SUMMER 2025

TEXAS
The University of Texas at Austin

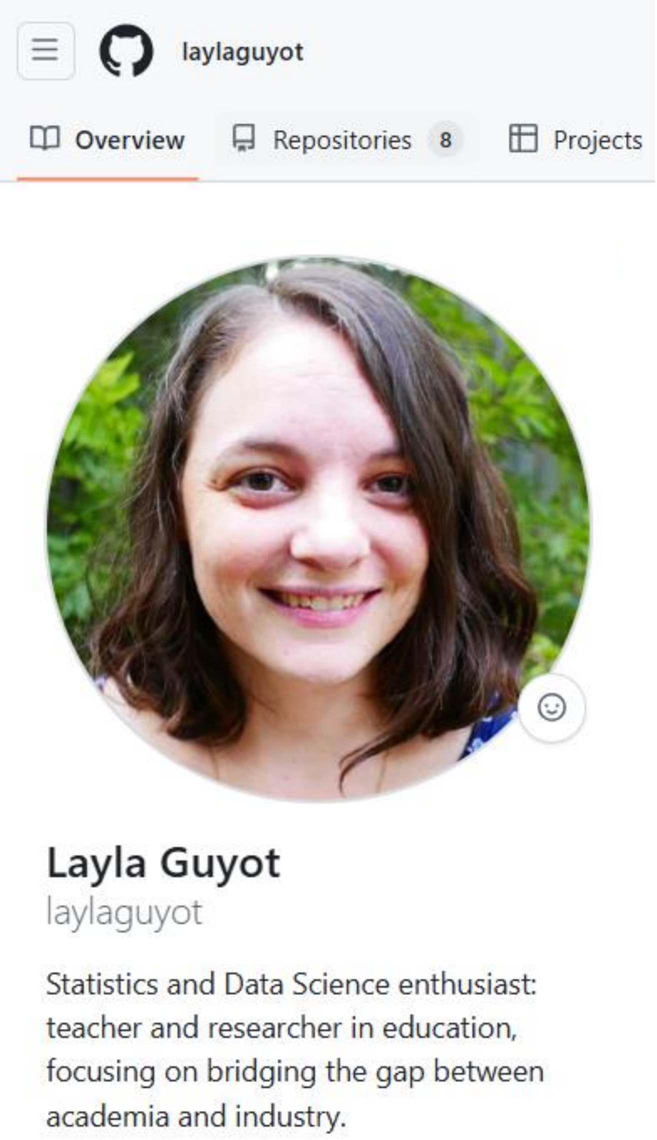
INTRODUCTION TO STATISTICAL MODELING

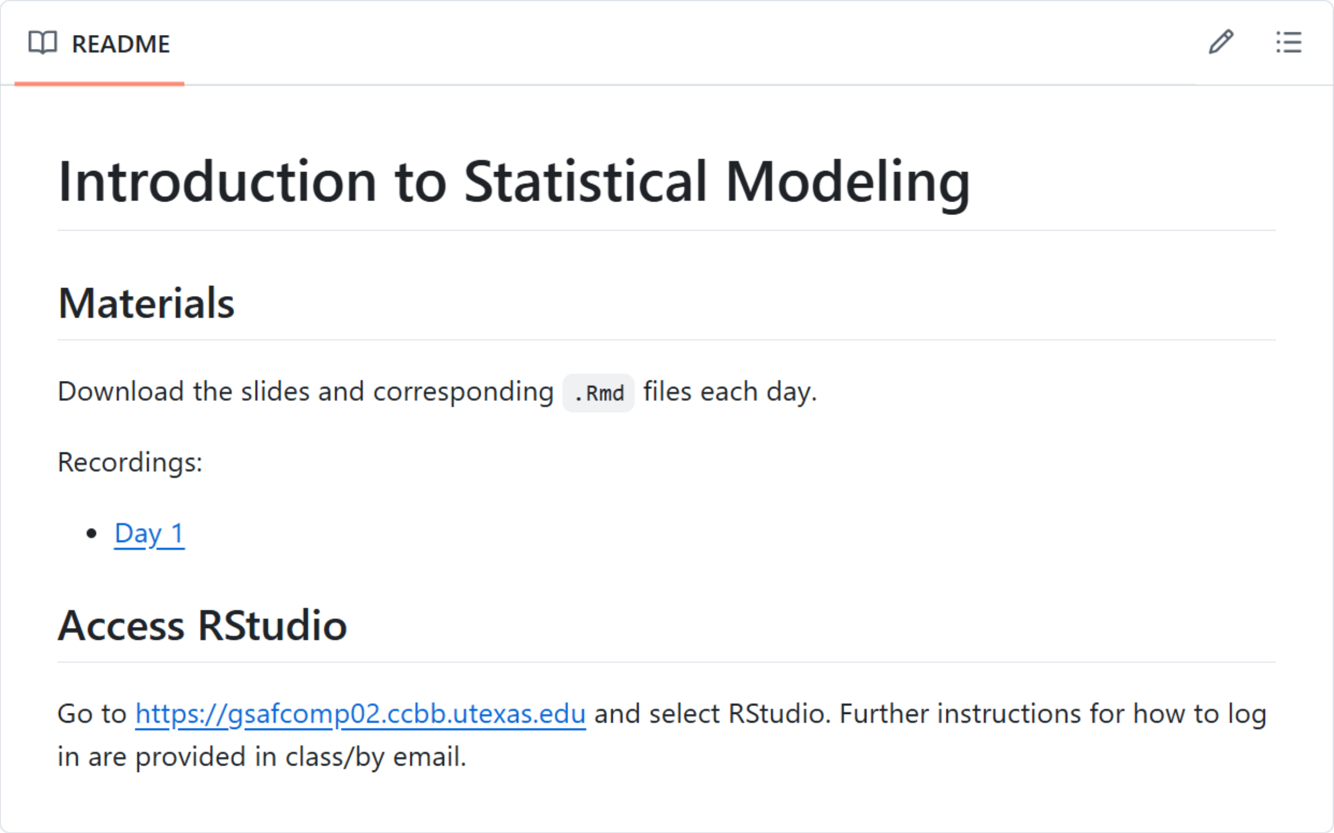
Center for Biomedical Research Support

LAYLA GUYOT
Assistant Professor of Instruction, Ph.D.
Department of Statistics and Data Sciences
The University of Texas at Austin

Access materials

https://github.com/laylaguyot/CBRS_Intro_Statistical_Modeling




README

Introduction to Statistical Modeling

Materials

Download the slides and corresponding `.Rmd` files each day.

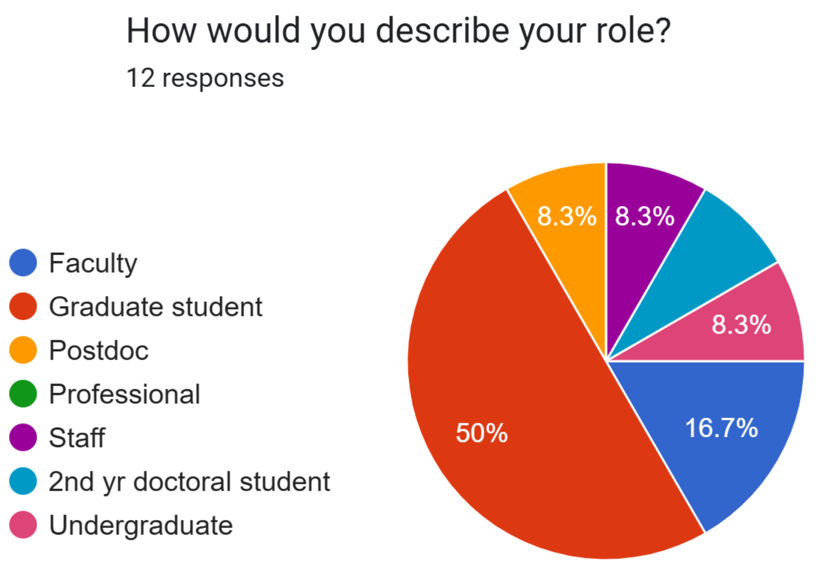
Recordings:

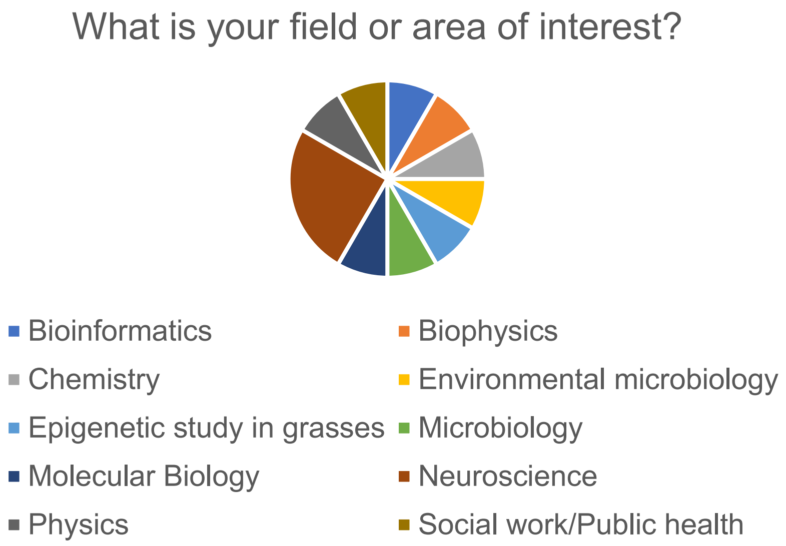
- Day 1

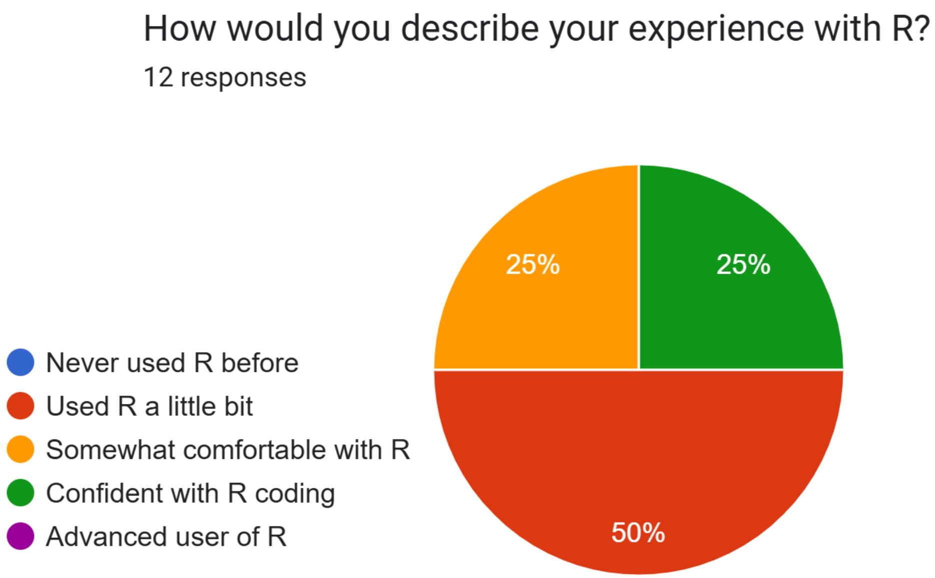
Access RStudio

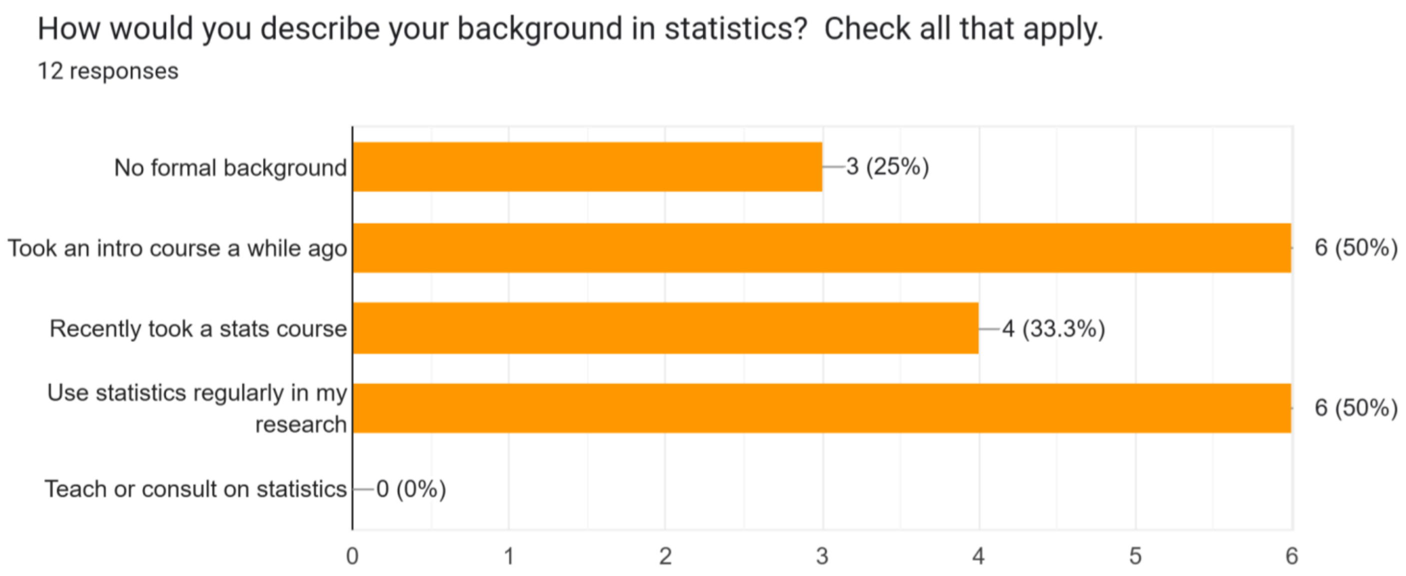
Go to https://gsafcomp02.ccbb.utexas.edu and select RStudio. Further instructions for how to log in are provided in class/by email.

