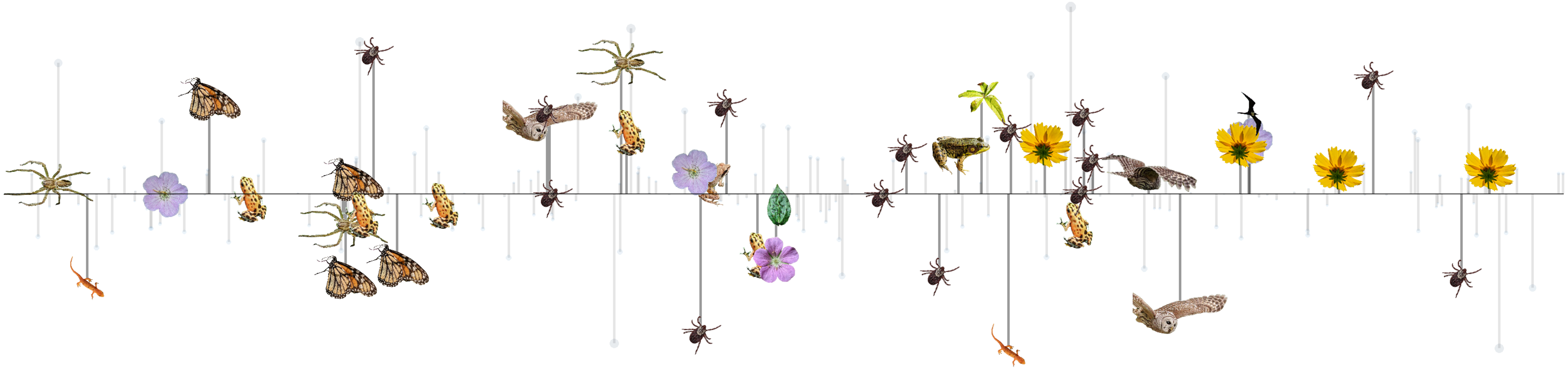
Intro to Quantitative Ecology
UMass Amherst – Michael France Nelson

Deck 11 – Frequentist Statistics



Announcements

A few reminders

- **Default groups:** double check your group assignment submissions. If your group has submitted as part of the default group, i.e. you didn't use the group self-selection tool, we cannot grade your work and you will get a zero.
- **Late work:** A quick reminder of the late work policy:
  - Assignments are accepted up to 1 week late with a 25% penalty.
  - Assignments are not accepted more than 1 week late.
  - If you need more time on an assignment, make sure you contact us **before** the due date.
- **Office hours:** if you're struggling with an assignment or general course concepts, make sure you attend office hours or schedule a session with Ana or myself. Remember that it's worth 5% of your grade!

Let’s Review Model Selection

How do we choose a good model?

- What part of a model summary tells us how much variation is explained?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 19.98859    2.93234   6.817 3.82e-07 ***
PercVeg      0.30111    0.03628   8.299 1.20e-08 ***
Dist2Road    4.86408    3.68759   1.319    0.199    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.957 on 25 degrees of freedom
Multiple R-squared:  0.7338,    Adjusted R-squared:  0.7125 
F-statistic: 34.45 on 2 and 25 DF,  p-value: 6.544e-08
```

How do we choose a good model?

- What part of a model summary tells us how much variation is explained?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.98859    2.93234   6.817 3.82e-07 ***
PercVeg      0.30111    0.03628   8.299 1.20e-08 ***
Dist2Road    4.86408    3.68759   1.319    0.199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.957 on 25 degrees of freedom
Multiple R-squared:  0.7338,    Adjusted R-squared:  0.7125
F-statistic: 34.45 on 2 and 25 DF,  p-value: 6.544e-08
```

How do we choose a good model?

The adjusted R-squared value tells us approximately how much of the variation in the response is explained by the predictors.

- Adding more predictors ***always*** increases the R-squared
- Does adding more predictors necessarily make a better model?

The Akaike Information Criterion (AIC) is an objective value that captures the tradeoff between model complexity and understanding.

- Low values are better
- Rewards models that explain the data
- Penalizes complicated models

Complicated Models

- Is the model with the highest R-squared value always the best one?
- Why?

?

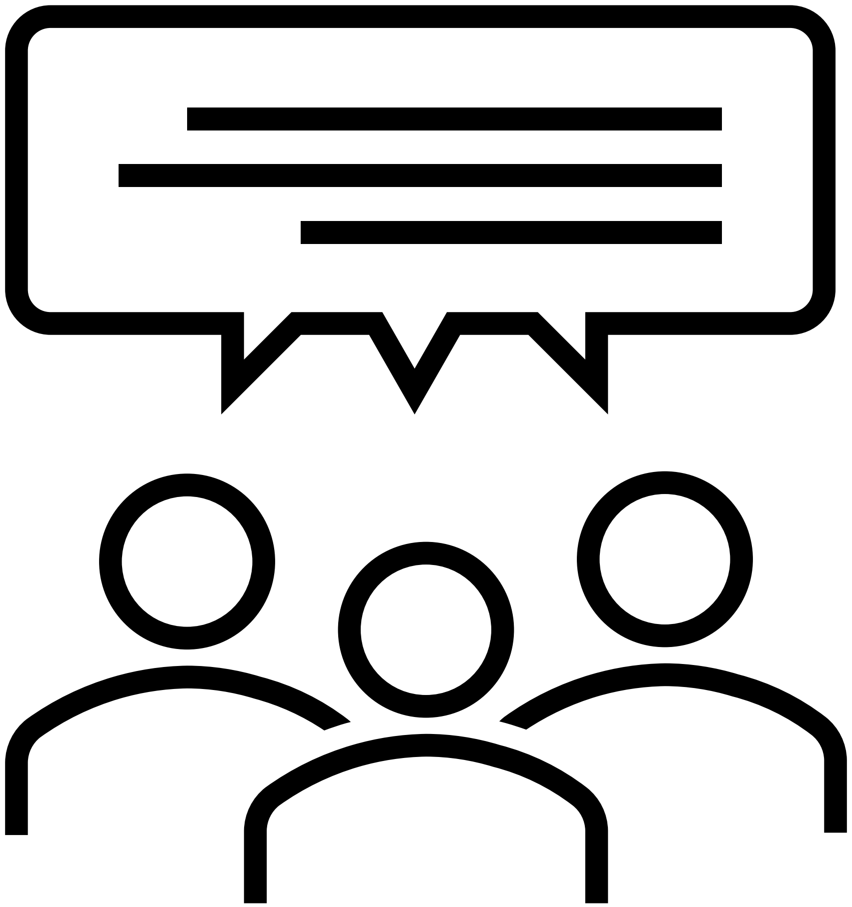
ANOVA Questions Inspired By The Salamander Assignment

Review Topics

Degrees of Freedom

Test statistics: the F-ratio

Differences between models



Degrees Of Freedom

Formulas And Graphical Intuition

Degrees Of Freedom

How do we define the degrees of freedom?

The salamander dataset contains:

- 273 salamander observations.
- 3 levels for `Sex`: female, male, unknown
- 4 levels for `Site`: A, B, C, D

Degrees of Freedom: Concept

Degrees of freedom is an *adjusted* version of the sample size.

We have to make an adjustment because we are working with *finite-sized* samples. More on this when we talk about Frequentism.

The idea is we make an assumption that our observations are independent. That is a key assumption that allows us to do inference.

