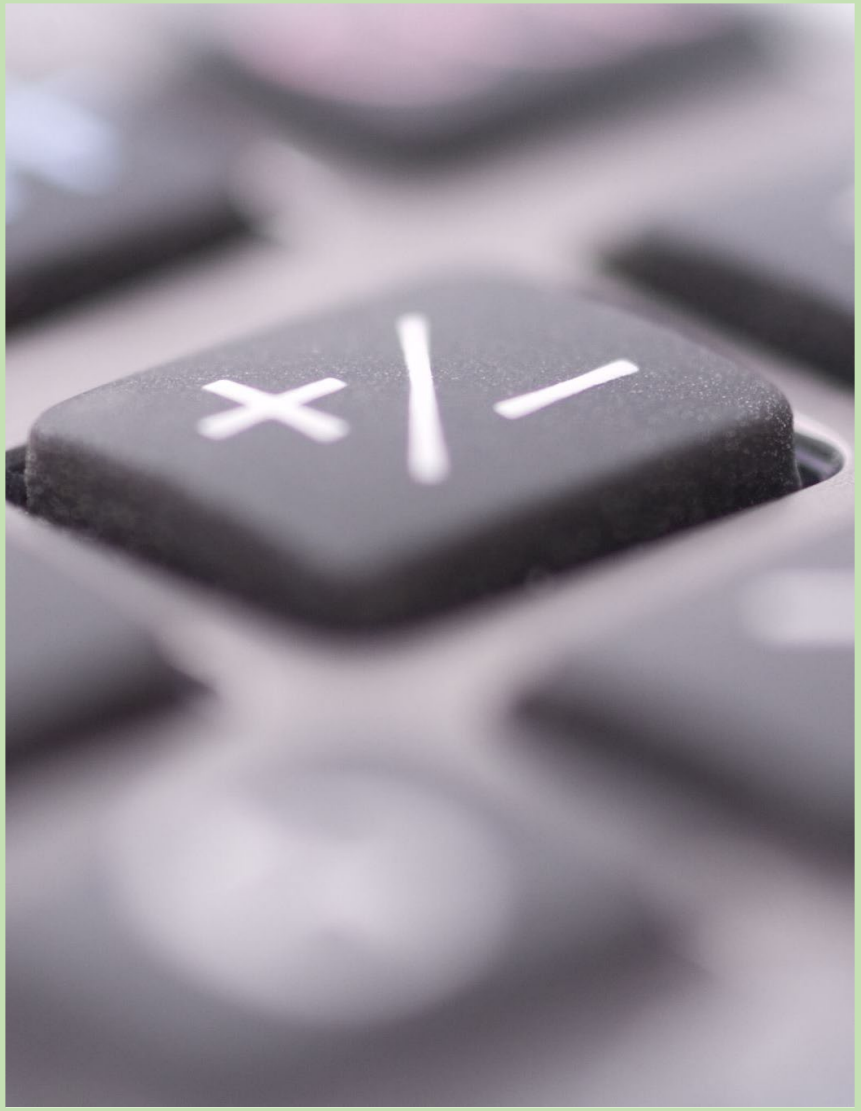
- Constructing a 95% confidence interval does **not** mean you are 95% sure your interval contains the true value!
  - It either does or does not, but you can't know.
- "If I were to repeat the experiment many times, approximately 95% of the CIs I construct would contain the true population parameter"

Sample standard deviation

Sample Standard Deviation: a measure of the spread of data

- "The square root of the average squared deviation from the mean."
  - That's a lot. We'll temporarily get rid of the outer square root to simplify the concept.
- We'll look at variances first. Standard deviation is just the square root.
- Population variance: $var(X) = \sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$
- Sample variance: $var(X) = s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

Sample variance



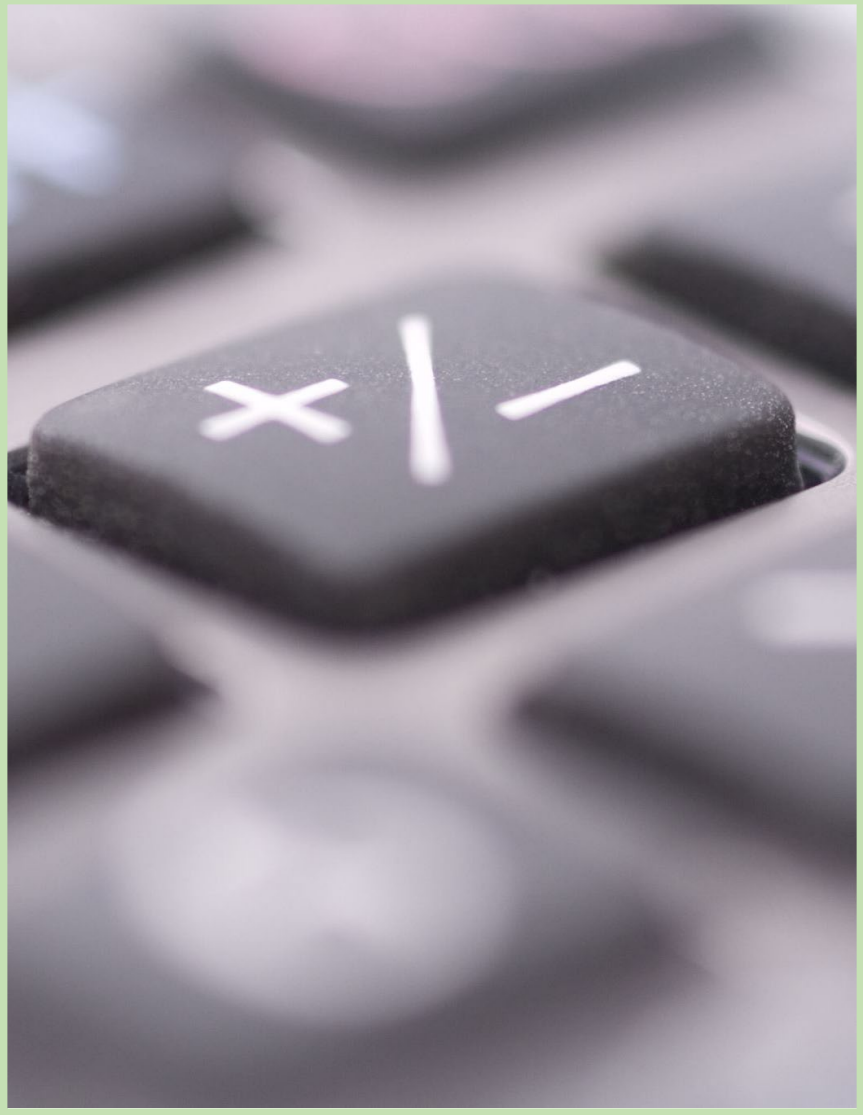
Let's dissect the formula for some insight

$$var(X) = s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

We recognize some parts:

- A residual: $x_i - \bar{x}$ and squared residual $(x_i - \bar{x})^2$
  - Residual: difference between an observed value and an expected value
  - Why do we square the residuals?
- Sum of squared residuals: $\sum(x_i - \bar{x})^2$
- Sample size: $n - 1$