Who is participating to this workshop?









Tentative Schedule

Day 1
Exploring Data

- Study design and variables
- Descriptive statistics and visualizations
- Introduction to hypothesis testing

Day 2
Making Inferences

- Probability, random variables, and common probability distributions
- Sampling distributions and Central Limit Theorem
- Confidence intervals, t-tests, ANOVA, and Chi-square tests

Day 3
Linear Regression

- Simple Linear Regression
- Multiple Regression with different types of predictors
- Model assumptions, evaluation, and comparisons

Day 4
Logistic Regression

- Odds
- Logistic Regression
- Model evaluation with ROC curves or confusion matrix

Day 5
Model Building

- Underfitting, overfitting, and cross-validation
- Regularization with Lasso and Ridge
- Missing data

Probability

What is a probability?

To define probabilities, we consider the possible outcomes of an experiment.

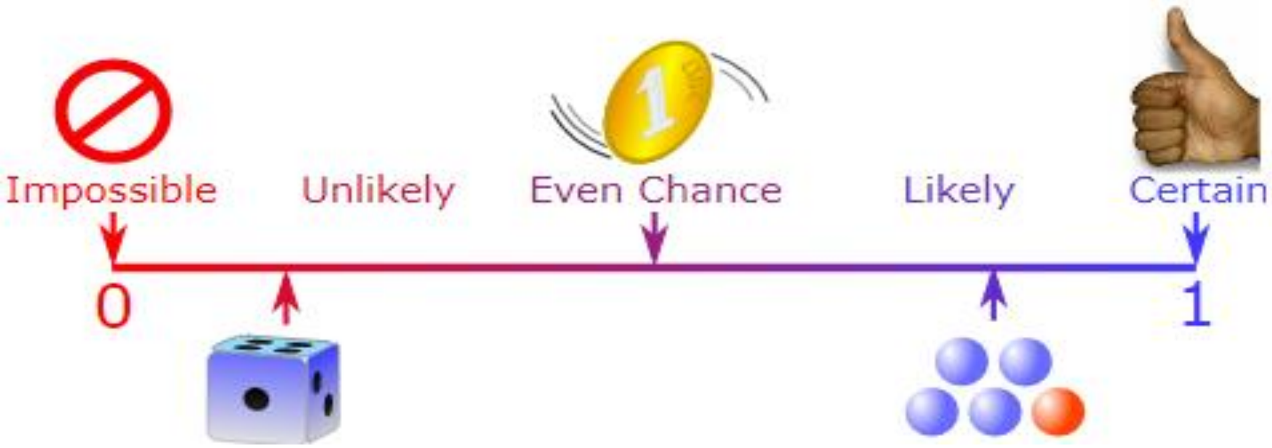
<u>Example</u>: Toss a coin and record the outcome.

- ❑ **What are the possible events?**
- ❑ **What is the probability of getting heads?**

Probability

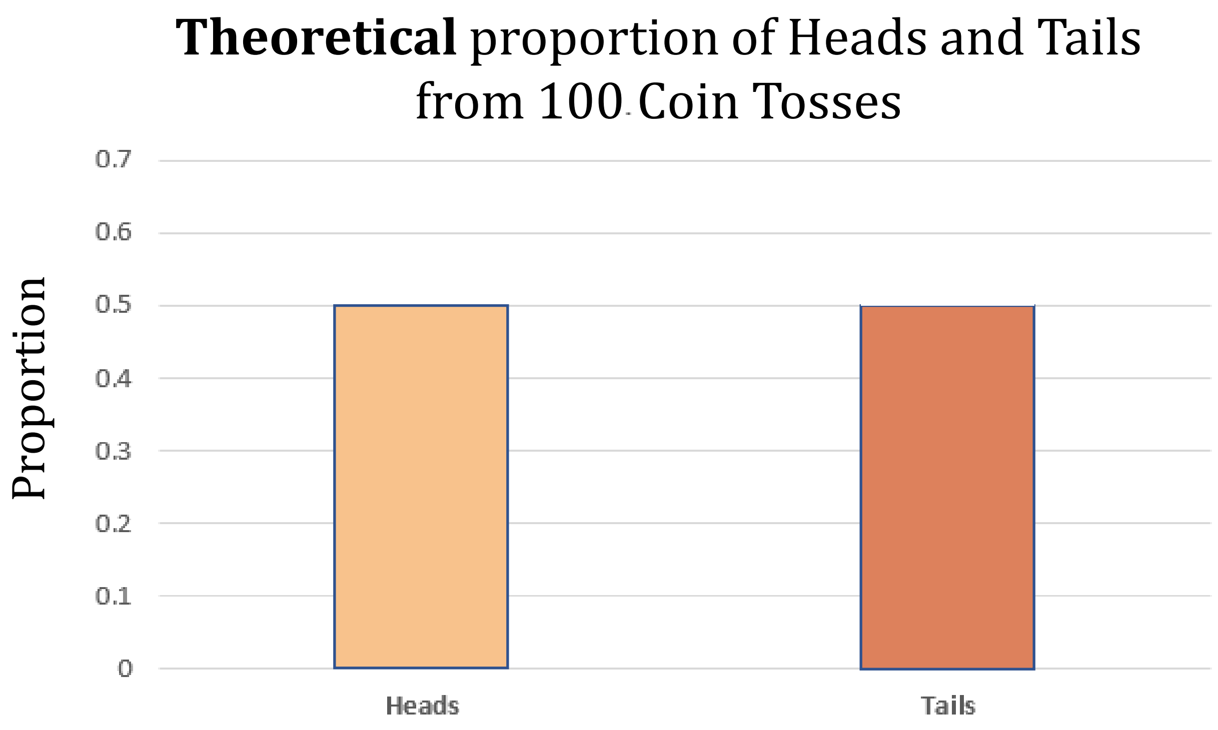
What is a probability?

A probability has a value between 0 and 1.



Probability

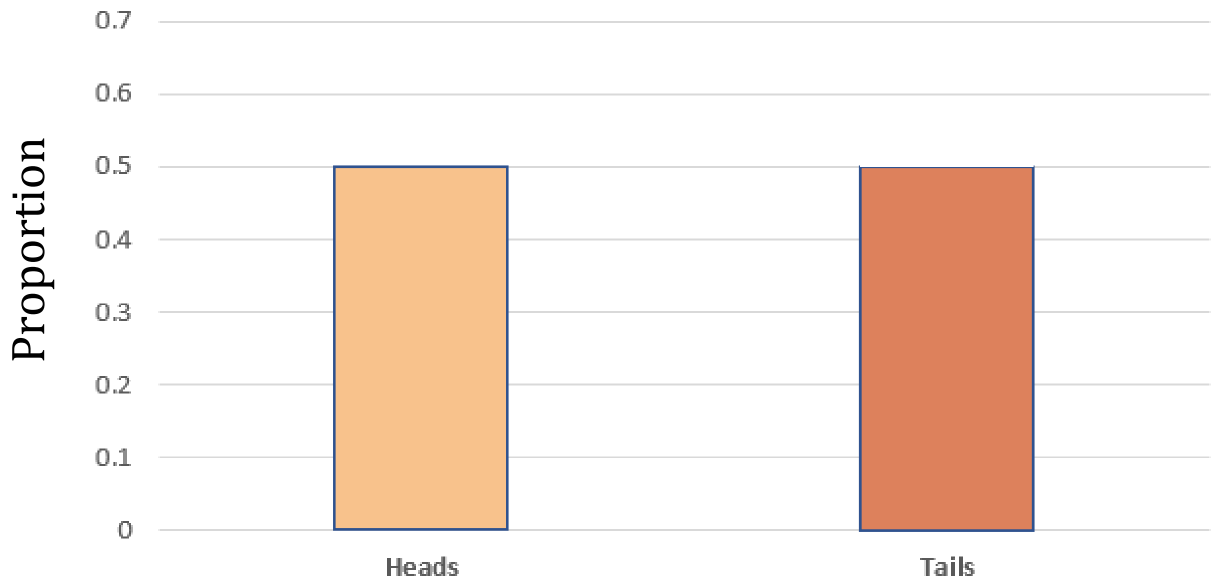
What happens if we repeat an experiment?



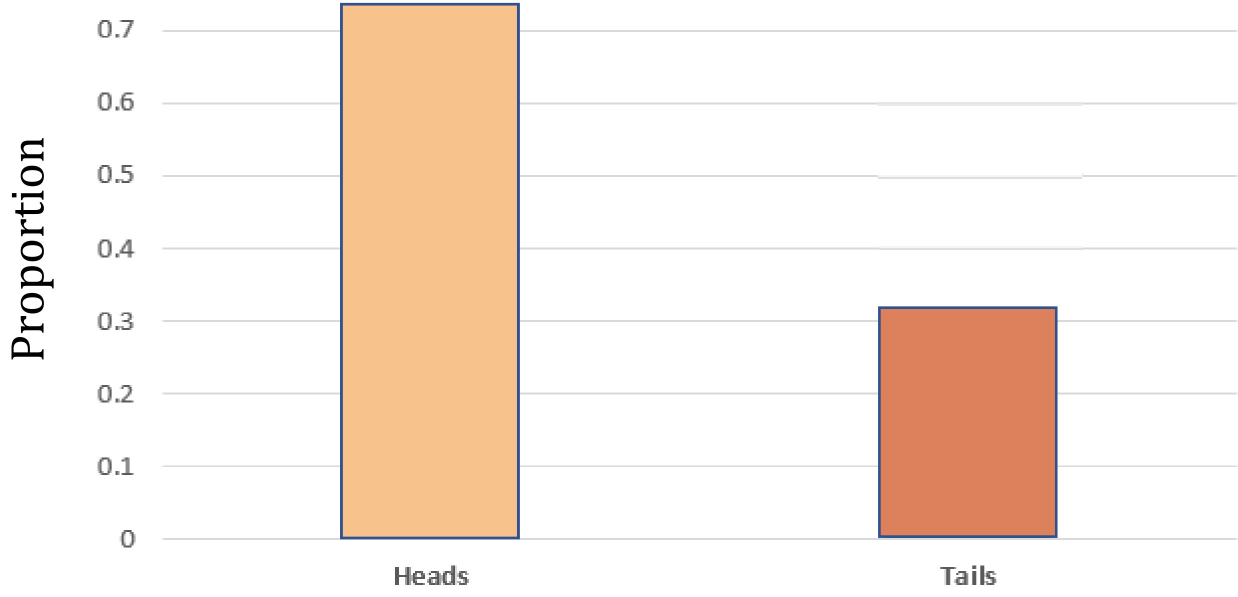
Probability

What happens if we repeat an experiment?