Let's say we have a collection of five numbers with a mean of 5. The first four numbers are 3, 4, 6, and 7. By the definition of mean, we know that the sum of all the terms has to add up to 20. The last number has to be 5, it cannot vary.

Degrees of freedom are a measure of how many *independent* pieces of information we have.

Degrees of Freedom Symbols:

We use the following abbreviations for the number of observations and groups:

- `n` is the total number of observations
- `g` is the number of groups within a grouping factor

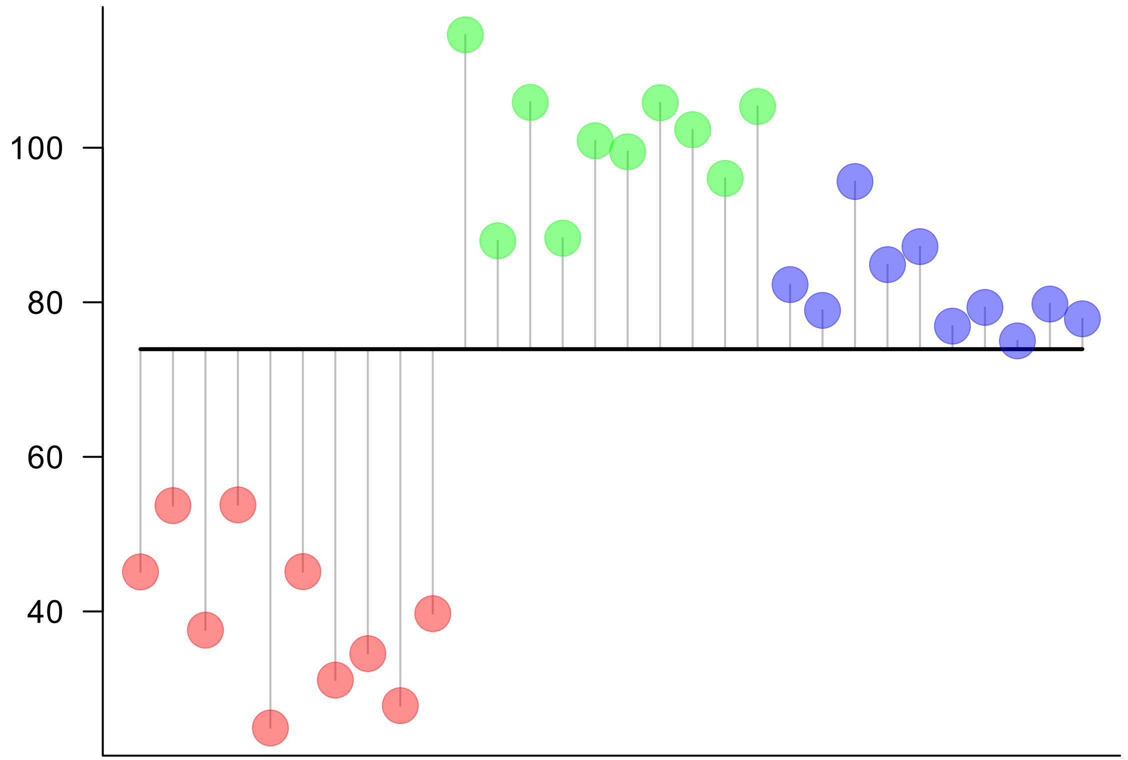
Usually you lose one degree of freedom for each *summary statistic* you estimate.

Note that you have 2 grouping factors in a 2-way ANOVA

Degrees Of Freedom: Total

You lose one degree of freedom because you are have to calculate the mean of the squared differences for *all observations*.

- This mean value is the *grand mean*.
- $df_T = n - 1$



Degrees Of Freedom: Between-Group

Between-group is also called *Among Group*.

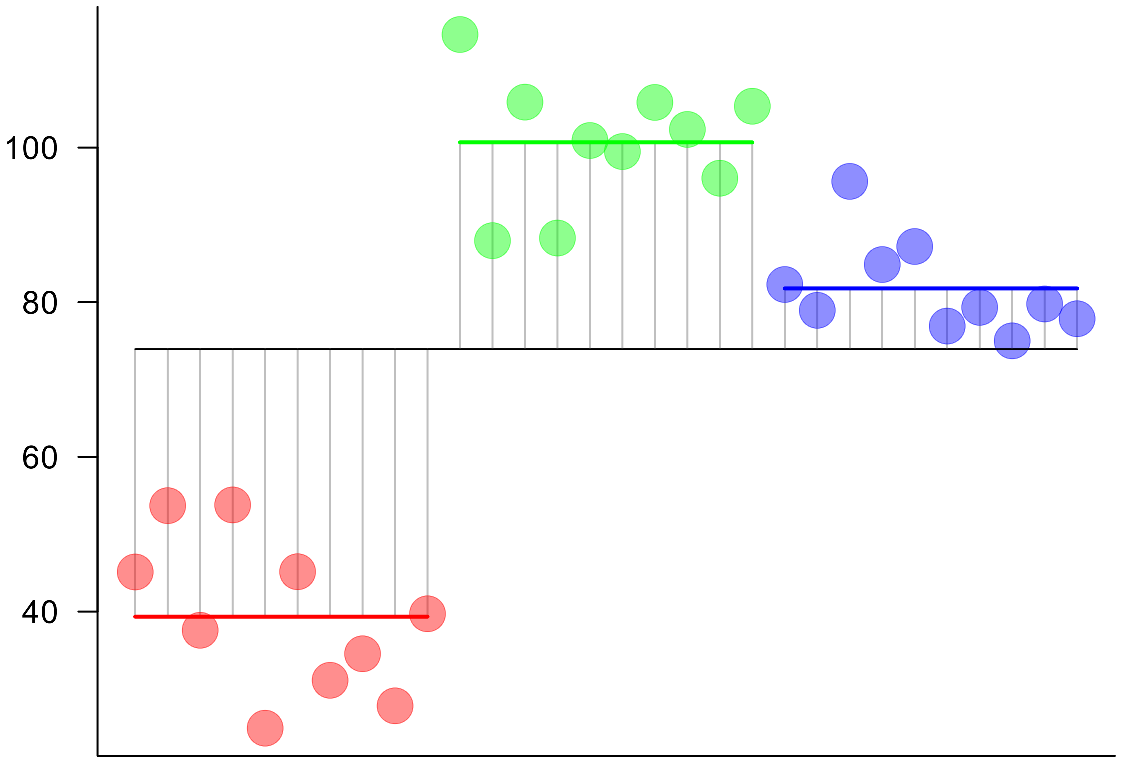
You need to calculate mean values for observations within each group. However, you only lose g - 1 degrees of freedom.

- Since you are calculating a mean value within each group, you should lose 1 degree of freedom for each group, right?
- Remember that you already calculated the *grand mean*. If you calculate the means within all the groups, except for the last one, you can derive the value of the last group mean using the other group means and the grand mean.
  - Grand mean was already accounted for with the n-1
  - You only need to calculate g-1 means: the other mean is determined by the g – 1 that you already calculated.

Degrees Of Freedom: Between-Group

We don't have to use the values of individual observations in the last group, so they are all free to vary.

- $df_B = g - 1$



Degrees Of Freedom: Within-Group (Residuals)

Also called the *Residual Degrees of Freedom*.

We lost 1 df from the grand mean calculation, and $g - 1$ df for the among-group calculations.