Sample variance



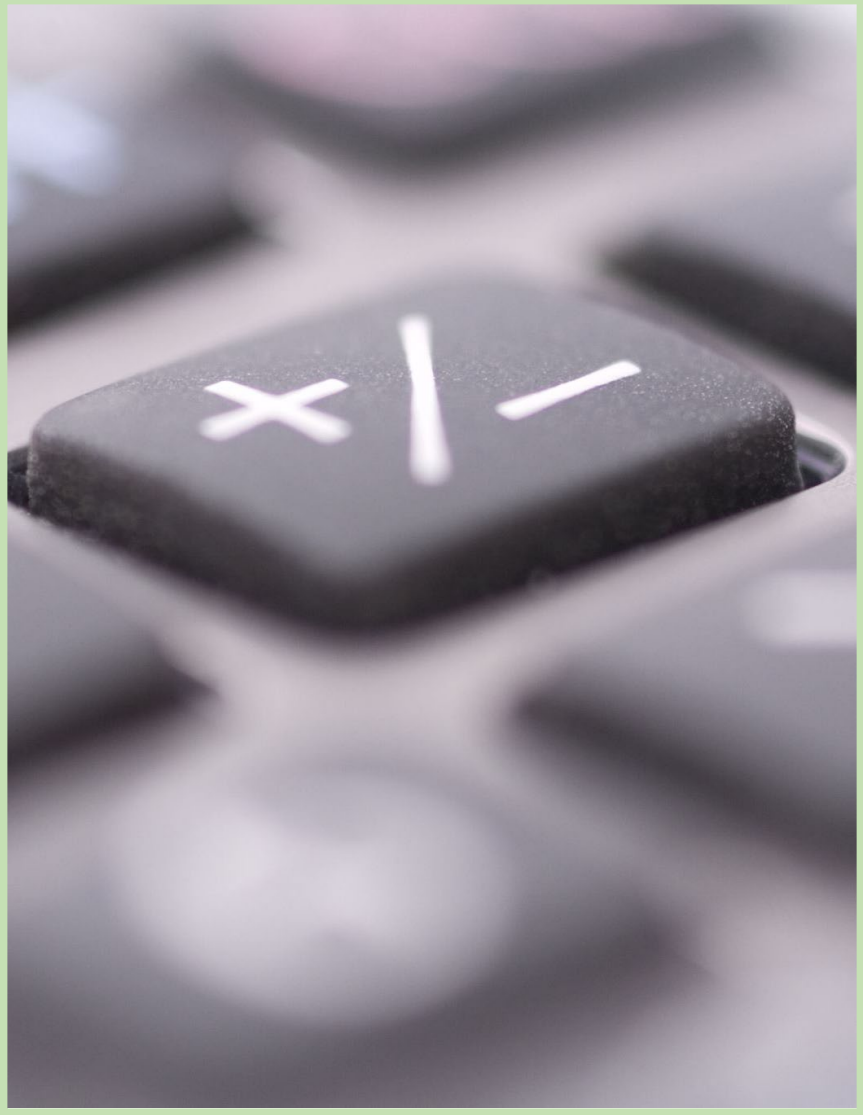
Let's dissect the formula for some insight

$$var(X) = s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

In words:

- **A residual:** the difference between what we observe and what our model predicts.
- **Squared residual.** Squaring results in:
  - All values become positive.
  - Large residuals contribute a **lot** more to the sum than small values.

Sample variance



Let's dissect the formula for some insight

$$var(X) = s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

- Large residuals contribute *much* more to the sum than small residuals.
  - Large residuals are highly **influential**, or highly **penalized** when we use sums of squares for optimization.
- Normalizing by the population size:
  - Variance can be interpreted as *a measure of* how much observed values differ from expected values.

Sample variance: normalizing by the sample size

Normalization

Sum of residuals and population size: The sum of the squared residuals is *normalized* by the population size.

- It's like the *average* squared residual.

This may seem unimportant, but *normalizing* is key to gaining insight into the sampling distribution.

- By normalizing the sum of squared residuals by the sample size, it means that as sample size grows, the sample variance *stabilizes* around a fixed value (hopefully the population value).

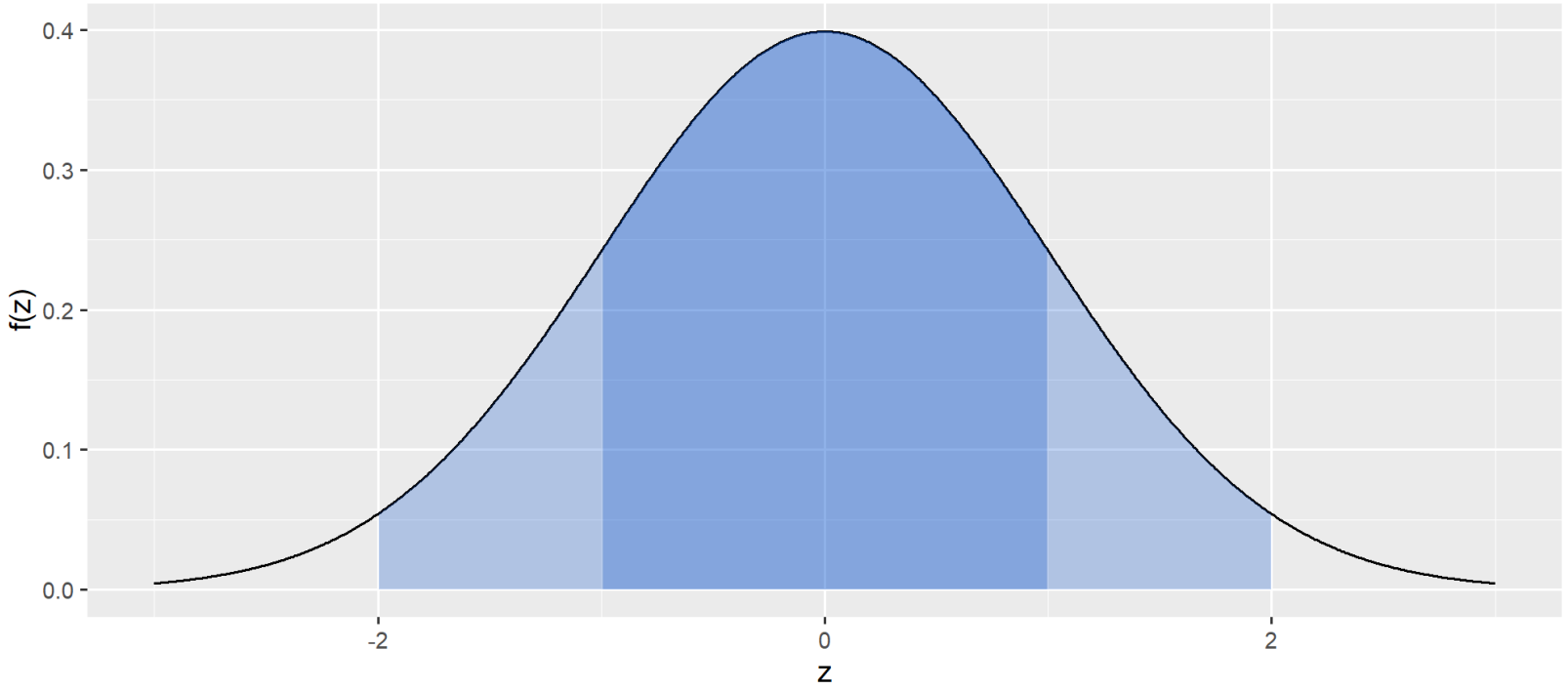
Standard deviation: It's just the square root of variance

Why transform by the square root function?