Theoretical proportion of Heads and Tails from 100 Coin Tosses



Empirical proportions of Heads and Tails from 100 Coin Tosses



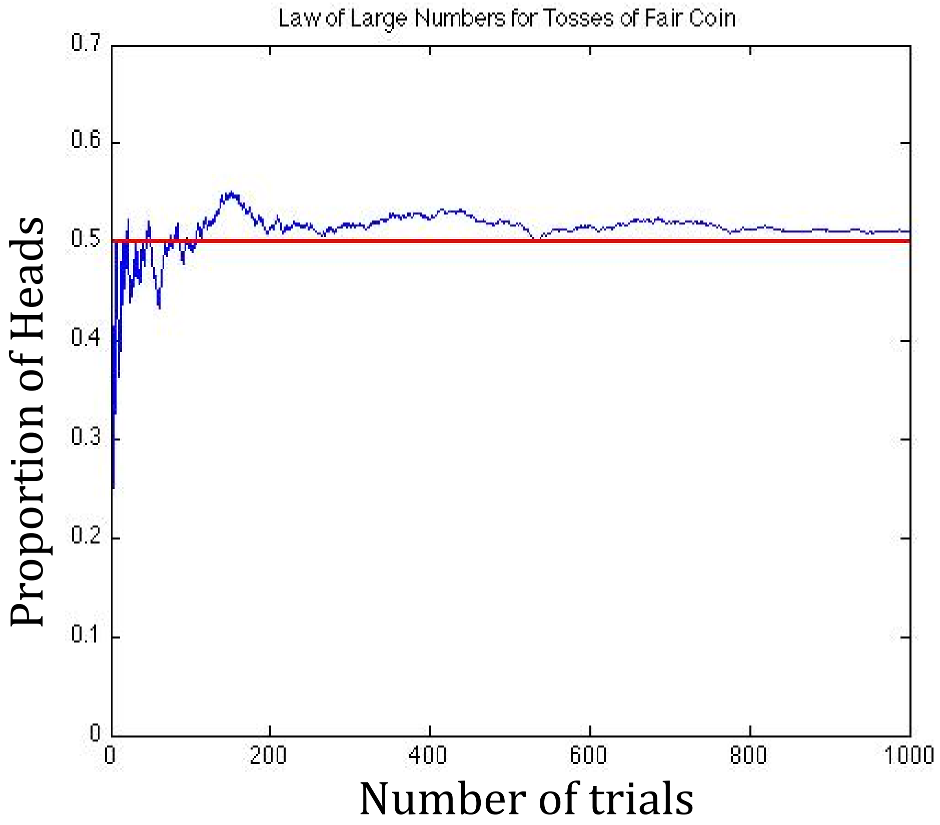
Each experiment of 100 trials would result in different proportions

Probability

What happens if we repeat an experiment?

Law of Large Numbers

If we repeat an experiment <u>a large number of times</u>, then the **empirical** probability of a particular outcome is likely to be close to the **theoretical** probability of the outcome.



Probability

What is a conditional probability?

When repeating an experiment, the next outcome could rely on what outcome previously occurred.

Example: Toss a coin: I get heads. Toss it again.

What is the probability of getting heads this second time?

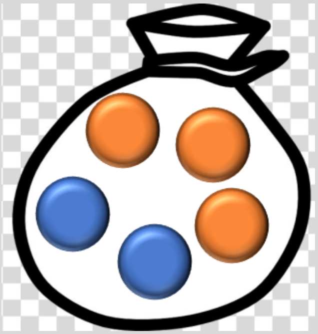
Probability

What is a conditional probability?

When repeating an experiment, the next outcome could rely on what outcome previously occurred.

Example: Pick a marble from a bag with 2 blue marbles and 3 orange.

- ❑ **What is the probability of picking a blue marble?**
- ❑ **If I first picked a blue marble, put it in my pocket, what is the probability that I pick a blue marble on second pick?**



Probability

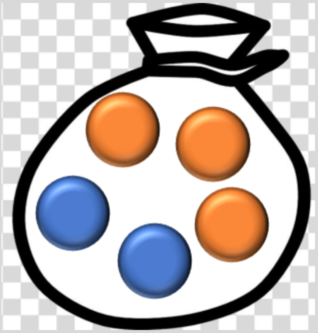
What are independent events?

Events are independent if the probability of one event does not impact the probability of another event.

Example: Pick a marble from a bag with 2 blue marbles and 3 orange.

Which events are independent?

- ❑ **Picking a blue marble, then drawing another without replacement (put it in my pocket)**
- ❑ **Picking a blue marble, then drawing another with replacement (put it back in the bag)**



Probability Distribution

What happens if we repeat an experiment with independent events?

Let's toss a coin twice.

- ❑ **What is the probability of getting 2 heads?**
- ❑ **What is the probability of getting exactly 1 head and 1 tail?**

Probability Distribution

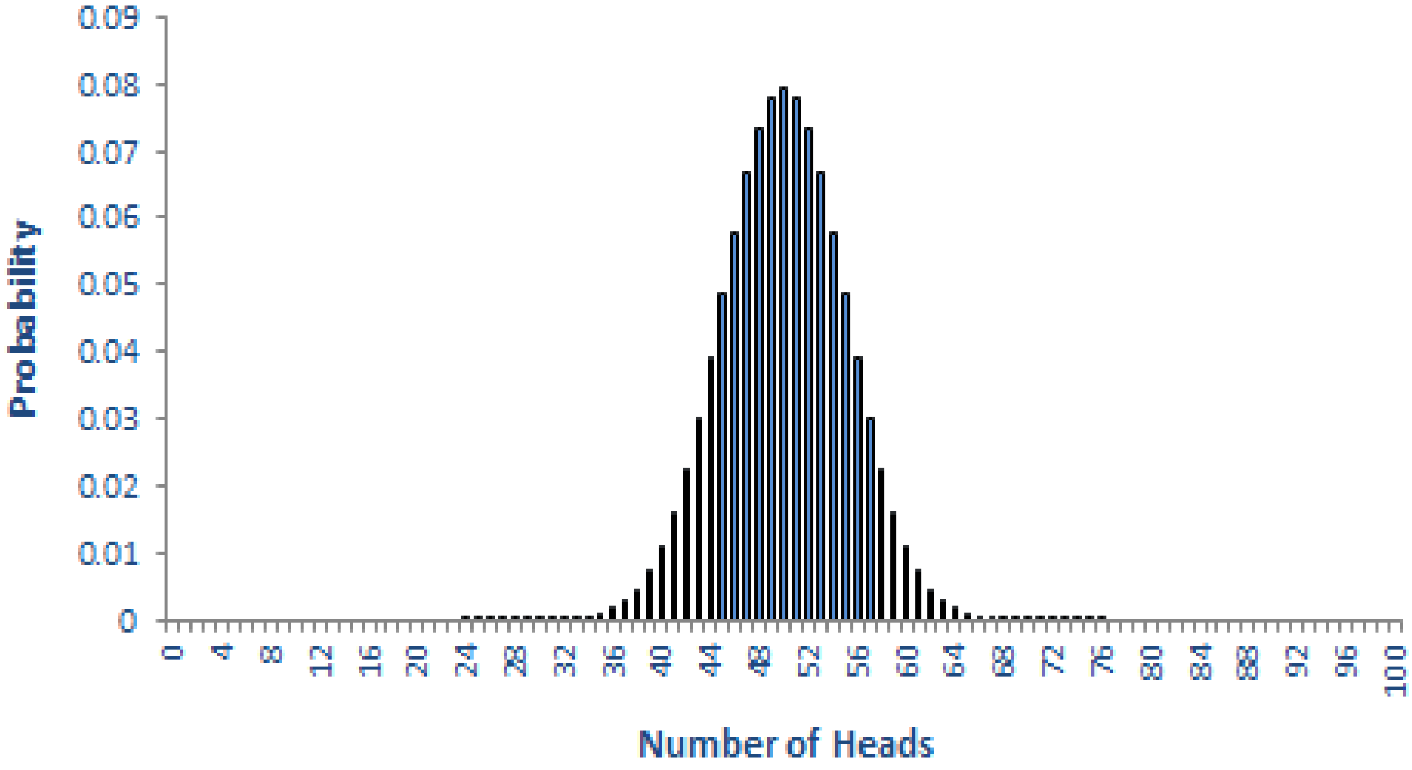
What happens if we repeat an experiment with independent events?

Let’s toss a coin 100 times and only keep track of the probability of heads.

From the graph, what is:

- ❑ **the probability to get exactly 50 heads?**
- ❑ **the probability to get exactly 2 heads?**
- ❑ **the probability to get at least 2 heads?**

Expected probability of Heads
from 100 Coin Tosses



Normal distribution

What is a normal distribution?

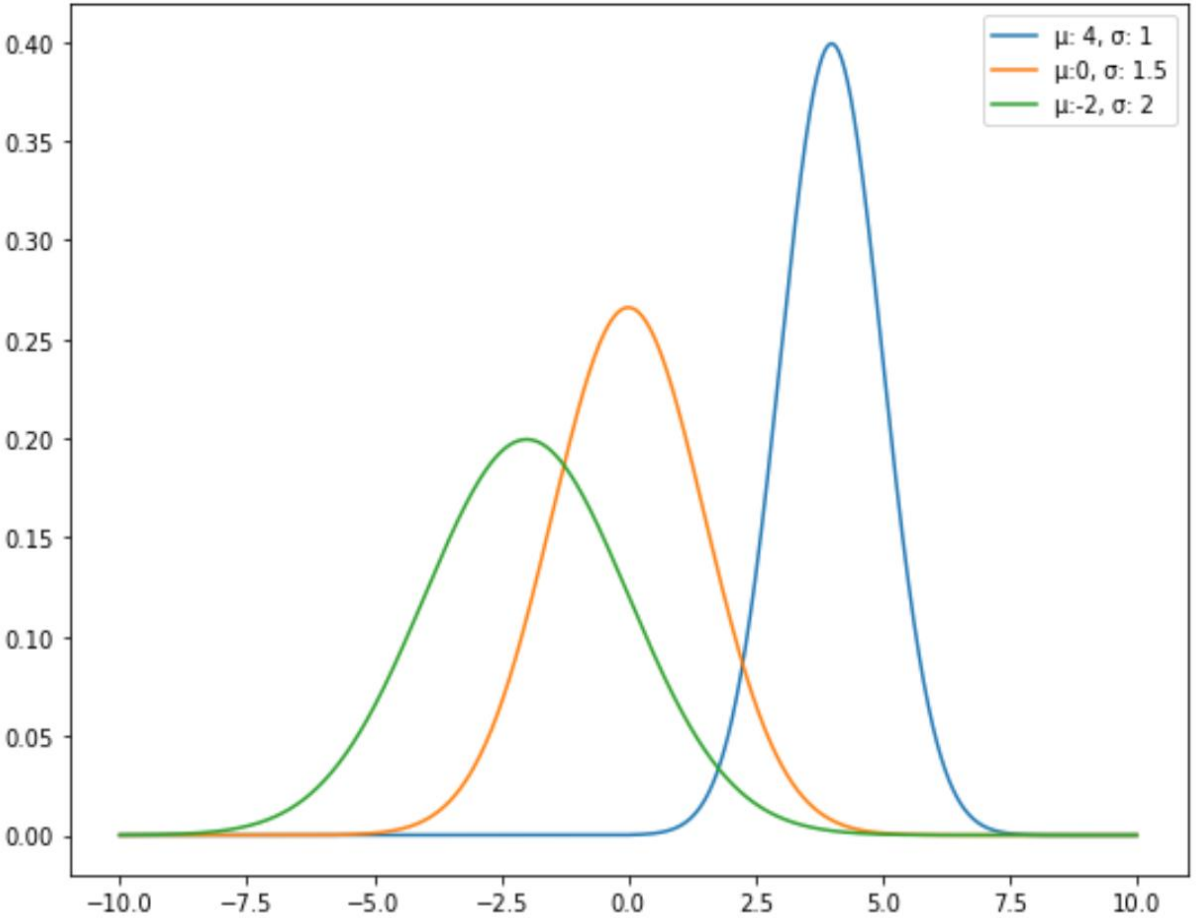
A normal distribution represents the probability of a continuous variable that is symmetric and depends on two characteristics:

Distribution center

$\boldsymbol{\mu}$: $\boldsymbol{mean}$

Distribution spread

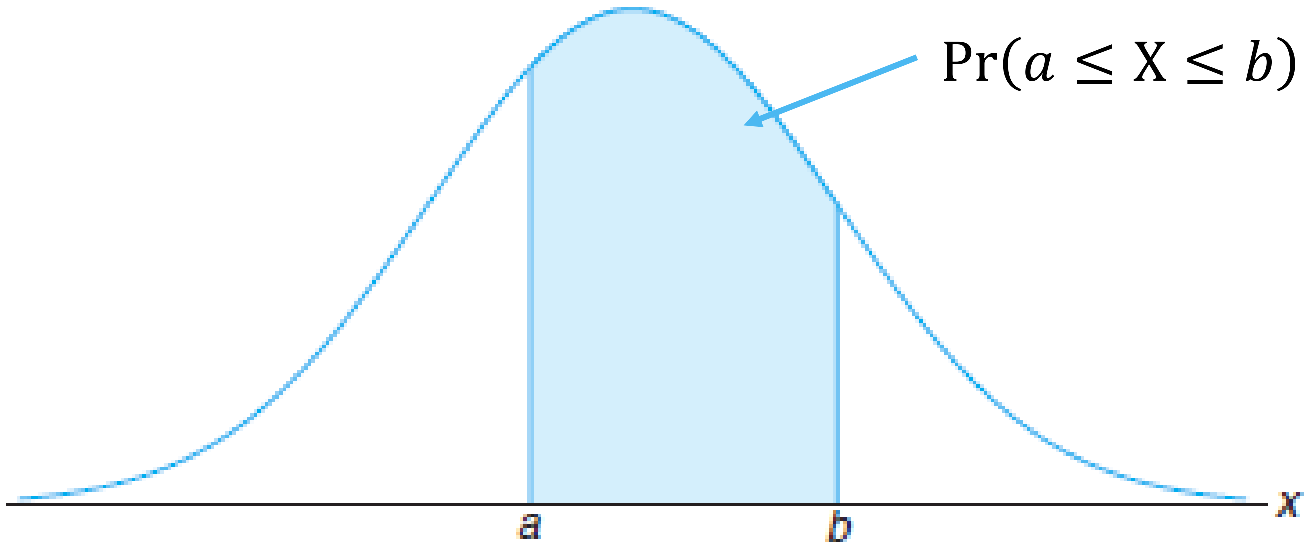
$\boldsymbol{\sigma}$: $\boldsymbol{standard\ deviation}$