We know the total degrees of freedom must equal the sum of the between-group and within-group degrees of freedom;

- We can rearrange to find the within-group (residual) degrees of freedom:

$$df_T = df_W + df_B$$

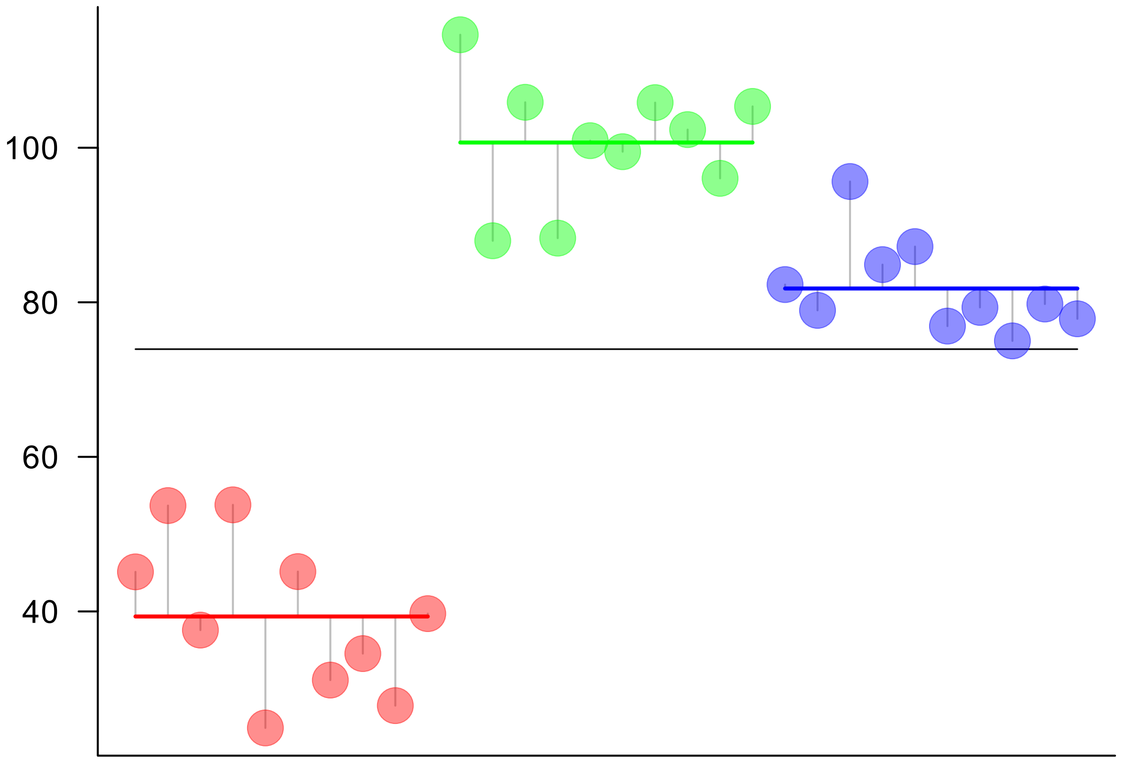
$$df_W = df_T - df_B = (n - 1) - (g - 1) = n - g$$

Degrees Of Freedom: Within-Group (Residuals)

It's like the mean for each group, but without using the grand mean. That's why you lose one degree of freedom for each group:

$$df_W = n - g$$

- Within each group you need $n_i - 1$
- All of those -1 terms add up to g!



Test Statistics

The F-Ratio

ANOVA Main Effect: F-statistics

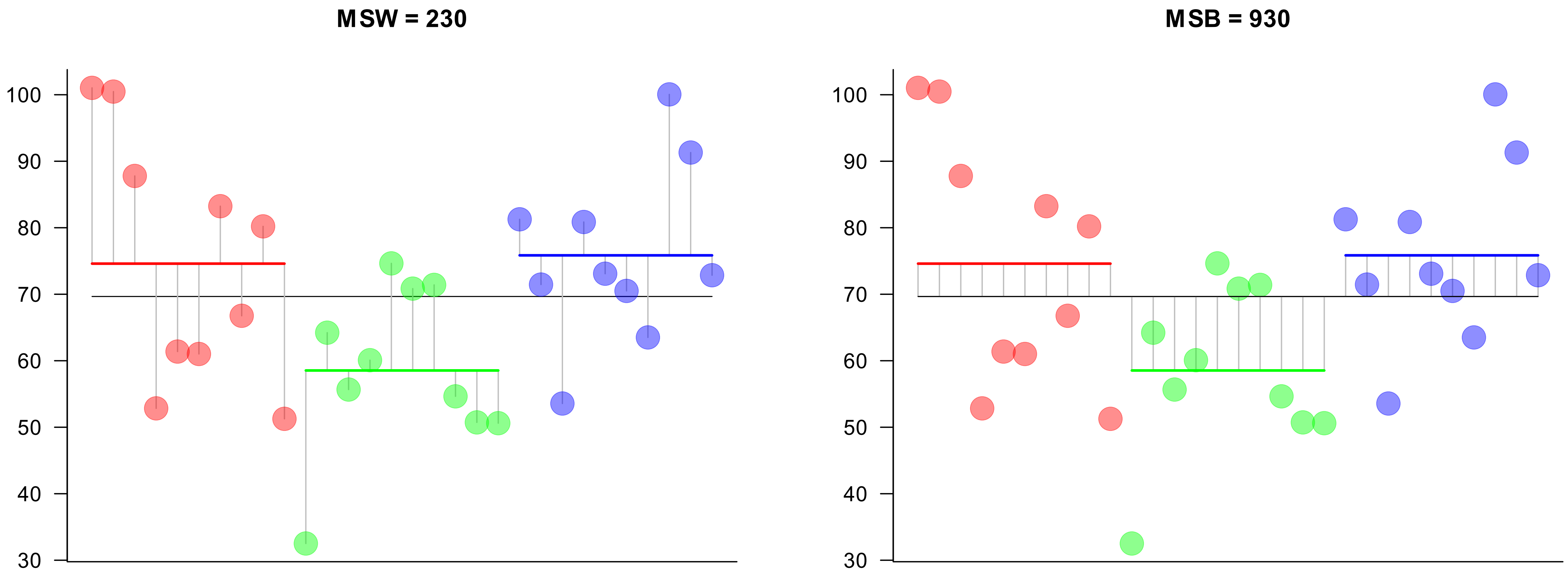
We use the F-statistic to calculate the *significance* of a grouping factor in an ANOVA.

F is the ratio of the mean between-group squared deviation MS_B to the within-group (residual) squared deviation MS_W.