- What are the units of *variance*?

Standard deviation has a very nice interpretation for a *Standard Normal* distribution:

- 68% of observations are within 1 sd from the mean, 95% within 2 sd:



The sampling distribution

The sampling distribution is crucial to Frequentist inference

- Confidence intervals are cool and all, but the sampling distribution does all the work.

The standard error is the key to the sampling distribution

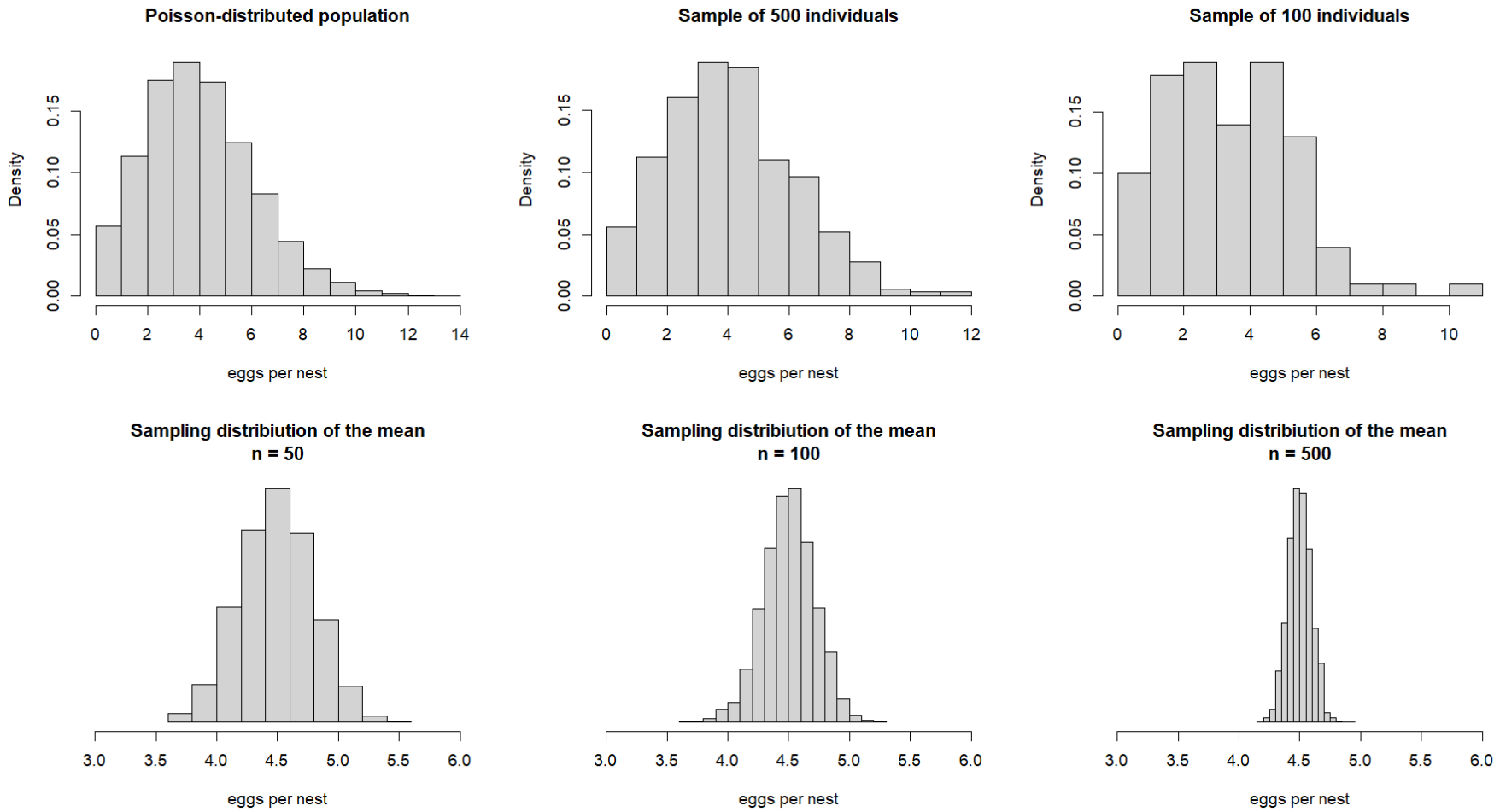
- The standard error describes how we feel about our estimate of the mean (or another *sample statistic*).

Standard error vs sample standard deviation

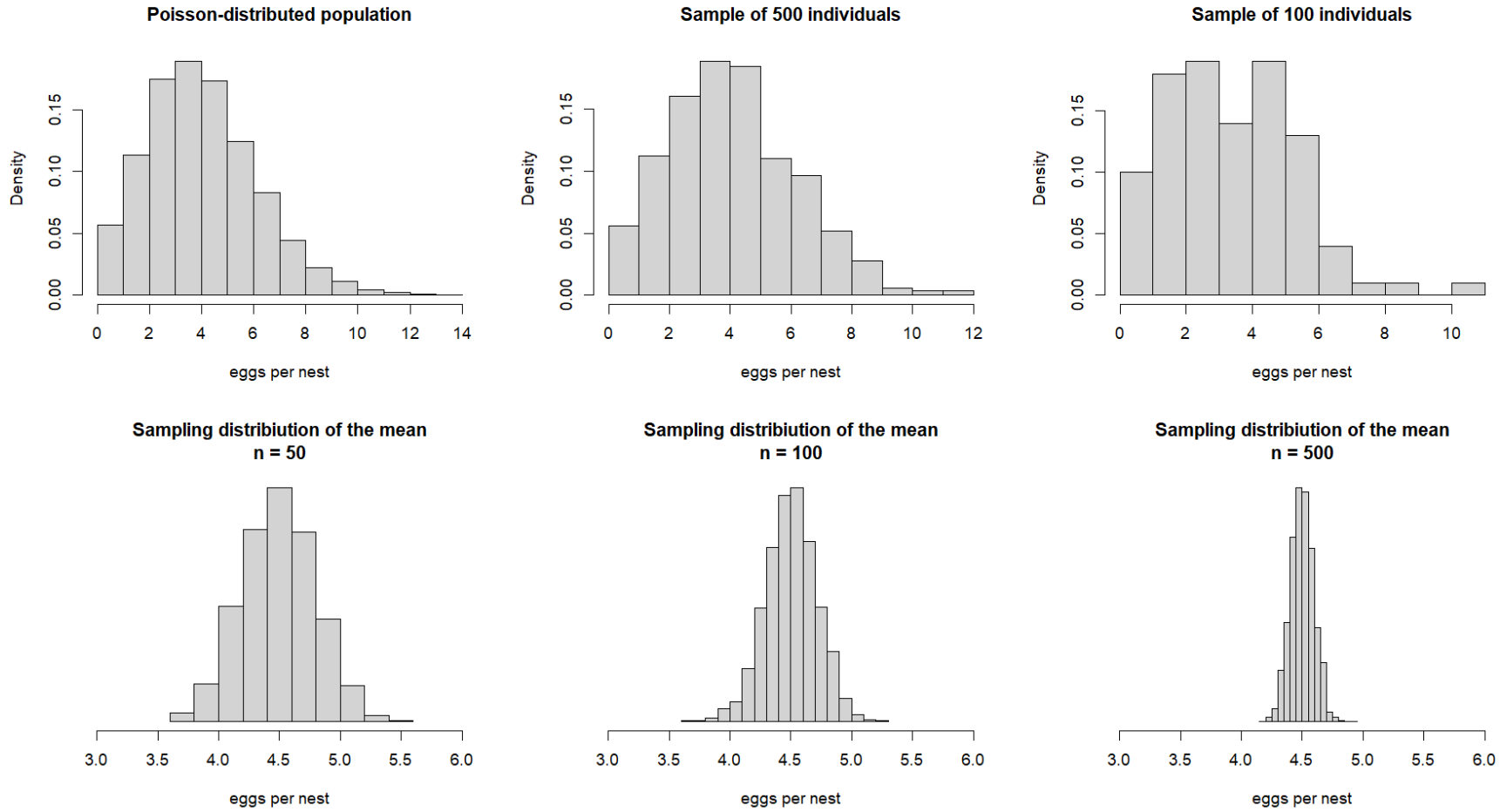
- Standard deviation describes how we feel about the **individuals in the population or sample** we have collected.
- Standard error describes how we feel about the **population of possible samples** we could collect.
- This is not a simple distinction. The similar terminology doesn't help.