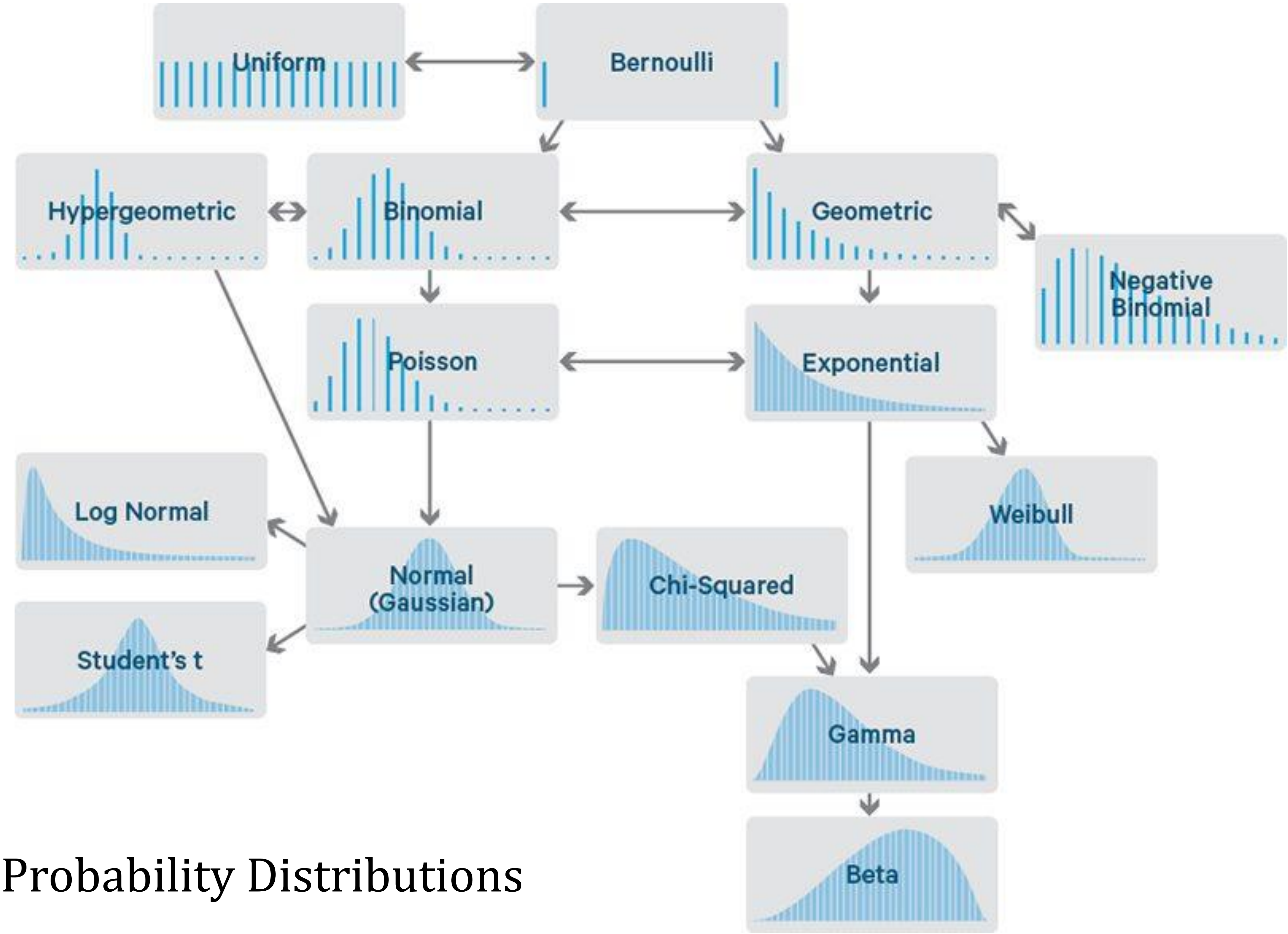
Normal distribution

What is a normal distribution?

For a continuous variable, there are infinitely many possible outcomes. This means the probability of observing any exact value is essentially zero and instead, we find the probabilities over a range of outcomes.



$$\Pr(a \leq \mathrm{X} \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



Common Probability Distributions

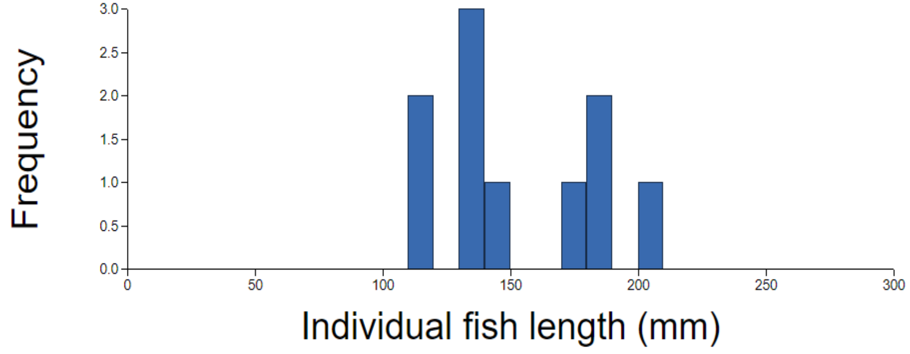
USING R AND RSTUDIO

Sampling distribution

Differences between…

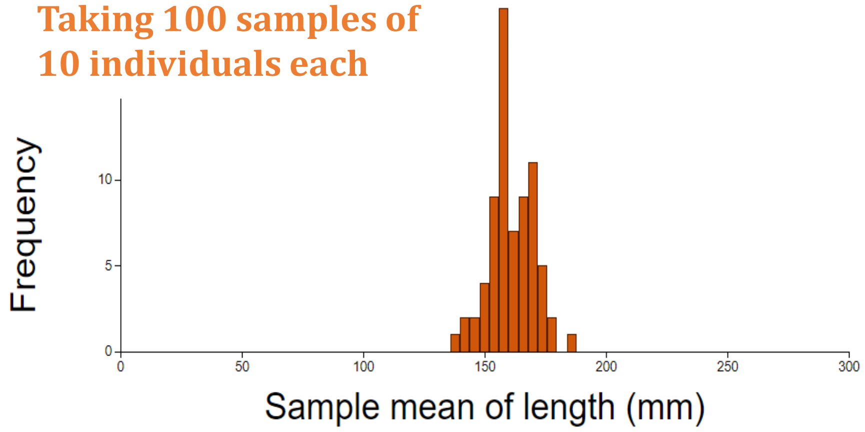
Distribution in a sample:
How values differ from individual to individual in a sample

Taking 1 sample of 10 individuals



Sampling distribution:
How means differ from sample to sample

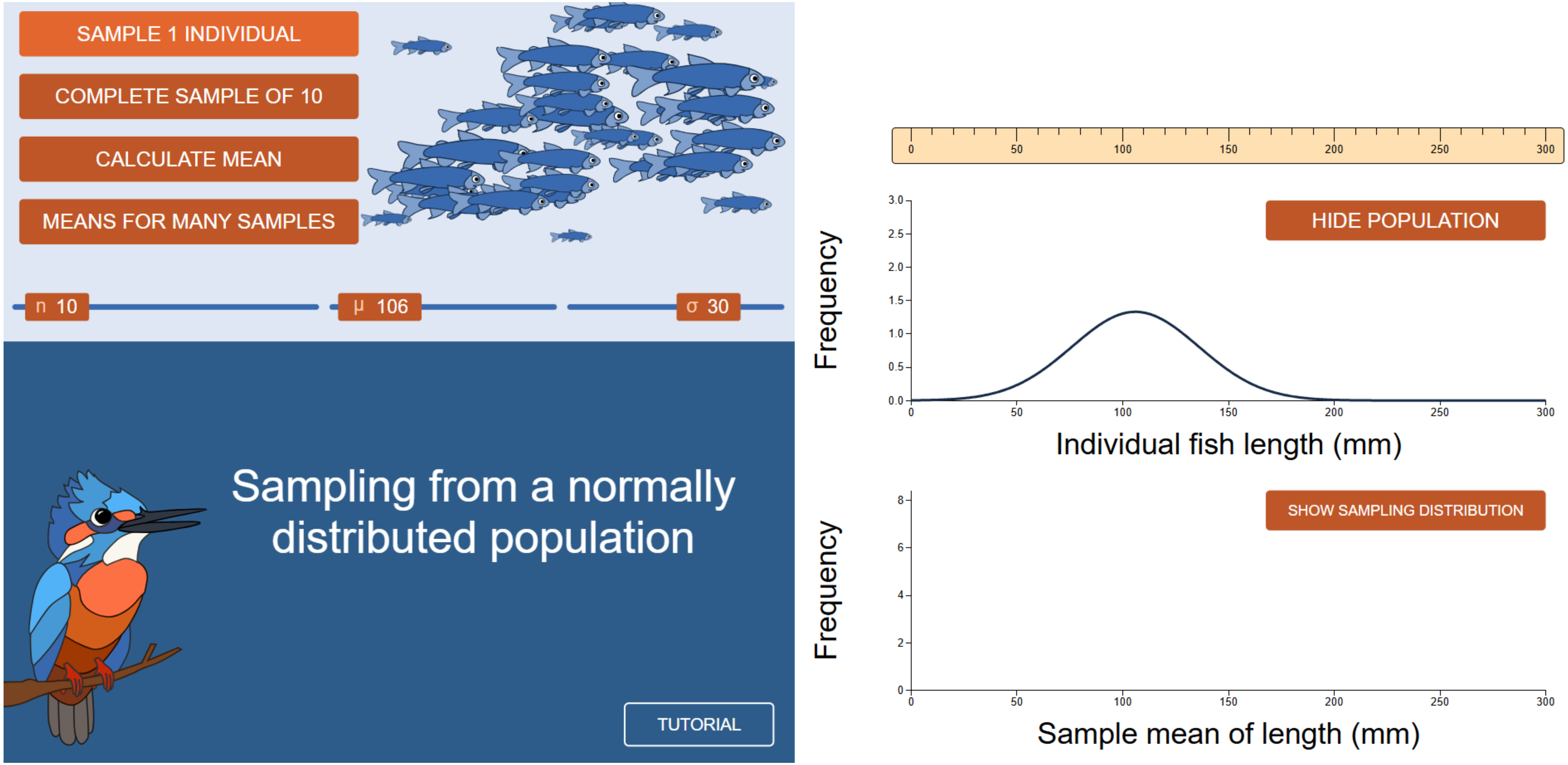
Taking 100 samples of 10 individuals each