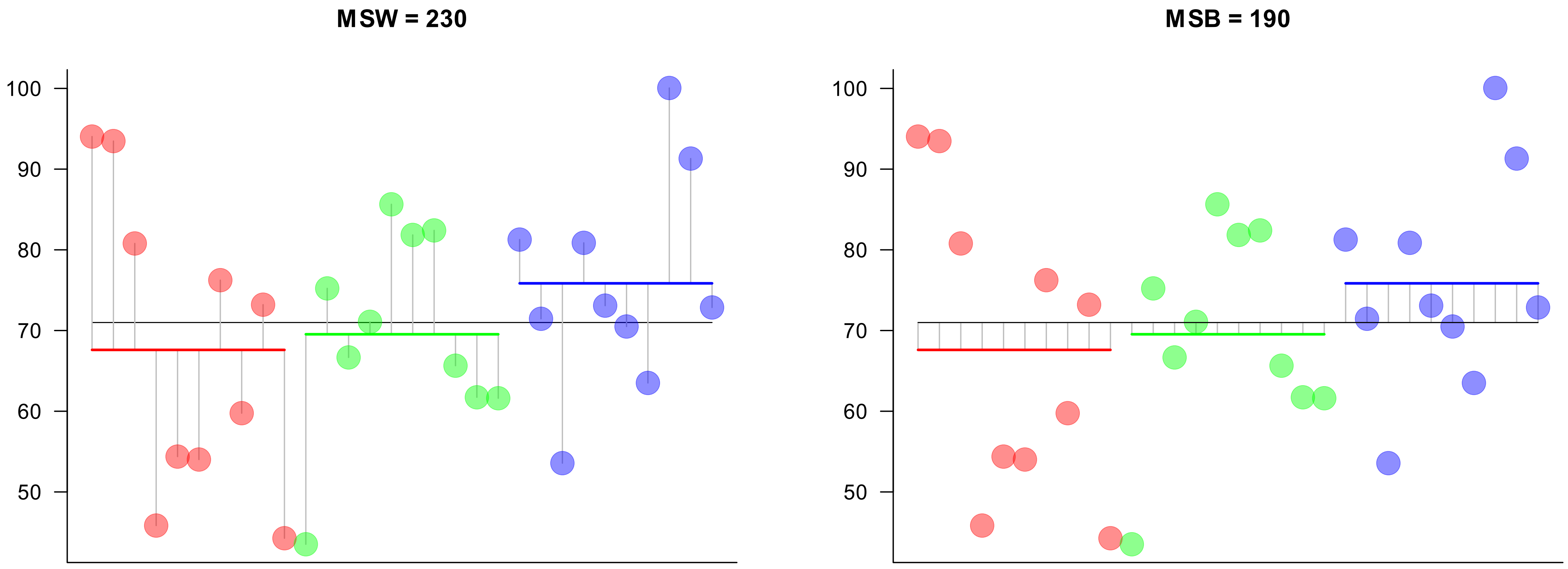
If MS_B is high (compared to MS_W), it means there was a large difference between group means: adding the grouping factor improved the model.

- The variation in group means was greater than the variation of individual observations within groups.

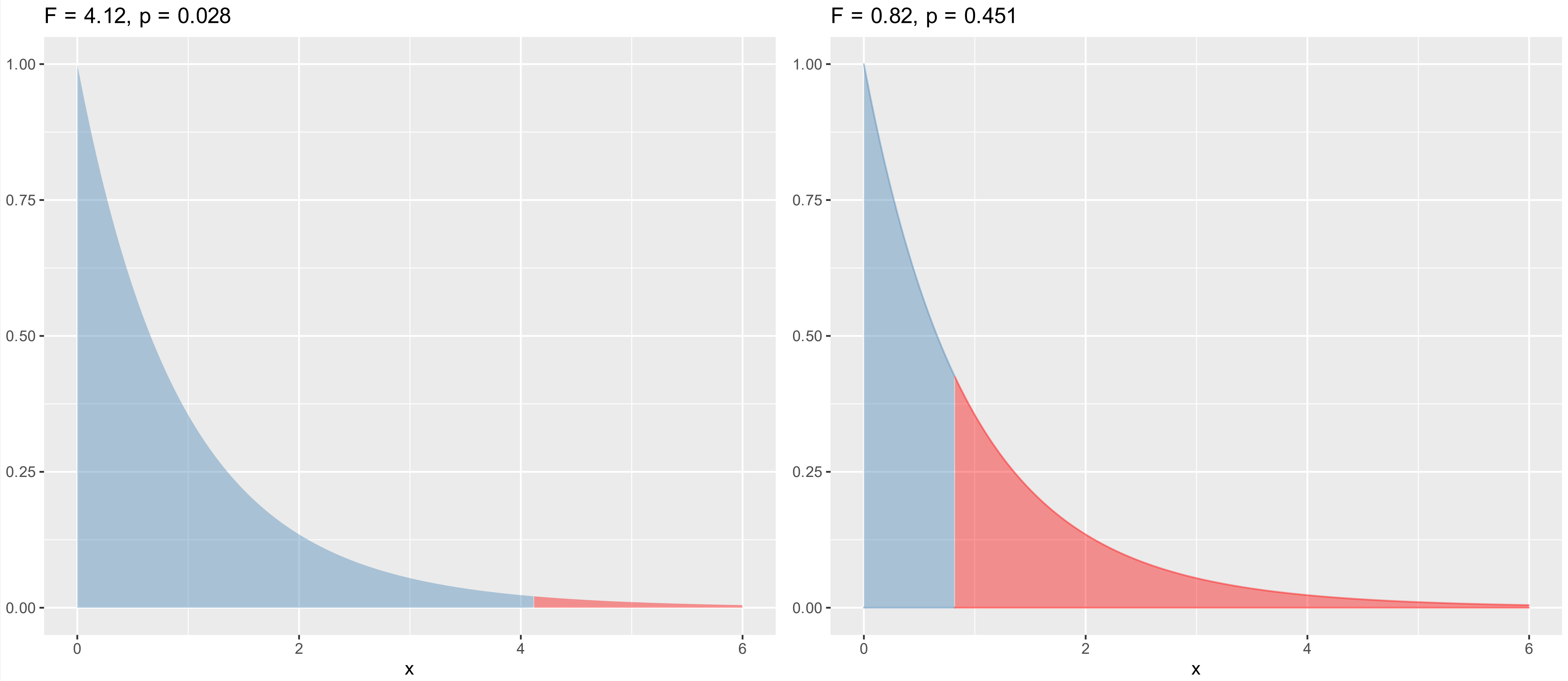
ANOVA Main Effect: Large F-ratio (MSB/MSW)



ANOVA Main Effect: Small F-ratio (MSB/MSW)



ANOVA: Significant + Non-Significant F-ratios



Differences Between Models

F and p

Why Do F And p Change?

You created two one-way ANOVAs and one two-way ANOVA. You probably noticed the that the F and p-values for the factors changed slightly between the one-way and two-way ANOVA.

Does this matter?

- Usually no. If the p-values in both the one- and two-way ANOVAS are below the threshold (0.05), the interpretation doesn’t change.

The calculations for different types of models vary, so even with the same data we don’t expect the value of a test statistic to stay exactly the same. For example, both the sum of squares and the degrees of freedom of the residuals change. This alone changes the F- and p-values.

One-Way ANOVA for Site

```
Analysis of Variance Table

Response: Total_length
           Df Sum Sq Mean Sq F value   Pr(>F)
Site        3   2877  959.12  4.9504 0.002317 **
Residuals 269  52118  193.75
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-Way ANOVA for Site and Sex

The F- and p-values changed, but they are still significant.

```
m2 = lm(Total_length ~ Site + Sex,data = sals)
anova(m2)
Analysis of Variance Table


Response: Total_length
           Df Sum Sq Mean Sq F value    Pr(>F)
Site        3   2877   959.1  7.9174 4.452e-05 ***
Sex         2  19774  9886.8 81.6144 < 2.2e-16 ***
Residuals 267  32345   121.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thursday: Frequentism In-Class Assignment

To be ready for the Thursday group assignment, you'll need to check out the Seeing Theory site. It contains excellent visualizations of many statistical concepts that can help build your intuition.