Population, sample, and sampling distributions

Let's return to our graphic to build some intuition:



Population, sample, and sampling distributions



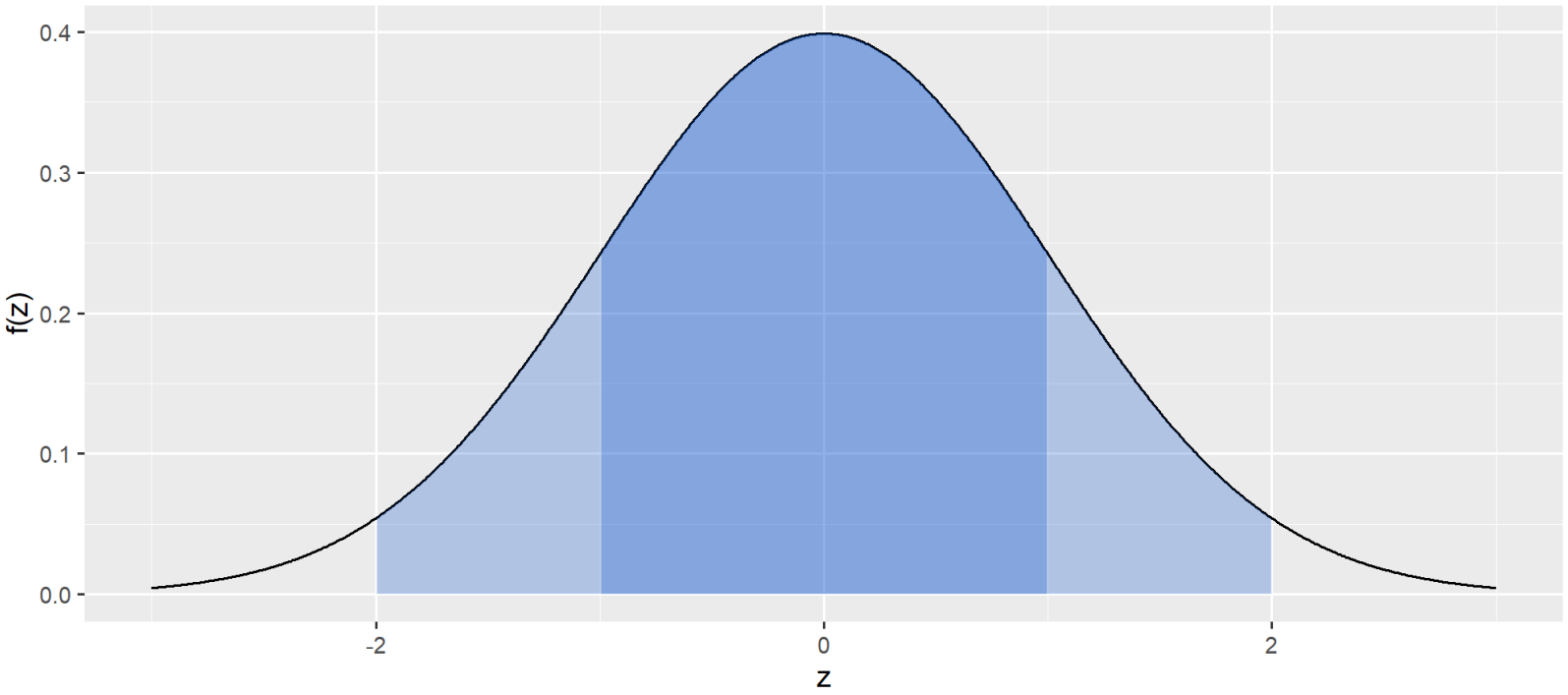
- Note: how did the population and sample distributions differ from one another?
- Note: how did the population/sample distributions differ from the sampling dists.?

Calculating the intervals… is much less important than understanding what they mean

Software will almost always do this for us.

When we focus on constructing the intervals themselves, we lose focus of the *much* more important *sampling distribution* context.

Remember our standard normal:

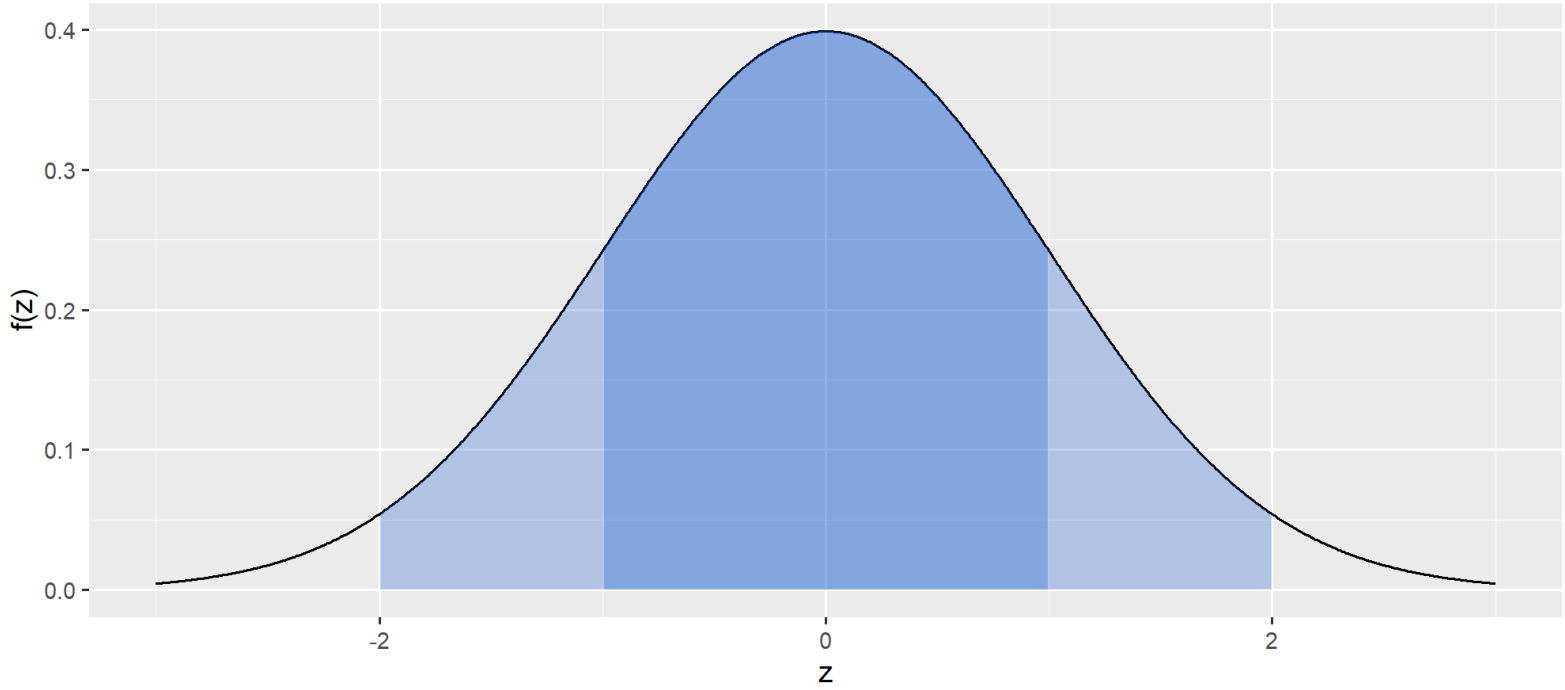


Calculating the intervals

Per the Central Limit Theorem, we can consider the sampling distribution to be normal

- If needed, review details in the previous lecture/slides on standard errors and sampling distributions.

We can use the nice properties of the standard normal to calculate CIs.



Key Concepts and Terms

Population distribution

Sample distribution

- Sample means
- Sample variance/

Sampling distribution

Standard error

[Hypothetical] repeated sampling

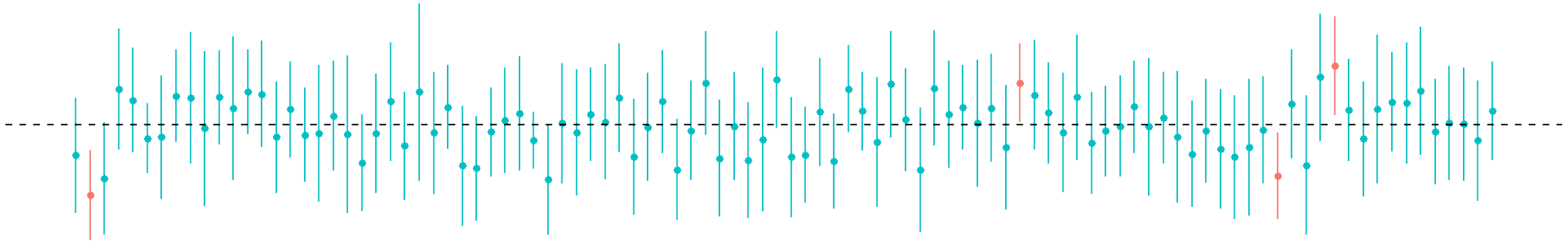
Confidence Intervals 2

Calculating the Interval

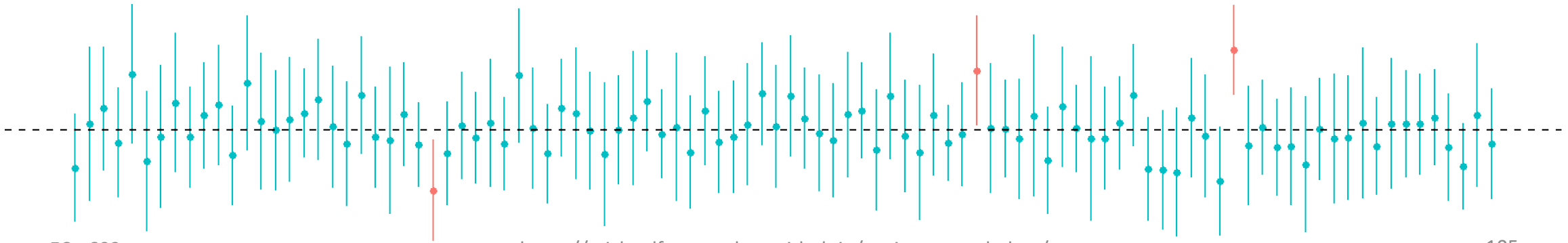
Outline

Confidence intervals aren’t that interesting…

Note on terminology: frequentist *confidence* and *significance*



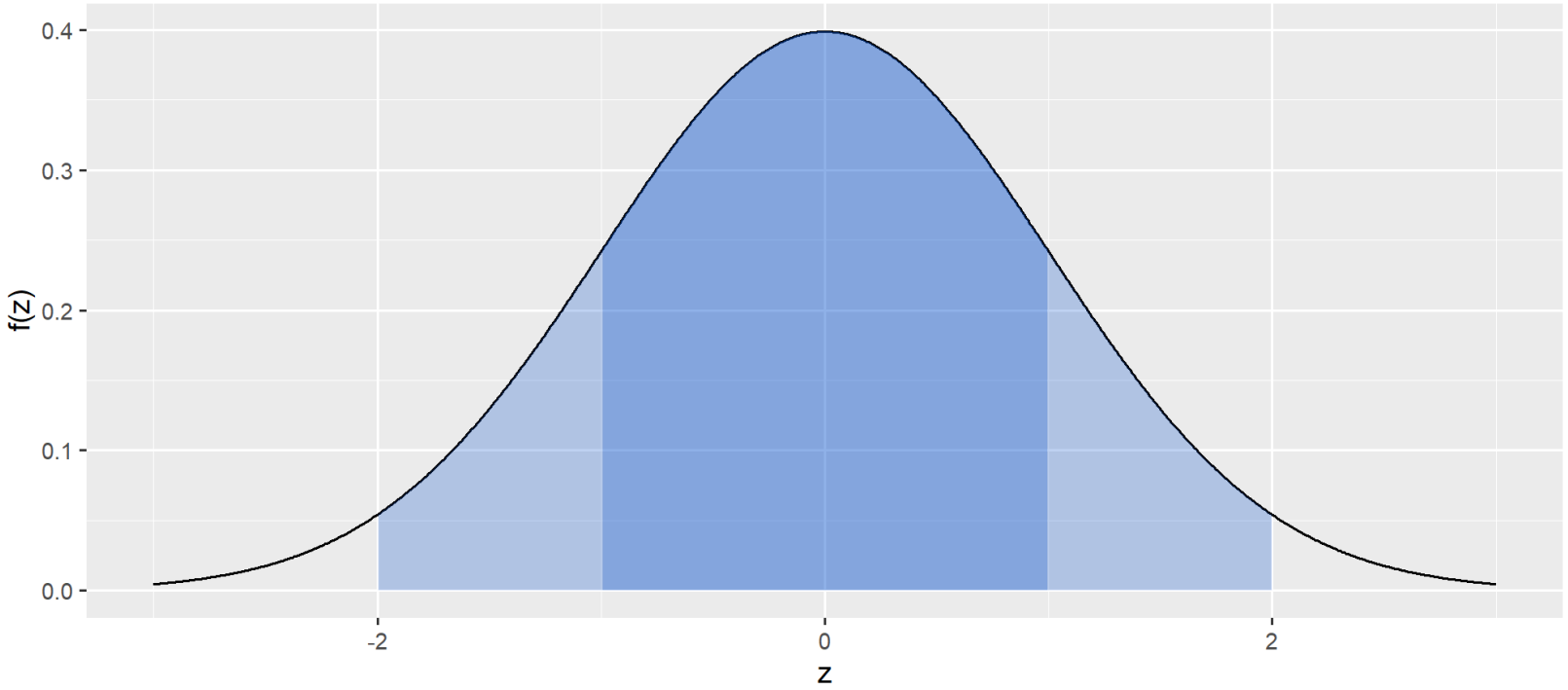
But it will help to build intuition if we work through the calculation on some real data.