Distribution in a population: not usually known

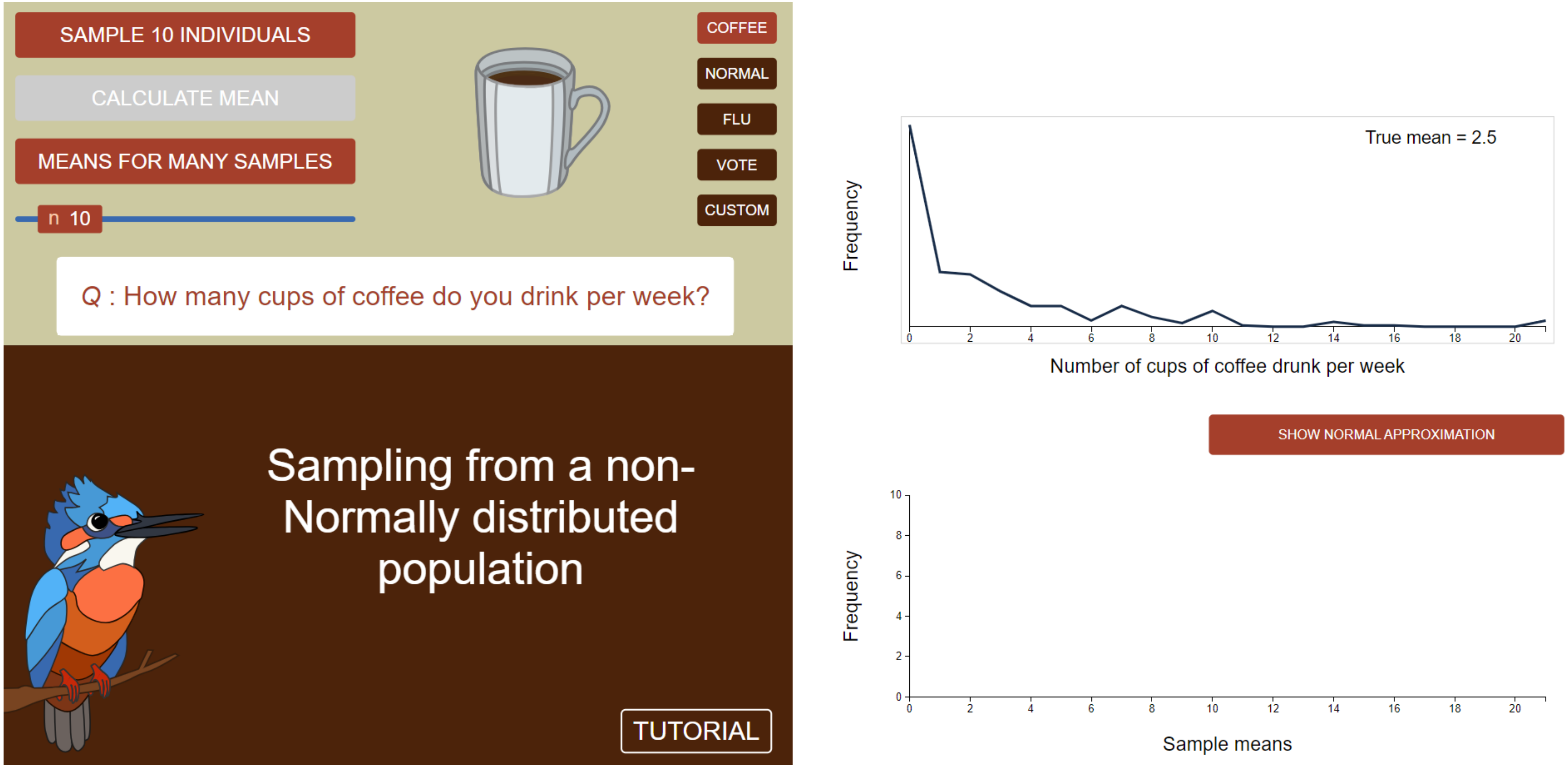
?

Sampling from a Normally Distributed Population



[Sampling from a Normal Distribution]()

Sampling from a non-Normally Distributed Population



Sampling means from a Non-normal Distribution

Central Limit Theorem

MOST IMPORTANT

Regardless of the population distribution, as the **sample size** *n* increases, the samples means of the **random** samples drawn from the population will approach a normal distribution, specifically:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Central Limit Theorem

⚠ MOST IMPORTANT

Regardless of the population distribution, as the **sample size** n increases, the samples means of the **random** samples drawn from the population will approach a normal distribution, specifically:

a normal distribution

$\sim$ follows

σ population standard deviation

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$\frac{\sigma}{\sqrt{n}}$ is called the standard error

$\bar{X}$ sample means

μ population mean

n sample size

n usually needs to be at least 30

USING R AND RSTUDIO

Making Inferences

If we know what we expect the **distribution of the sample means** to be for many random samples, we can determine:

1. How we can estimate the population mean, more or less.
2. How likely (or unlikely) it is to get a sample mean like we got *if* the population mean equals a claimed value.

Confidence Interval

Inference for a population mean

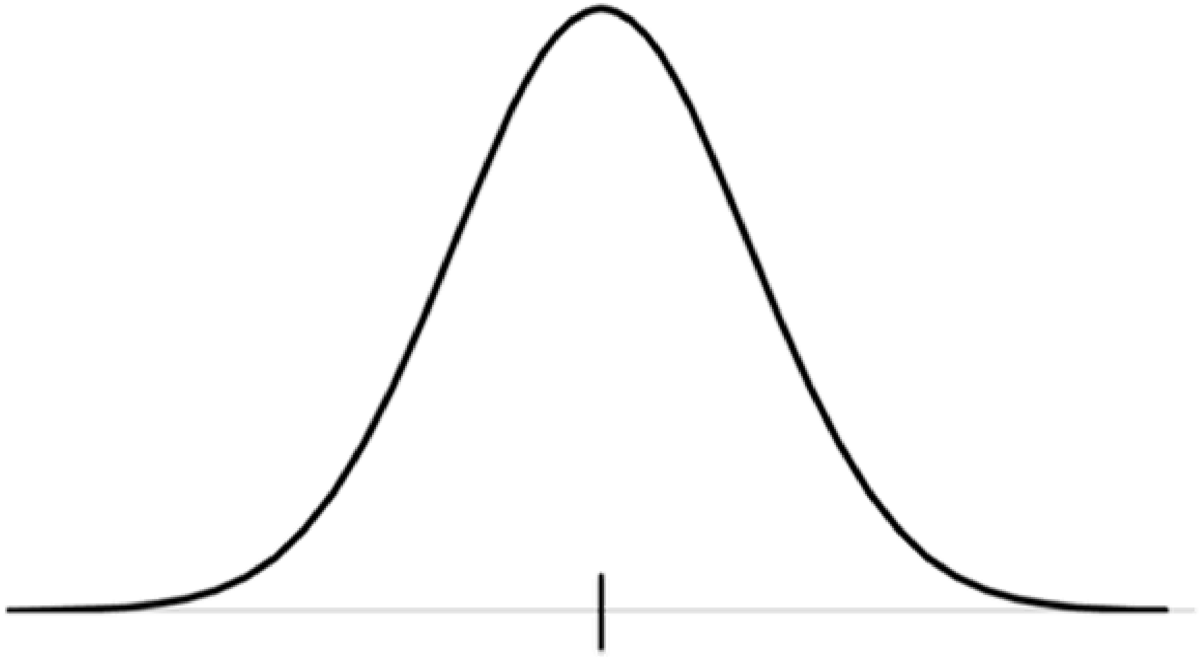
According to the Central Limit Theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Sample	**Population**
Statistics $\bar{X}$	**Parameters** μ

$\bar{X} \longrightarrow \mu$

Estimate μ by $\bar{X}$



$\mu \approx \bar{X}$

Confidence Interval

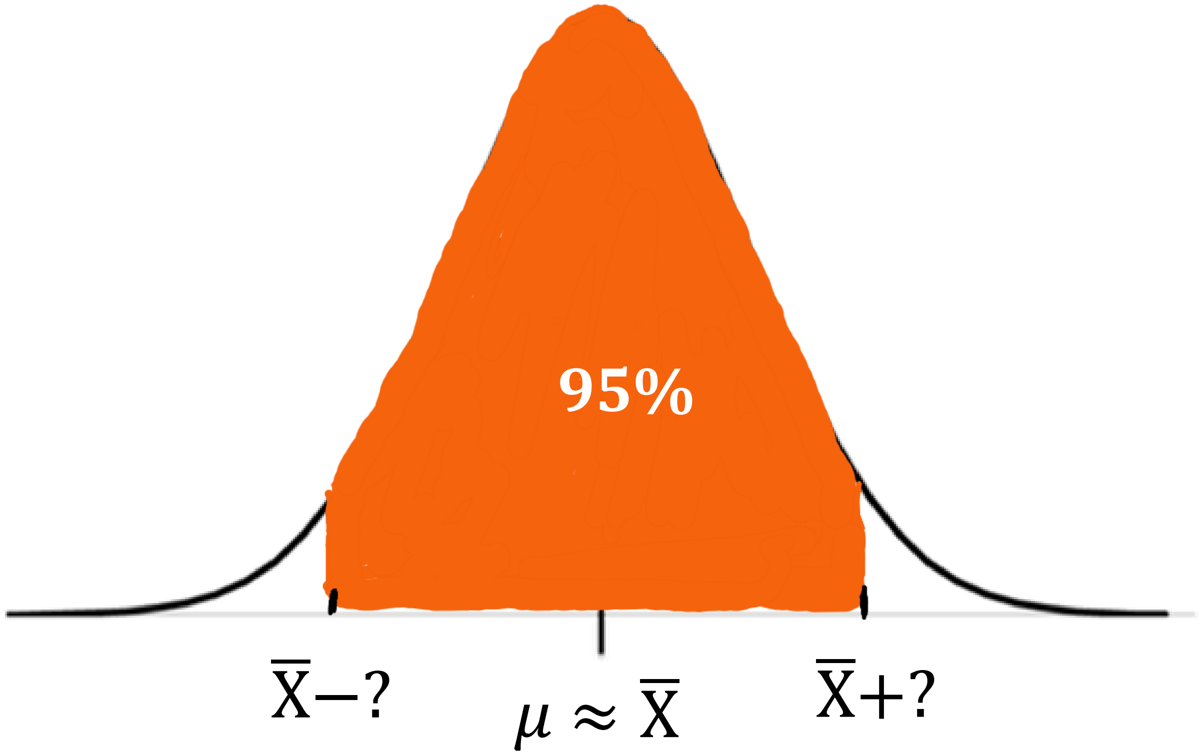
Inference for a population mean

According to the Central Limit Theorem

$$\bar{X} \sim N\left(\mu\,, \frac{\sigma}{\sqrt{n}}\right)$$

Sample	**Population**
Statistics $\bar{\mathbf{X}}$	**Parameters** $\boldsymbol{\mu}$

Estimate μ by $\bar{X}$



Confidence Interval

Inference for a population mean

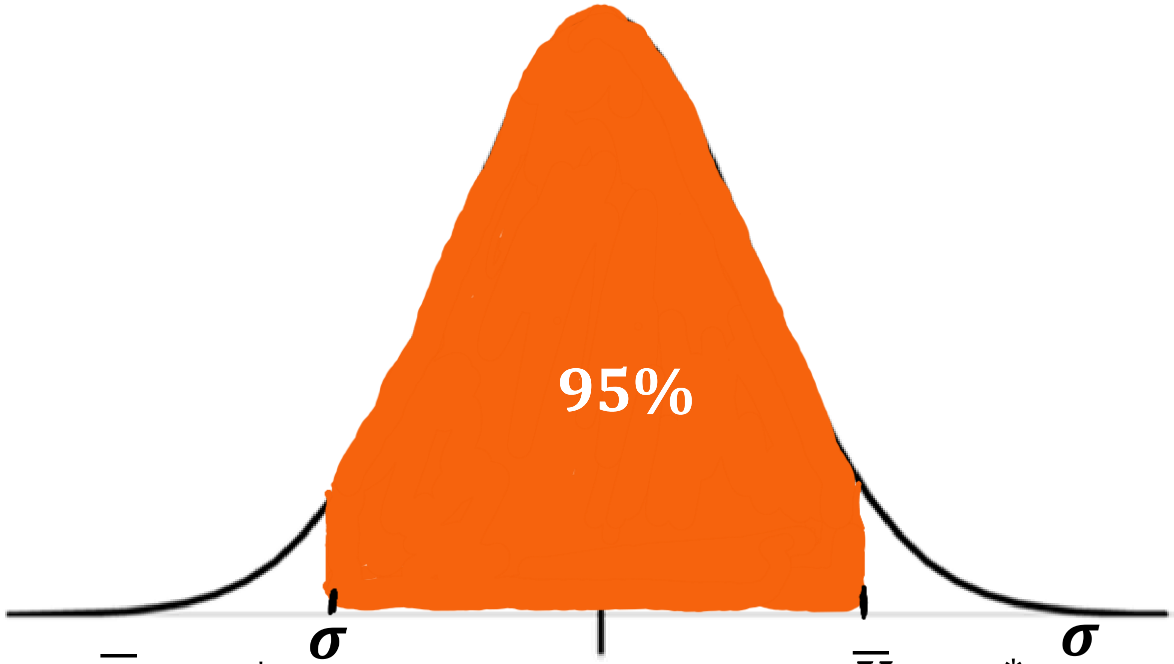
According to the Central Limit Theorem

$$\bar{X} \sim N\left(\mu\,, \frac{\sigma}{\sqrt{n}}\right)$$

Sample	**Population**
Statistics $\bar{\mathbf{X}}$	**Parameters** $\boldsymbol{\mu}$

$\bar{\mathbf{X}} \longrightarrow \boldsymbol{\mu}$

Estimate μ by $\bar{\mathrm{X}}$



$$\bar{\boldsymbol{X}} - \boldsymbol{z}^* \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} \qquad \mu \approx \bar{\mathrm{X}} \qquad \bar{\boldsymbol{X}} + \boldsymbol{z}^* \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$$

$\boldsymbol{z}^*$ is the critical value that corresponds to the z score splitting the 95% middle of the normal distribution