Before Thursday, individually explore the Seeing Theory site. Specifically check out the following sections:

- Probability Theory Chapter: Central Limit Theorem
- Frequentist Inference Chapter: Point Estimation
- Frequentist Inference Chapter: Confidence Interval

Key Concepts and Terms

- What is the Frequentist paradigm?
- Assumptions
- Frequentist pros and cons
- Confidence and significance

Nota Bene

Frequentism is not your enemy!

Much of the following slides will 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 fashionable to disparage Frequentist statistics

- Some criticism is justified

Frequentism

An inference *paradigm*

It's a framework that contains a set of tools and assumptions that we can use to do inference.

Frequentist statistics is the dominant paradigm

It's the kind of inference you [usually] learn in a first course on statistics.

- It's not the *only paradigm*!

Frequentism

Some key features

- Makes no assumptions about prior knowledge of a system.
- 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.

Frequentism: Commonly Used Tools

Most of the tools you have probably used in previous courses or research have been frequentist versions of:

- Linear regression
- ANOVA
- T-tests
- Generalized Linear Models

Frequentism Assumptions

Some of the conceptual weirdness in Frequentism comes from the assumptions.

Key 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 general, we can't observe every *sampling unit* in a population.
- The population exists and is has true properties.
- We have to 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.

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

Benefits 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

Drawbacks of the Frequentist framework

- Repeated sampling assumption: 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).
- Does not explicitly take into account the results of prior experiments.
  - But it can incorporate this information in the design of our experiments and the form of our hypotheses.

Announcements

- Emails and default groups – thank you!
- RMarkdown 2 assignment is not ready, we'll delay for one week. Details to follow.
- No week 12 pre-class questions
- No assignments due Monday
  - I hope we all can relax a bit over the long weekend!
- Final Projects: assignments and details will be released next week

Confidence and Significance

More Frequentist Weirdness

Point and Interval Estimates

Point estimates, which we calculate from samples, provide an estimate of population parameters.

Because we're sampling, our estimates have an associated uncertainty.

Interval estimates help us quantify our uncertainty

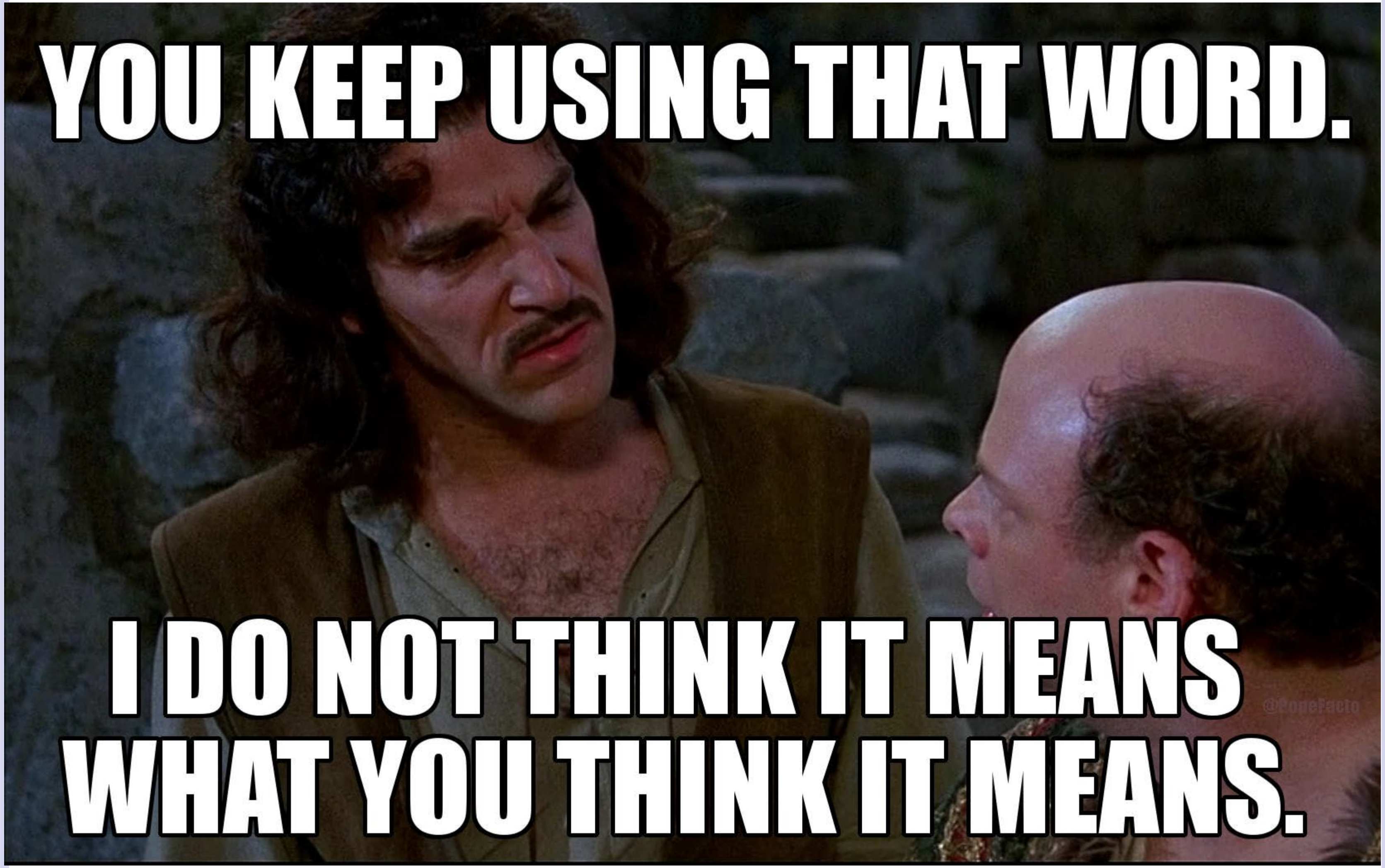
- Wide intervals correspond to high uncertainty
- Narrow intervals correspond to lower uncertainty

Confidence intervals of the mean provide a measure of uncertainty of our estimates of the population mean value.

They are related to sampling distributions (we'll talk about these soon).