Calculating the intervals… is much less important than understanding what they mean

The standard normal is our friend when we calculate CIs

- **Software will almost always do this for us.**
- When we focus on constructing the intervals themselves, we lose focus of the *much* more important *sampling distribution* context.
- **Remember the standard normal:**



The standard normal distribution

What does standardized mean?

Recall that the *standard normal* has $\mu = 0$ and $\sigma = 1$.

Z-standardization:

- We can convert any value from any normal distribution into the *standard normal* by:

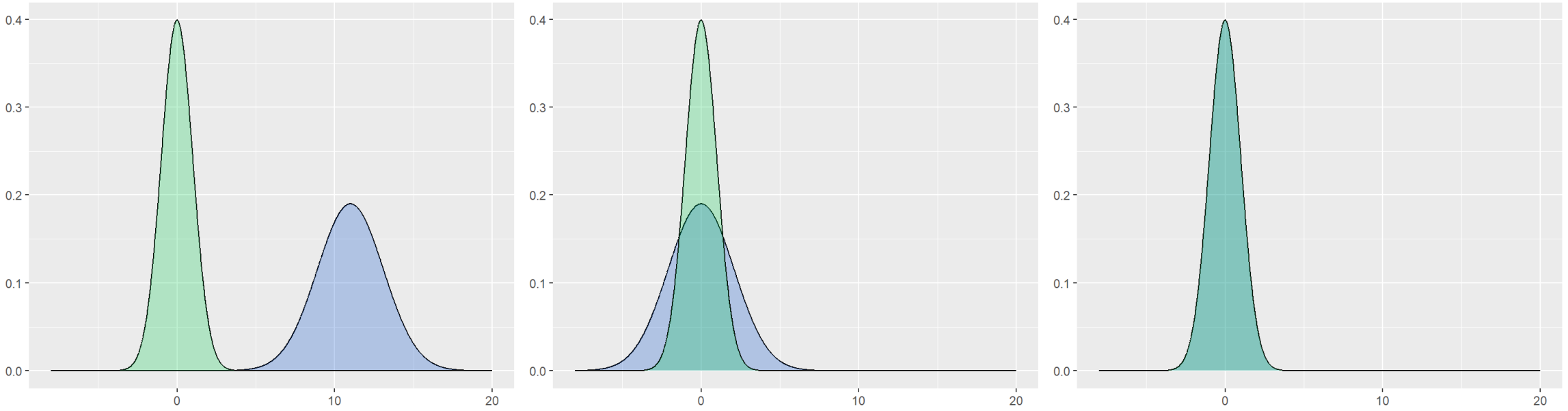
1. Subtracting the mean
2. Dividing by the standard deviation.

- We call a value standardized this way a *z-value*.
- It's very similar to how we calculate a *t-value*, more on that when we talk about t-tests and the t-distribution…

Z-standardization:

We can convert any value from any normal distribution into the *standard normal* by:

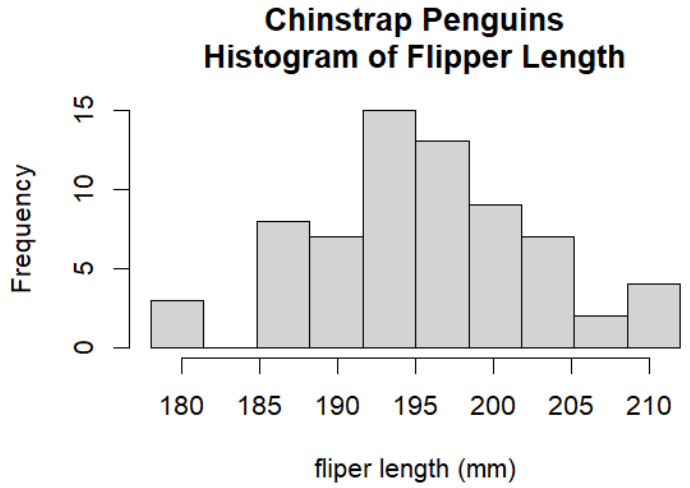
1. Subtracting the mean
2. Dividing by the standard deviation.



Standardizing example: Chinstrap penguins

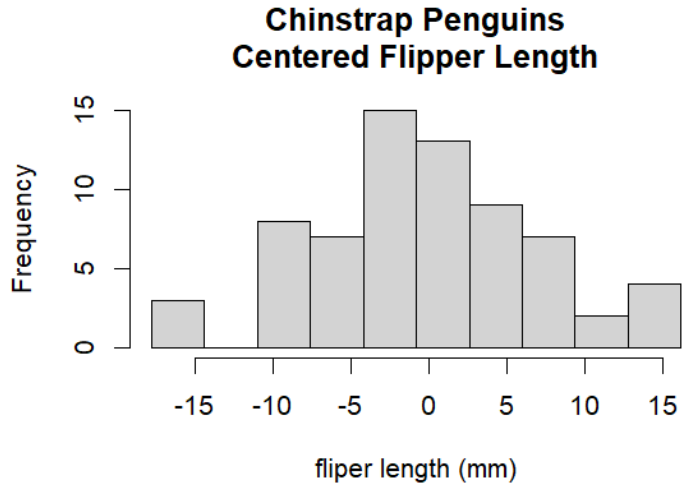
Z-standardizing the flipper lengths:

1. Calculate sample mean and standard deviation of original measurements.
2. Center the measurements: subtract the sample mean.
3. Standardize the measurements: divide the centered measurements by the sample standard deviation.