Confidence Interval

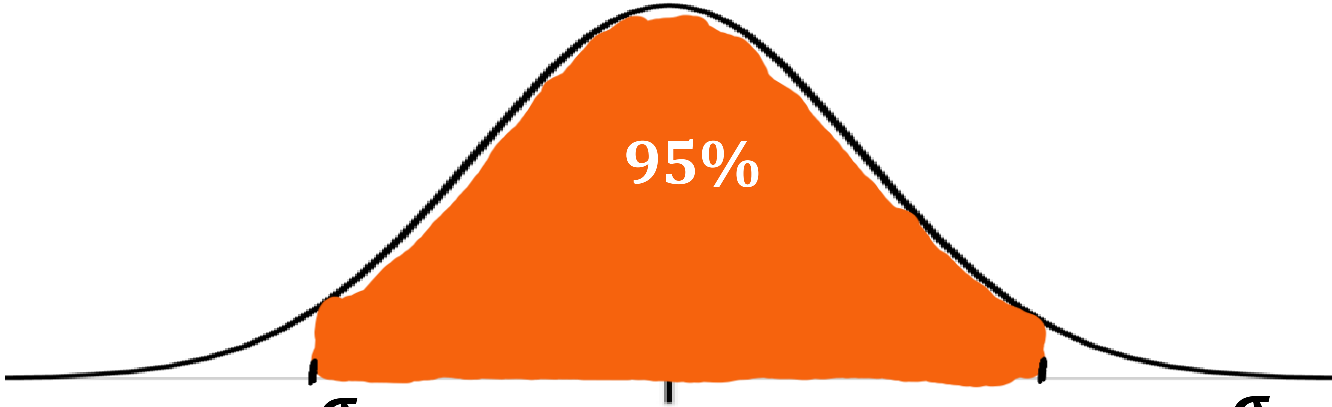
Inference for a population mean

According to the Central Limit Theorem, approximation with Student's t-distribution

$$\bar{X} \sim t_{n-1}\left(\mu\,, \frac{s}{\sqrt{n}}\right)$$

Sample	Population
Statistics	**Parameters**
$\bar{\mathbf{X}}$ →	$\boldsymbol{\mu}$
s →	$\boldsymbol{\sigma}$

Estimate μ by $\bar{\mathrm{X}}$
Estimate σ by s



$\bar{X} - t^*_{df}\dfrac{\sigma}{\sqrt{n}}$ $\mu \approx \bar{\mathrm{X}}$ $\bar{X} + t^*_{df}\dfrac{\sigma}{\sqrt{n}}$

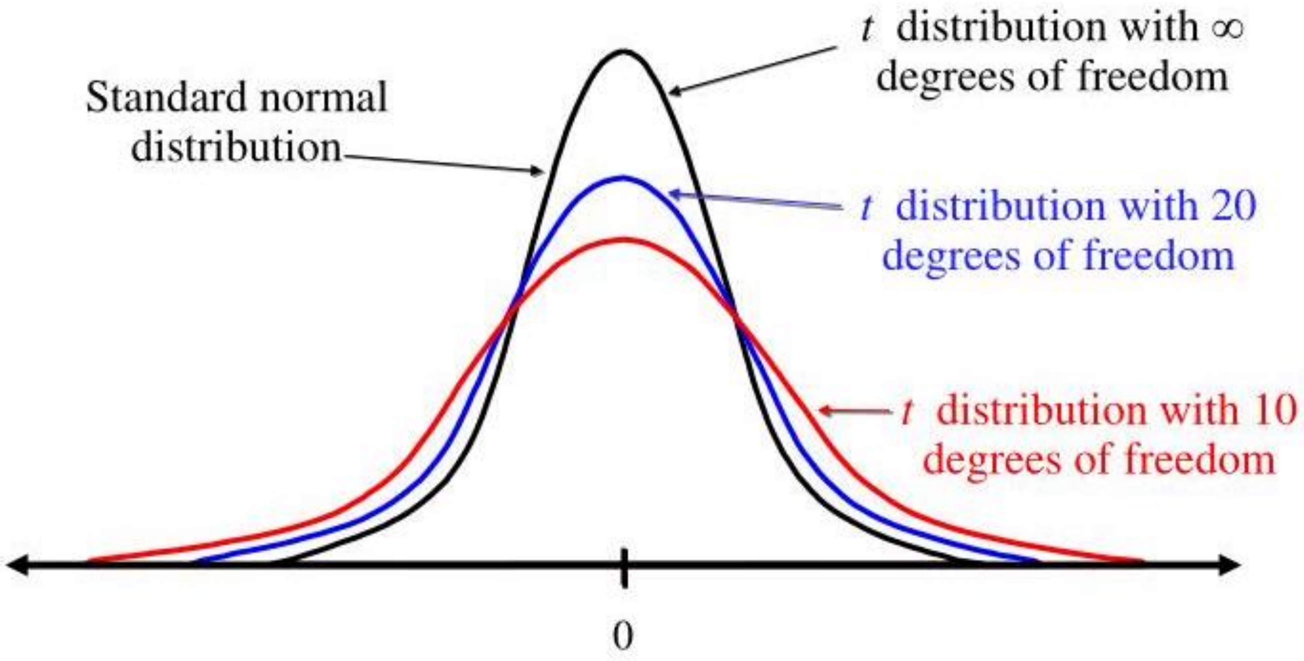
t^*_{df} is the critical value that splits the 95% middle of the Student's t-distribution

Student's t -distribution

$$t_{n-1}\left(\mu\,,\frac{s}{\sqrt{n}}\right)$$

$\boldsymbol{n-1=}$ **degrees of freedom**

The Student's t-distribution is a more conservative form of the standard normal distribution: there is a lower probability to the center and a higher probability to the tails compared to the standard normal distribution.



Developed by [William Sealy Gosset] (pen name was Student), a pioneer in modern statistics. He worked at the Guinness Brewery in Dublin, IRL, and was interested in problems with small samples (for example, chemical properties of barley).

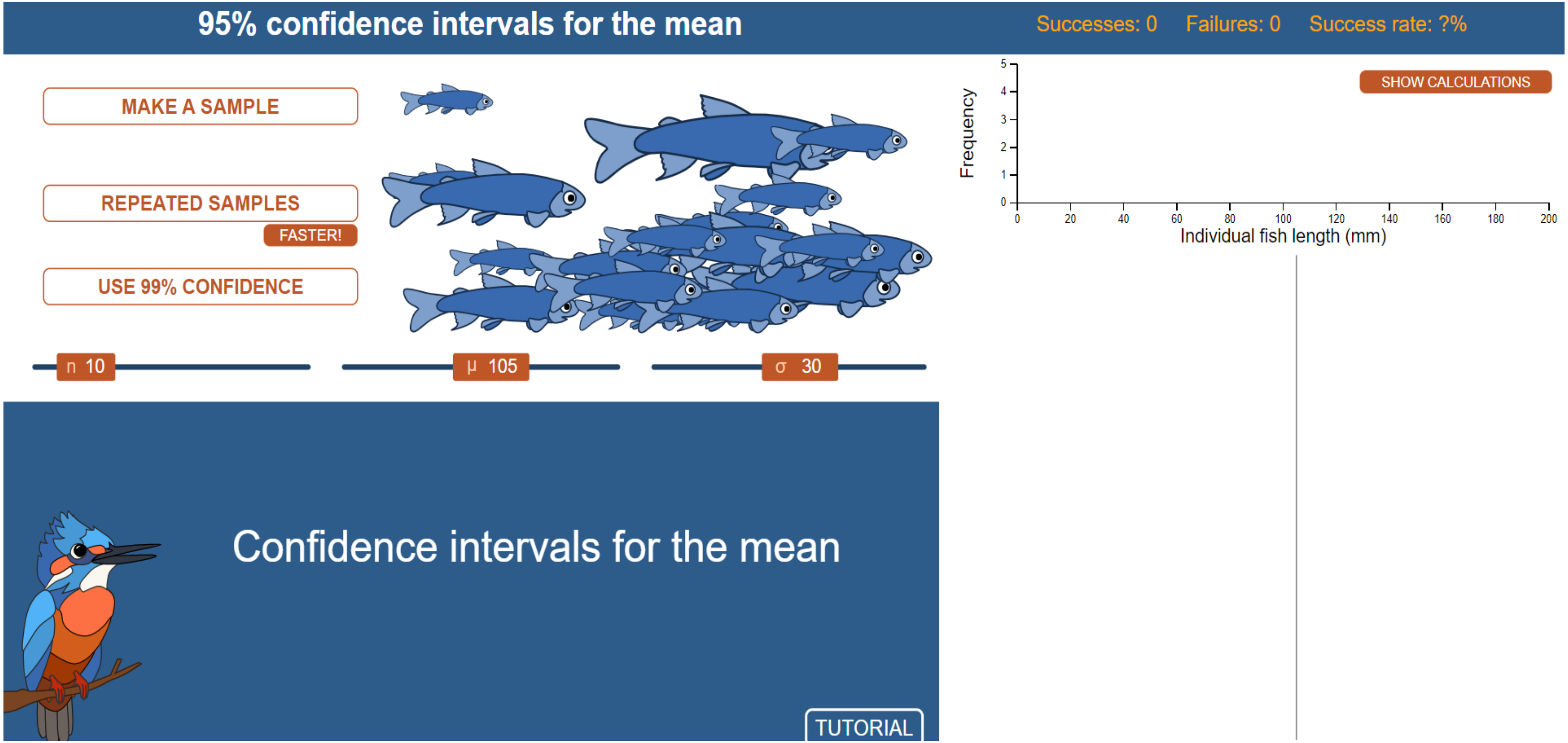
Confidence Interval

Confidence Interval for a **population mean**

$$statistic \pm margin\ of\ error$$

$$\bar{\boldsymbol{X}} \pm \underbrace{\boldsymbol{t}^{*}_{\boldsymbol{df}} \cdot \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}}_{critical\ value\ \cdot\ standard\ error}$$

Confidence Interval



[Confidence intervals of the mean]()

Hypothesis Testing

4 steps of hypothesis testing:

1. State a claim and counterclaim (hypotheses).

Null hypothesis: a statement of no difference, no effect, no relationship.
Alternative hypothesis: a statement that contradicts the null hypothesis.

2. Use sample data to calculate an estimate of a parameter.

Find descriptive **statistics**

3. Compare the estimate to the claim.

Suppose the **null hypothesis is TRUE**. If we had many random samples from the population, how **likely** were we to observe what we observed?

4. Make a decision: is there enough evidence to disprove the null hypothesis, or not.

Interpret the results of the test in context, citing appropriate statistics.

Hypothesis Testing

4 steps of hypothesis testing:

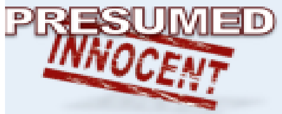
1. State a claim and counterclaim (hypotheses).

2. Use sample data to calculate an estimate of a parameter.

3. Compare the estimate to the claim.

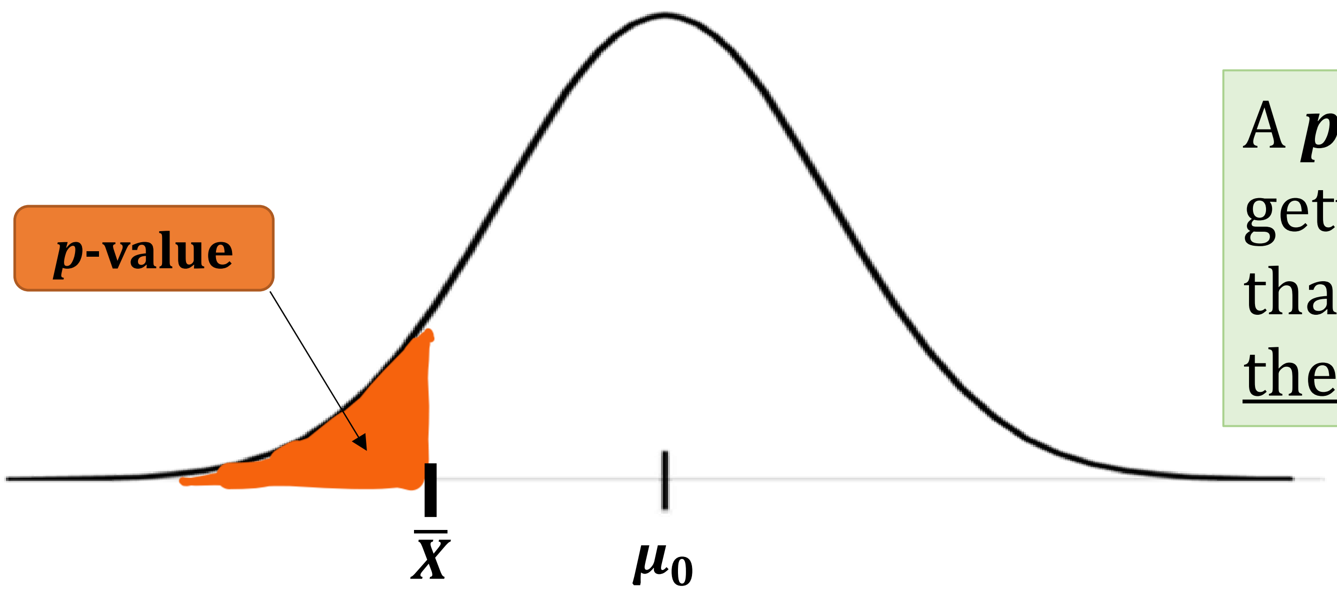
4. Make a decision: is there enough evidence to disprove the null hypothesis, or not.