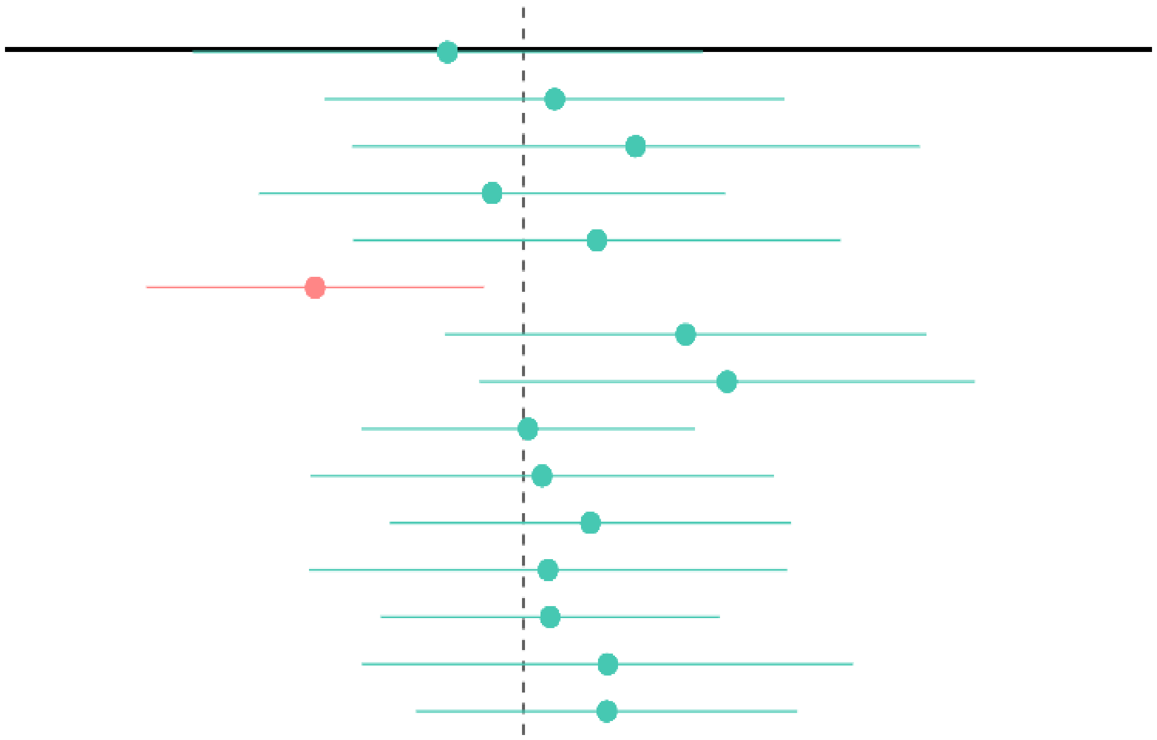
Frequentist Confidence and Significance: Collision with Everyday Usage

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



Interpreting a 95% confidence interval:

- “If I were to repeat the experiment many times, approximately 95% of the intervals contain the true value.”



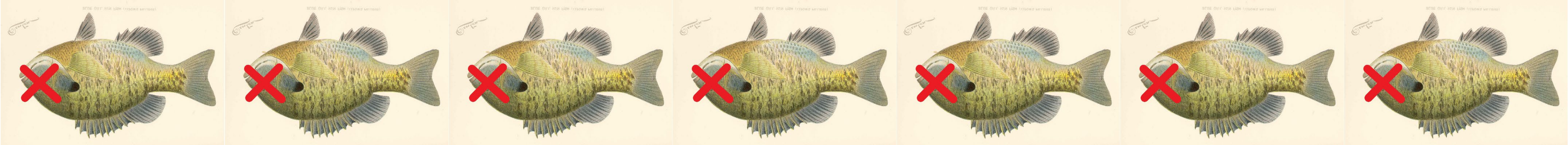
Frequentist Confidence and Significance

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

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

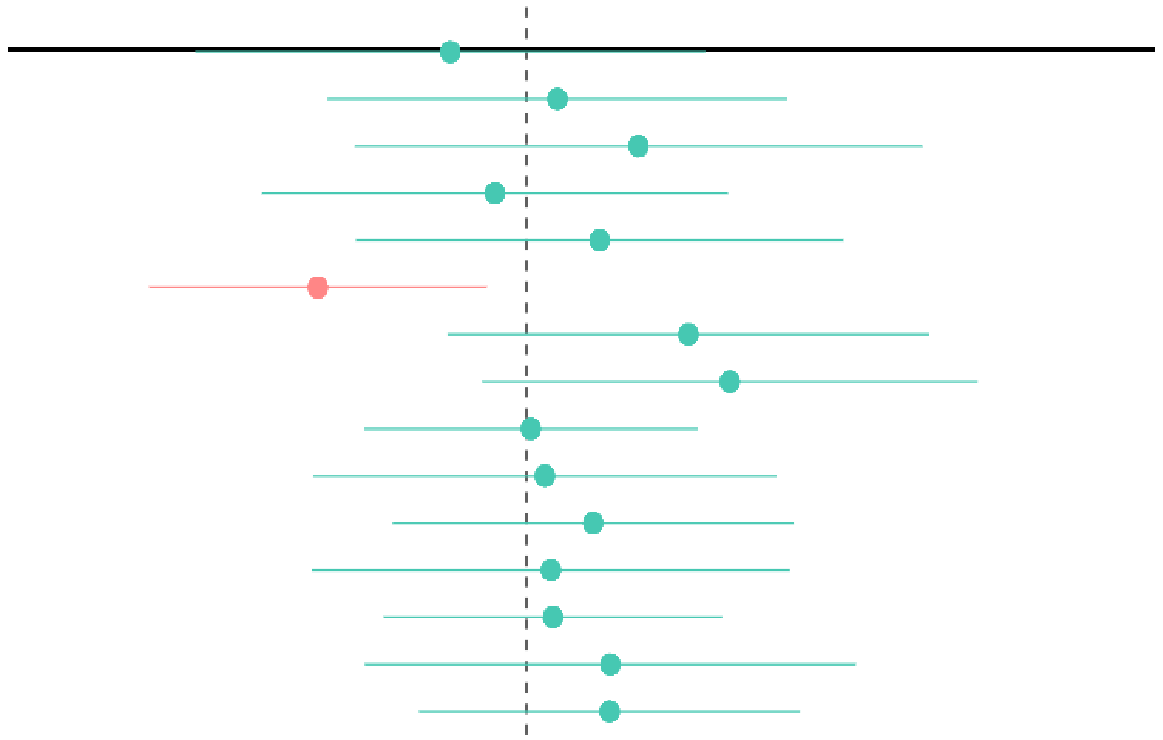
- This isn't exactly what we want from our analysis.
- It's difficult to explain the "if I repeated…" idea to a nonscientific audience.

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?
- Assumptions
- Frequentist pros and cons
- Confidence and significance



Sampling Distributions

What is a Sampling Distribution?

- It's the distribution of a *sample statistic*.
- If you took many samples and calculated a sample statistic, such as the mean, many times, the sample statistic would follow the sampling distribution.
- The Central Limit Theorem tells us that the sampling distribution of a statistic approaches a Normal distribution, no matter how the original data are distributed! [under certain conditions]

Sampling Distributions

Sampling distributions are **very** important, but often misunderstood.

Sampling distributions are **not** the same as the distribution of a population variable.

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 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 standard error **gets smaller** as the sample size increases!
- The sample standard deviation **stabilizes** as the sample size increases!