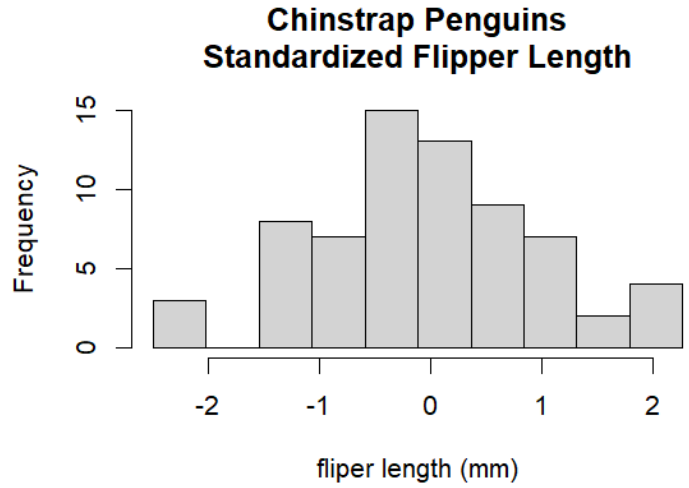
mean flipper length: 195.82

fipper length standard deviation: 7.13



mean flipper length: 0.00

fipper length standard deviation: 7.13



mean flipper length: 0.00

fipper length standard deviation: 1.00

Calculation steps

The general procedure is:

1. Calculate critical z-values for the standardized sampling distribution: use $alpha$ and the `qnorm()` R function.
    1. This is just the Standard Normal
2. Calculate sample mean and standard deviation
3. Calculate the sample standard error
4. Multiply the sample standard error by the critical z-value: This is the CI radius
5. CI is the sample mean $\pm$ the CI radius

We'll use the penguin flipper data to illustrate the procedure.

95% Critical z-values

We can use the quantile function to determine the critical z-values for a 95% confidence interval:

```
alpha = 0.05
z_crit = z_lower = qnorm(alpha/2, mean = 0, sd = 1)
z_upper = qnorm(1 - (alpha/2), mean = 0, sd = 1)

print(c(`Critical Z: lower tail` = z_lower,
        `Critical Z: upper tail` = z_upper), digits = 3)
```

```
Critical Z: lower tail Critical Z: upper tail
                 -1.96                  1.96
```

95% Critical z-values

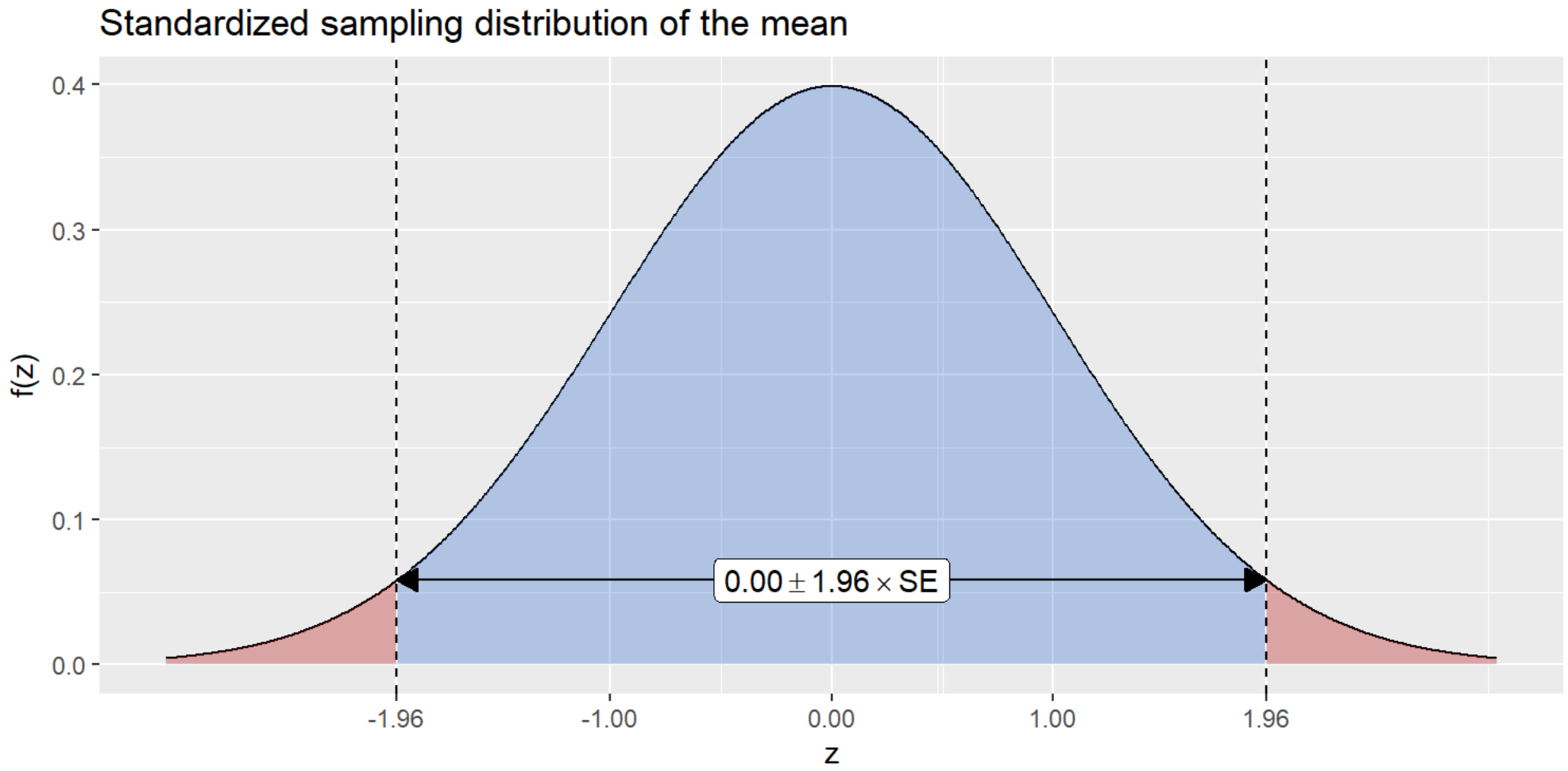
Verify that 95% of the normal density is within the upper and lower critical values using the cumulative probability function `pnorm()`

```
pnorm(z_upper)
pnorm(z_lower)
pnorm(z_upper) - pnorm(z_lower)
```

```
[1] 0.975
[1] 0.025
[1] 0.95
```

95% Critical z-values

Graphically:



Let’s take stock

Status

What we know so far:

- $alpha = 0.05$
- Critical Z values: ± 1.96

What we need:

- The sample mean
- The sample standard deviation
- The sample standard error

Calculation steps

The general procedure is:

1. Calculate sample mean and standard deviation
2. Calculate the sample standard error
3. Multiply the sample standard error by the critical z-value: This is the CI radius
4. CI is the sample mean $\pm$ the CI radius

Sample mean, standard deviation, standard error

Mean and sample SD are easy:

```
flipper_sample_mean = mean(dat$flipper_length_mm)
ssd = sd(dat$flipper_length_mm)
print(
  c(`Mean Flipper Length` =
      flipper_sample_mean,
    `Flipper Length Sample SD` =
      ssd),
  digits = 2)
```

```
      Mean Flipper Length Flipper Length Sample SD
                    195.8                       7.1
```

Sample mean, standard deviation, standard error

R does not have a built-in function for the standard error of the mean, but we know it's just $SSD / \sqrt{n}$

```
flipper_sample_se =
  ssd / sqrt(length(dat$flipper_length_mm))
print(
  c(`Sample Standard Error` = flipper_sample_se),
  digits = 2)
```

```
Sample Standard Error
                 0.86
```

- Astute students enrolled in the lab may notice that our calculate of SSE here does not take into account missing values!

Putting it together:

The last piece of the puzzle is to calculate the width of the interval.

We know:

- $SSE = 0.86$
- Critical Z values: ± 1.96

We can take advantage of the fact that the sampling distribution is normal, per the *central limit theorem*.