Analogy with a criminal trial:



Hypothesis Testing

4. Make a decision.



Sampling distribution **if** H_0 was true
(according to the Central Limit Theorem)

A ***p*-value** is the probability of getting a statistic more extreme than the one we observed <u>under the null hypothesis</u>.

Hypothesis Testing

How to conclude:

➢ If the p-value is small: there is <u>enough</u> evidence <u>against</u> H_0

Reject H_0 ⟶ **significant results**

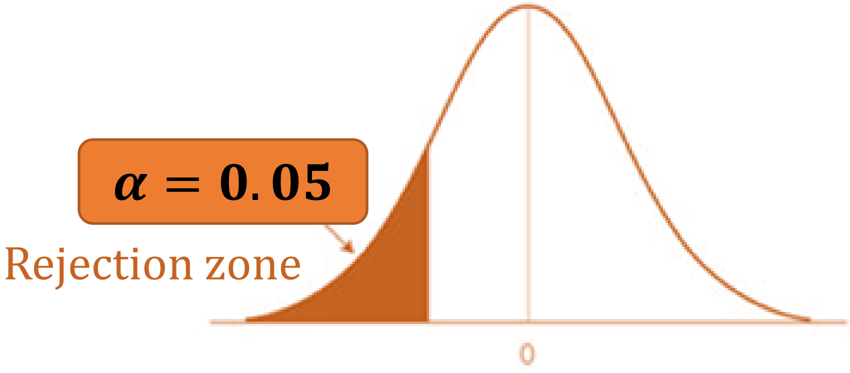
➢ If the p-value is not small: there is <u>not enough</u> evidence <u>against</u> H_0

Fail to reject H_0 ⟶ **not significant results**

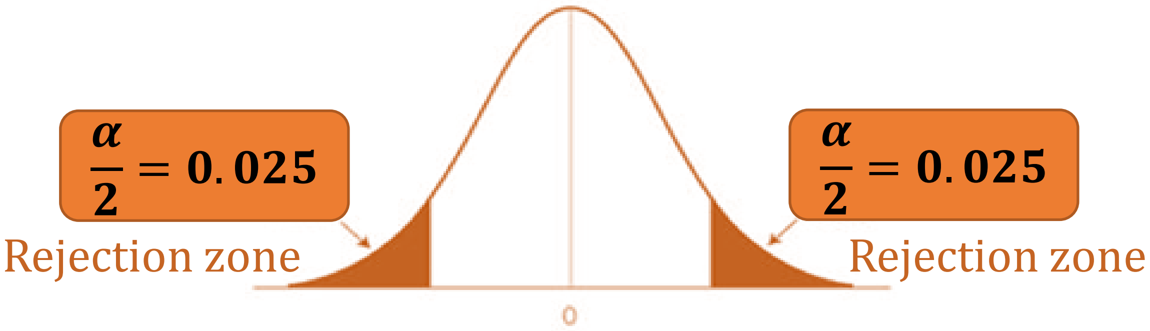
What is "small"?

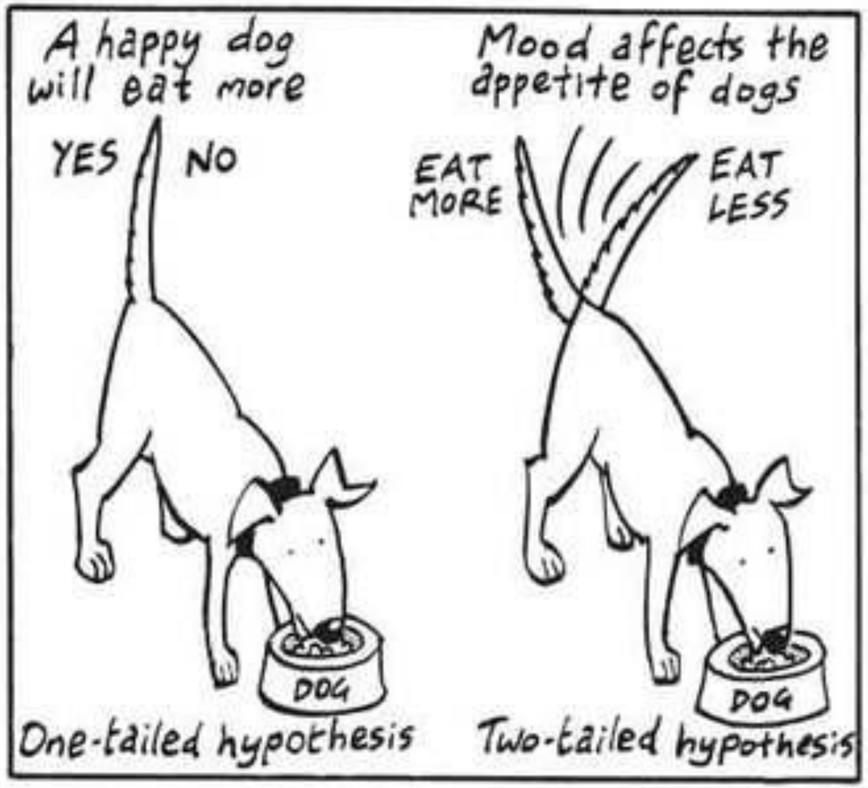
Hypothesis Testing

One-sided test: testing for extreme values on one side



Two-sided test: testing for extreme values on either sides





Why are one-sided tests not being accepted as frequently by most scientific journals?

Hypothesis Testing

Decisions can be correct or incorrect…

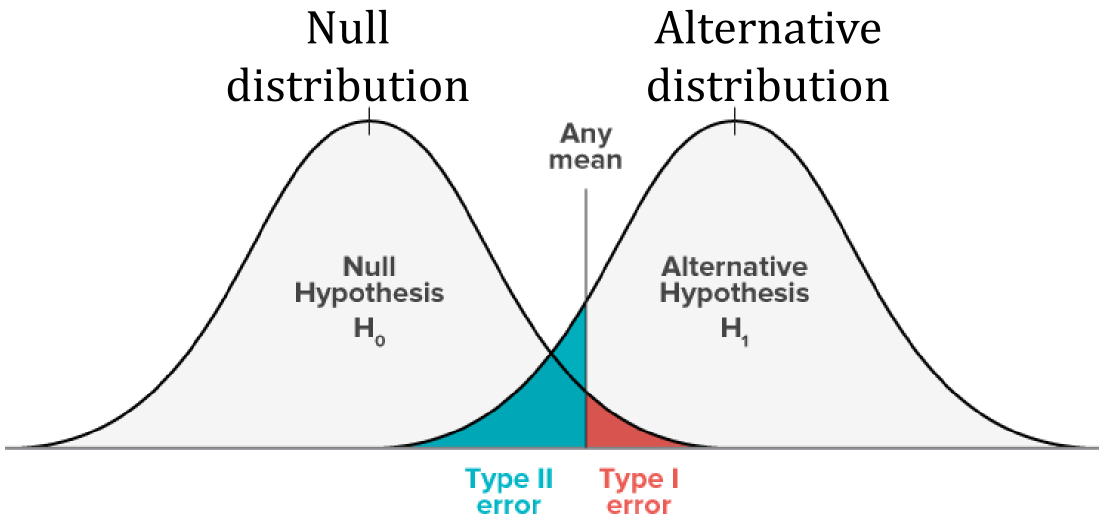
Reality

Test Decision	H_0 is true	H_0 is false
Fail to reject H_0	Correct Decision	Incorrect Decision **β = Type II Error**
Reject H_0	Incorrect Decision **α = Type I Error**	Correct Decision **Power 1-β**

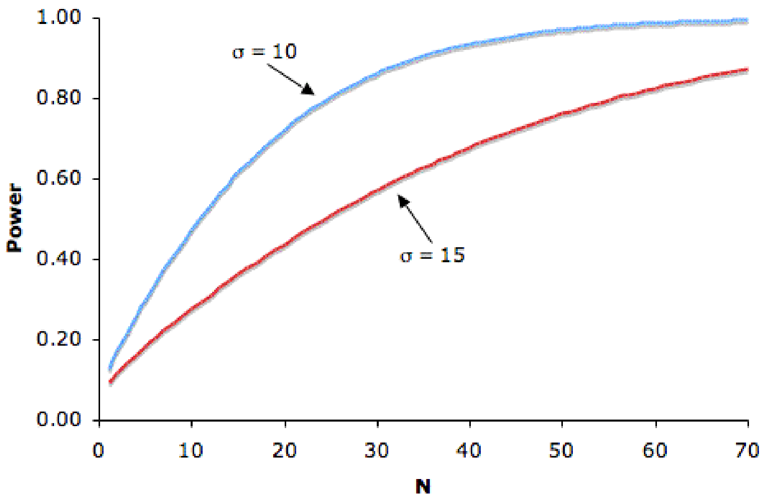
Based on the analogy with a criminal trial: what would it mean to make a Type I error? to make a Type II error? Which one would be "worse"?

Hypothesis Testing

- Type I and Type II errors are inversely related



- The power of a test increases as the sample size increases (and as the variation decreases)



Type I error (false positive) Type II error (false negative)

BREAK TIME

BACK AT …

One sample t-test

Comparing a population mean to a hypothesized value $\boldsymbol{\mu_0}$

1. State your hypotheses

In terms of the <u>difference</u> between the two variables:

H_0: The population mean **is the same** as the hypothesized value.

$$\mu = \mu_0$$

H_A: The population mean **is not the same** as the hypothesized value.

$$\mu \neq \mu_0$$

One sample t-test

Comparing a population mean to a hypothesized value $\boldsymbol{\mu_0}$

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{t}$ (based on sample data)

$$\boldsymbol{t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}}$$

One sample t-test

Comparing a population mean to a hypothesized value $\boldsymbol{\mu_0}$

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{t}$ (based on sample data)

3. Compare test statistic to null distribution (calculate $\boldsymbol{p}$**-value**)

4. Make a conclusion in context, reporting the appropriate statistics (t, df, p-value).

$$\boldsymbol{df = n_1 + n_2 - 2}$$

One sample t-test

Estimating a population mean

When reporting results of a significant test, also report a measure of the effect size with a **confidence interval** of the **population mean**:

$$\boldsymbol{\bar{X} \pm t^*_{df} \cdot \frac{s}{\sqrt{n}}}$$

One sample t-test

Comparing a population mean to a hypothesized value μ_0

Check assumptions:

✓ Random sample

✓ Independent observations

✓ The variable of interest is (approximately) normally distributed

USING R AND RSTUDIO

Independent t-test

Comparing two population means between 2 groups

1. State your hypotheses

In terms of the difference between the two variables:

H_0: The mean of group 1 **is the same** as the mean of group 2.

$\mu_1 = \mu_2$ equivalent to $\mu_1 - \mu_2 = 0$

H_A: The mean of group 1 **is not the same** as the mean of group 2.