Intuition: simulation experiments

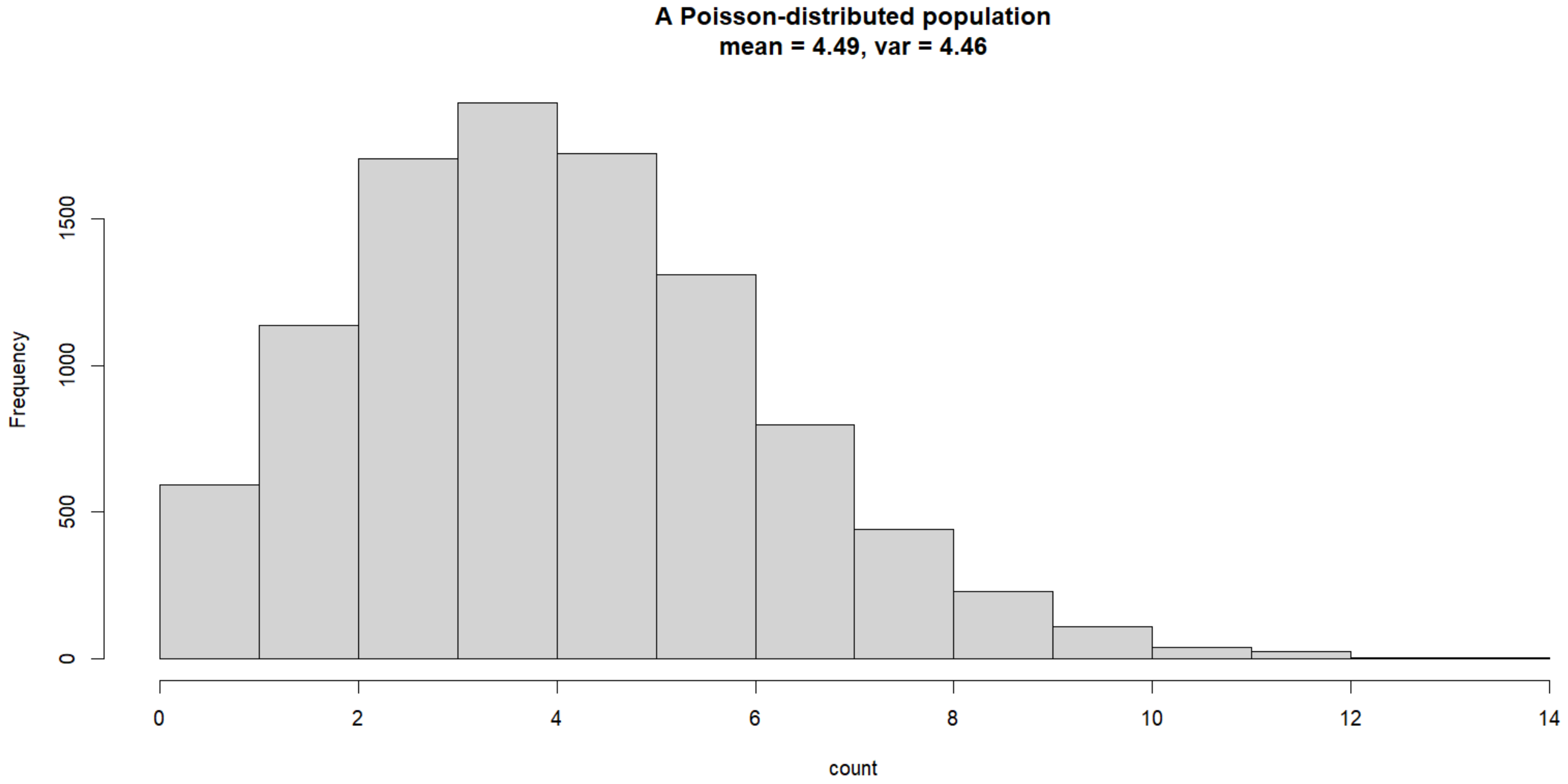
This is technical material. Let’s try to build some intuition with graphical examples.