Calculation steps

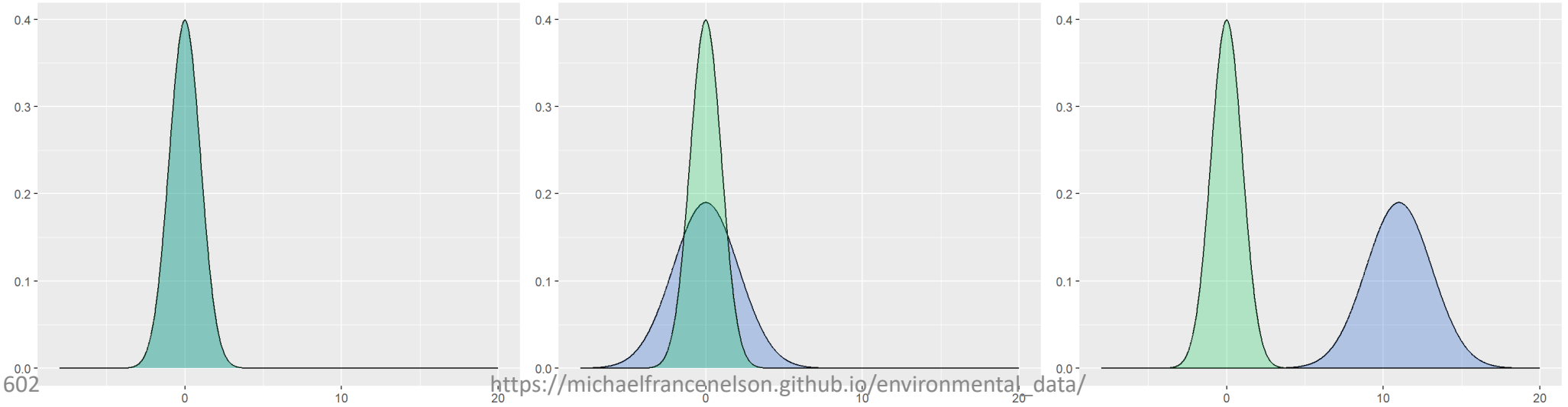
The general procedure is:

1. Multiply the sample standard error by the critical z-value: This is the CI radius
2. CI is the sample mean $\pm$ the CI radius

Putting it together:

We can use some of the nice features of the normal distribution to finish the calculation:

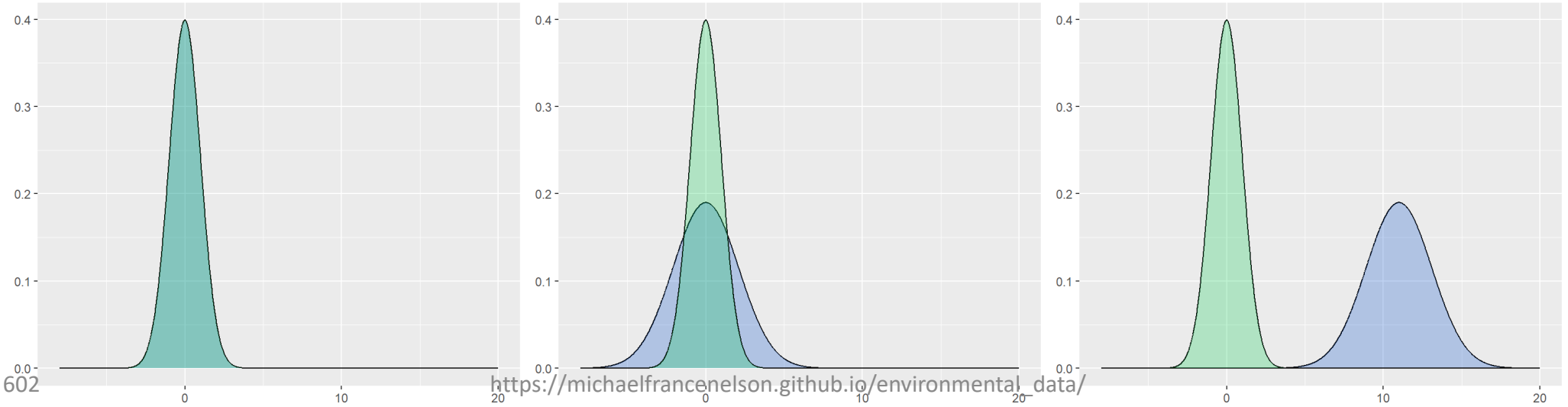
- It is symmetrical: we can use the same critical z-value for the upper and lower ends of the CI



Putting it together:

We can use some of the nice features of the normal distribution to finish the calculation:

- We can back-transform from a standard normal to a normal with our sample parameters:
  1. Multiplying by the sample standard deviation.
  2. Add the sample mean.



Confidence interval formula:

$$CI = \bar{x} \pm SSE \times Z_{crit}$$

Back-transform from the standardized sampling distribution to the sample sampling distribution:

```
ci_radius = abs(flipper_sample_se * z_crit)
```

- Add the sample mean

```
flipper_sample_mean - ci_radius
flipper_sample_mean + ci_radius
```

```
[1] 194.1284
```

```
[1] 197.5186
```

What if we had a larger sample?

What was the sample size for the penguin flipper data set?

```
length(dat$flipper_length_mm)
[1] 68
```

We know that sampling distributions are narrower if we have larger samples.

Let's pretend we had a sample size of 150

What if we had a larger sample?

Recall our CI:

```
[1] 194.13 197.52
```

```
n = 150
sse_150 = ssd / sqrt(150)
ci_radius_150 = abs(sse_150 * z_crit)
print(c(ci_radius, ci_radius_150), digits = 5)
print(c(flipper_sample_mean - ci_radius,
        flipper_sample_mean + ci_radius), digits = 5)
print(c(flipper_sample_mean - ci_radius_150,
        flipper_sample_mean + ci_radius_150), digits
= 5)
```

```
[1] 1.6951 1.1413
[1] 194.13 197.52
[1] 194.68 196.96
```

Small Sample Sizes: T-Distribution and CIs

- What if your sample size is small?
  - When n < 30, you can create more conservative Cis using the t distribution.
  - Central Limit Theorem states that for bigger sample sizes, the sampling distribution is approximately normal.
- You can also use the t-distribution, with appropriate degrees of freedom, to calculate a CI. The procedure is similar because the t distribution is similar to the Standard Normal.
- The calculation is very similar, except you calculate critical values from the t distribution (instead of the Standard Normal).
  - The r function is qt()
  - You have to remember the degrees of freedom!

Calculation steps

The general procedure is:

1. Calculate critical **t-value.**
    1. Use $alpha$ and the `qt()` R function.
    2. This is just a t distribution with the appropriate degrees of freedom: n-1
2. Calculate sample mean and standard deviation
3. Calculate the sample standard error
4. Multiply the sample standard error by the critical **t-value**: This is the CI radius
5. CI is the sample mean $\pm$ the CI radius

You should try out the procedure on the penguin flipper data.

Questions?

- Let’s go over your Deck 5 questions now!
  - Hypothesis Tests
  - Sampling Distributions
  - Confidence Intervals
  - Others

In-Class Confidence Intervals