A non-normal distribution: a Poisson-distributed popualtion

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

Poisson population

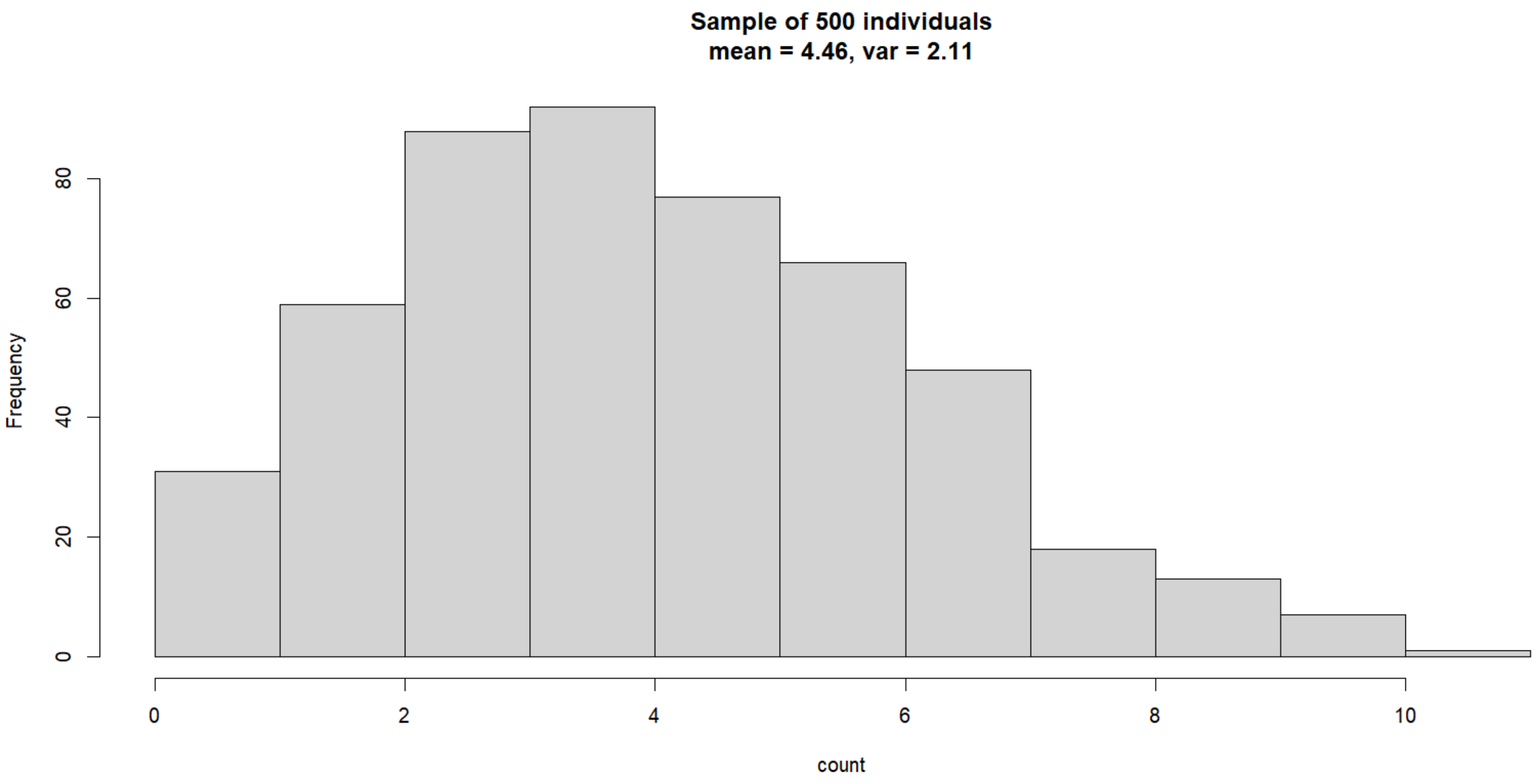
- 10000 individuals, lambda = 4.5



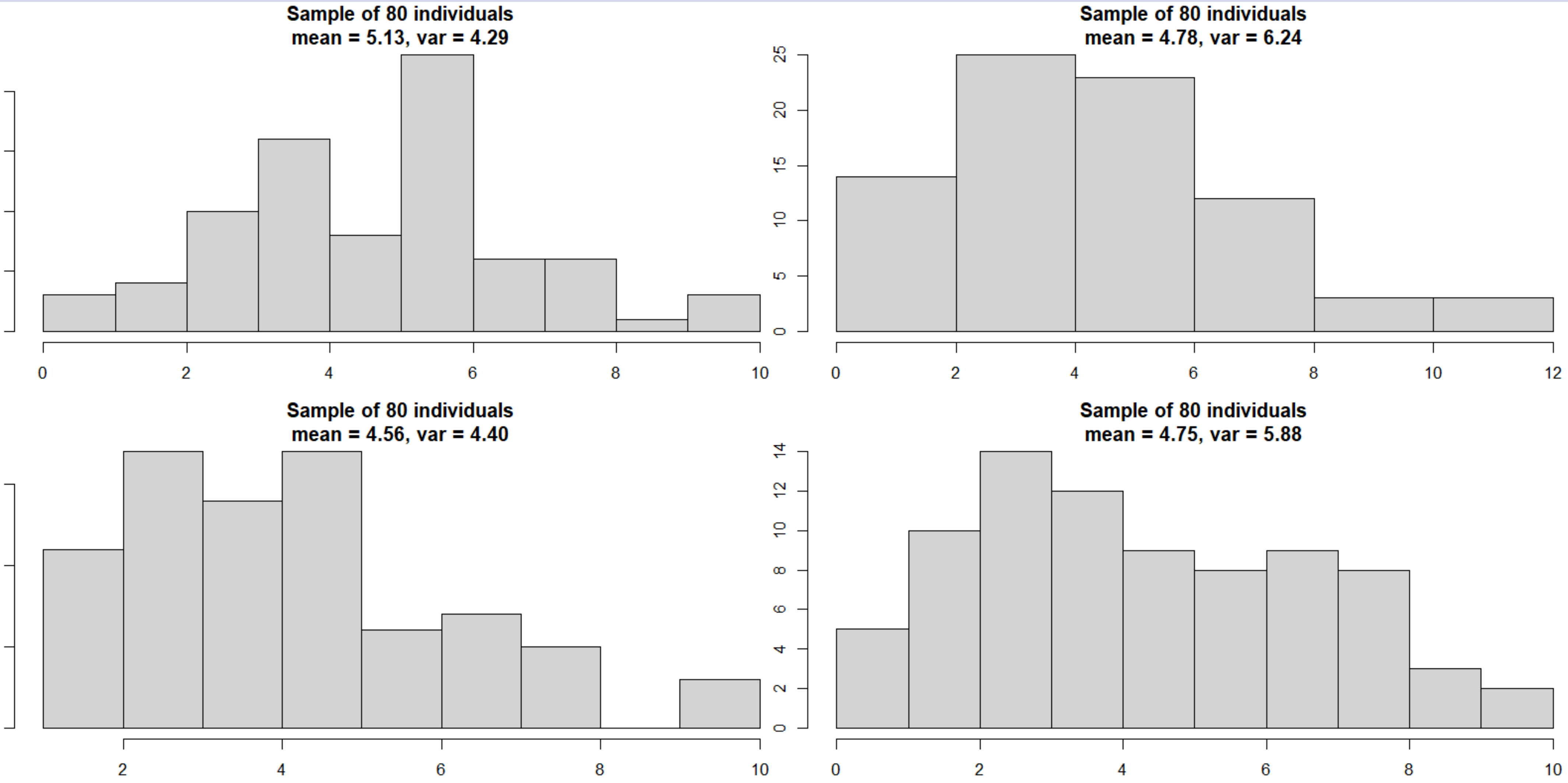
Sample from the population

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

Let's take a sample of 500:

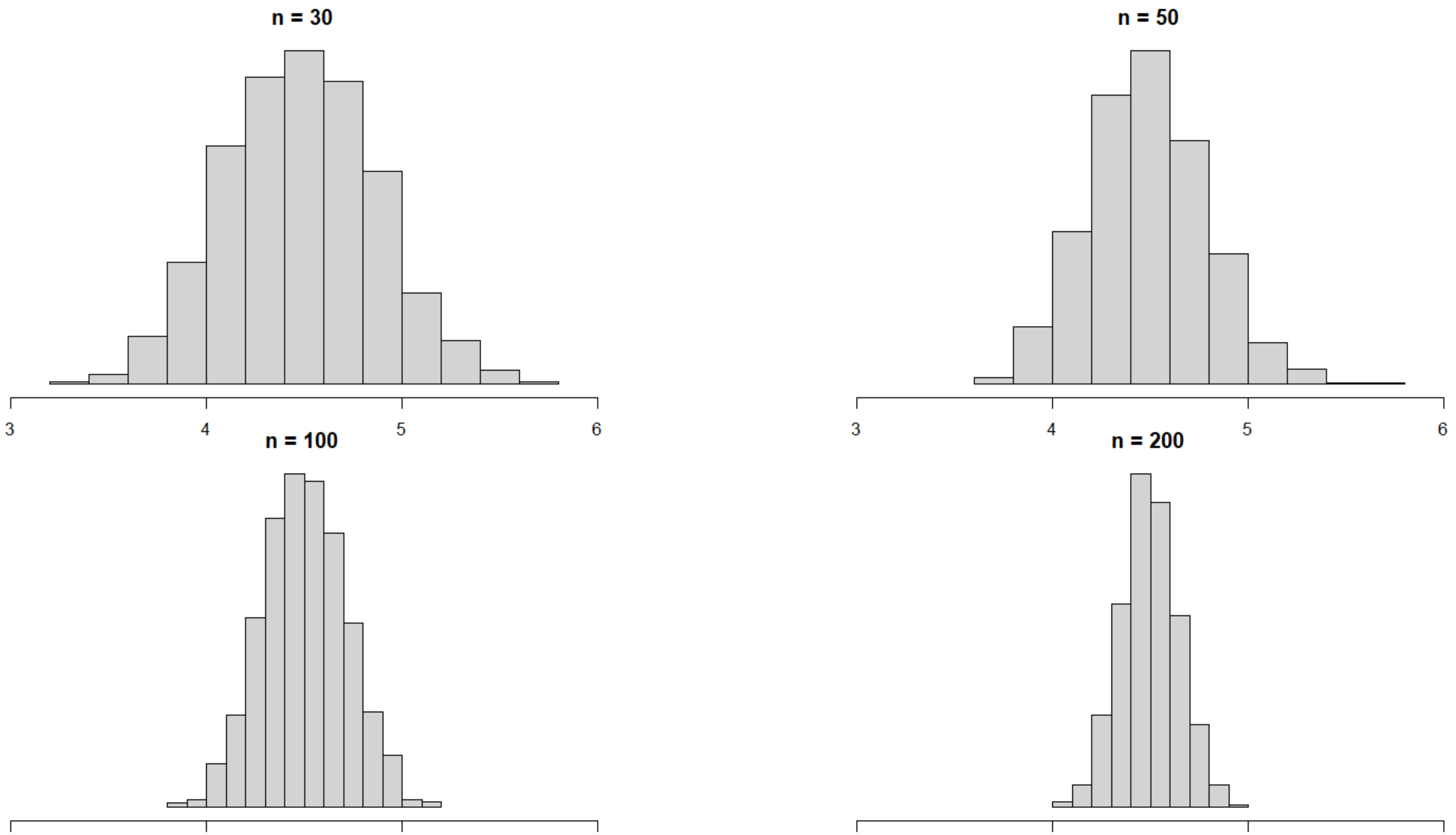


Sample from the population: 80 samples



Sampling Distribution of the mean

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



Key Points

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!