$\mu_1 \neq \mu_2$ equivalent to $\mu_1 - \mu_2 \neq 0$

Independent t-test

Comparing two population means between 2 groups

1. State your hypotheses
2. Calculate the test statistic $\boldsymbol{t}$ (based on sample data)

$$\boldsymbol{t = \frac{\bar{X}_1 - \bar{X}_2 - 0}{SE_{\bar{X}_1 - \bar{X}_2}}}$$

Independent t-test

Comparing two population means between 2 groups

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{t}$ (based on sample data)

$$SE_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

pooled variance: $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Independent t-test

Comparing two population means between 2 groups

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{t}$ (based on sample data)

3. Compare test statistic to null distribution (calculate ***p*-value**)

4. Make a conclusion in context, reporting the appropriate statistics (*t*, *df*, *p*-value).

$$\boldsymbol{df = n_1 + n_2 - 2}$$

Independent t-test

Comparing two population means between 2 groups

When reporting results of a significant test, also report a measure of the effect size with a **confidence interval** of the **true difference**.

$$\bar{X}_1 - \bar{X}_2 \pm t^*_{df} \cdot SE_{\bar{X}_1 - \bar{X}_2}$$

Independent t-test

Comparing two population means between 2 groups

Check assumptions:

✓ Random sample

✓ Independent observations

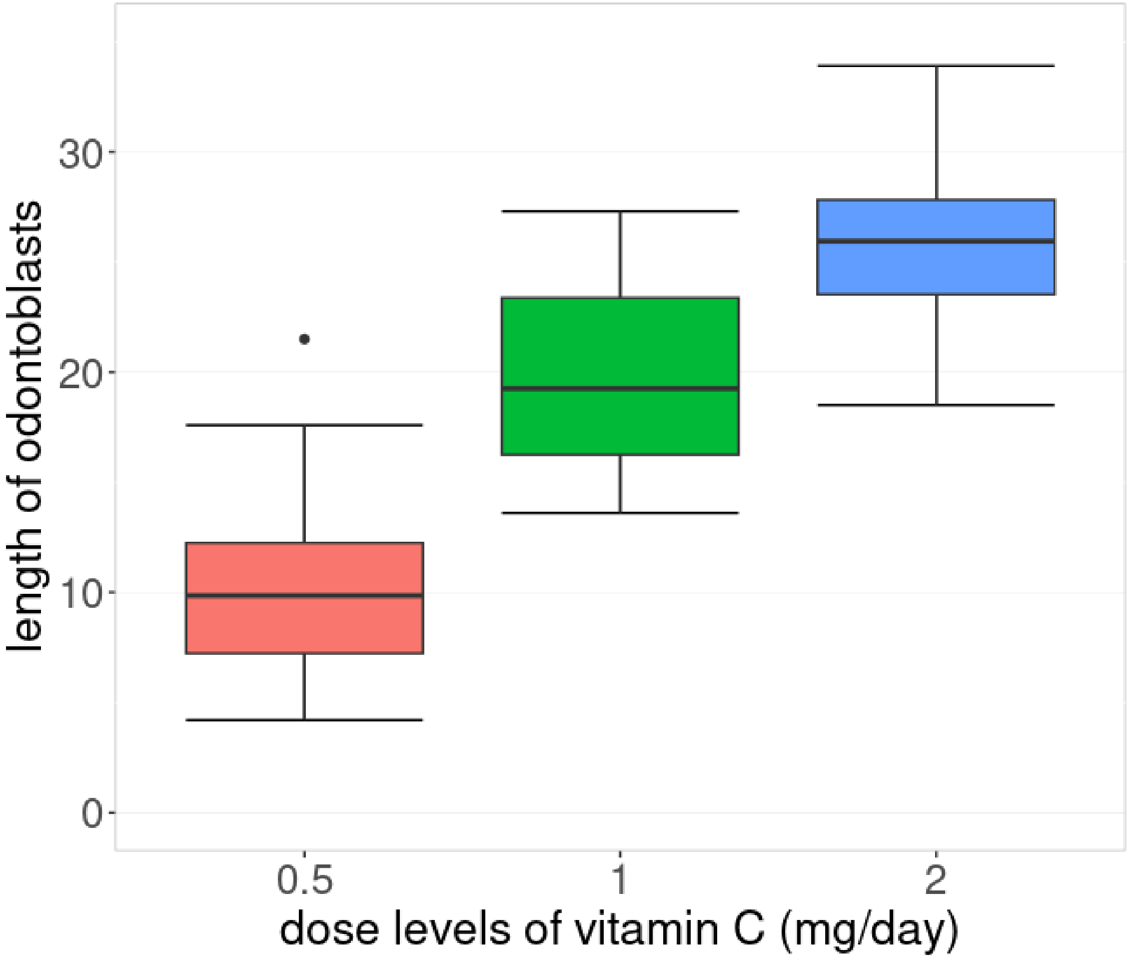
✓ The distribution in each group is normally distributed

✓ The two distributions have equal variance ($\sigma_1^2 = \sigma_2^2$)

USING R AND RSTUDIO

ANOVA (ANalysis Of VAriance)

Comparing population means across *k* groups

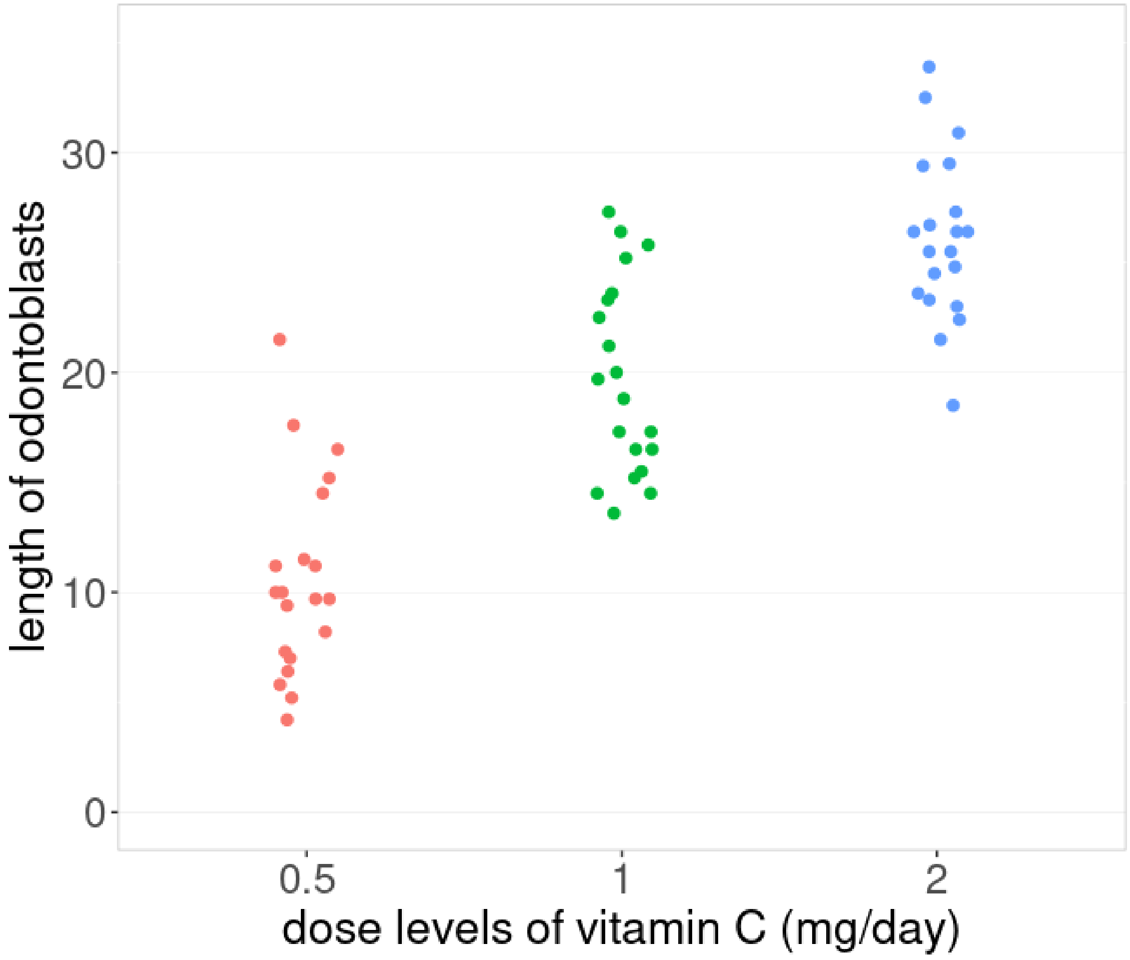


A researcher wants to compare the teeth length for the 60 guinea pigs which received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day).

Why can't we just conduct an Independent t-test in this scenario?

ANOVA (ANalysis Of VAriance)

Comparing population means across *k* groups

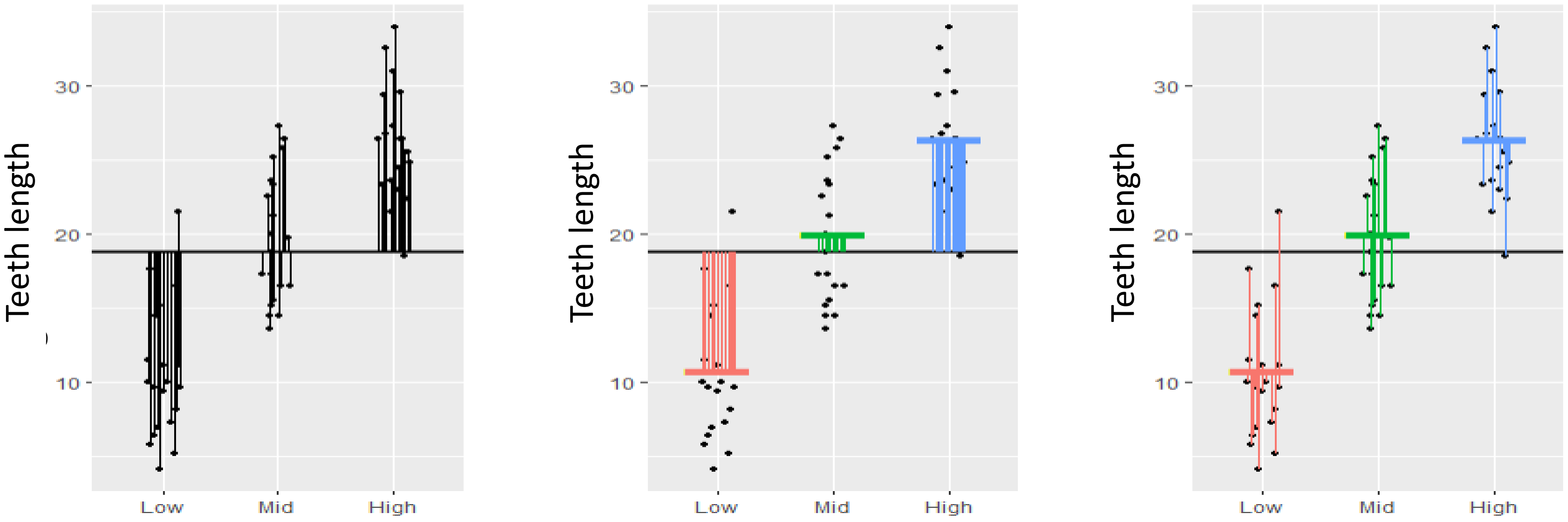


A researcher wants to compare the teeth length for the 60 guinea pigs which received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day).

Why can't we just conduct an Independent t-test in this scenario?

ANOVA (ANalysis Of VAriance)

Decomposing variation across k groups

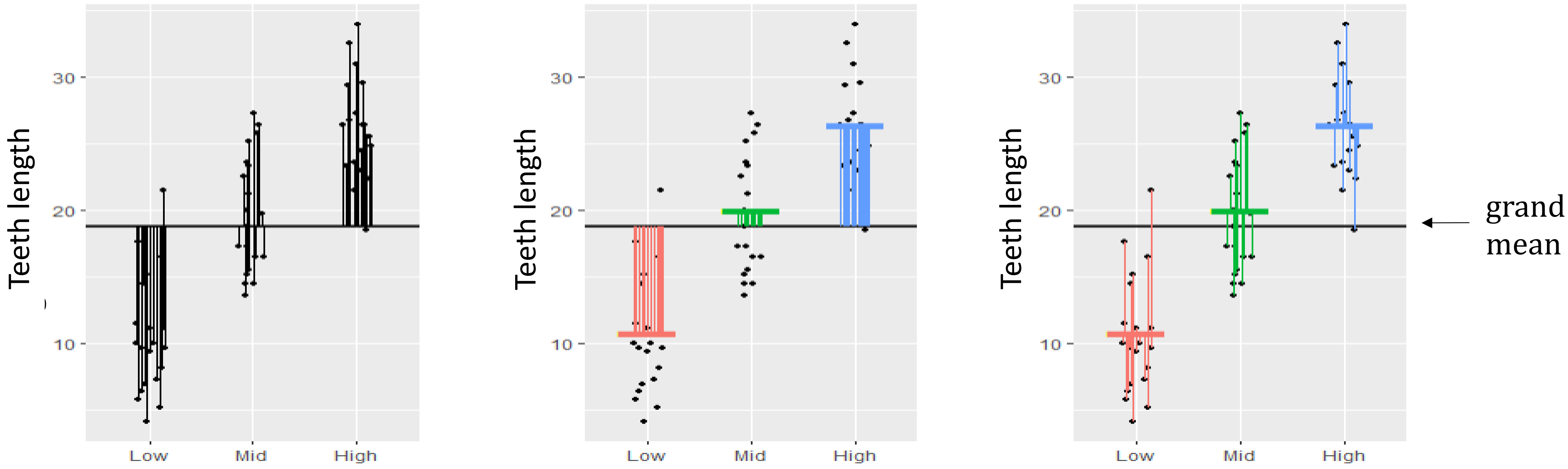


$$SS_{total} = SS_{group} + SS_{error}$$

Total Variation = **Variation between groups** + **Variation within groups**

ANOVA (ANalysis Of VAriance)

Decomposing variation across k groups

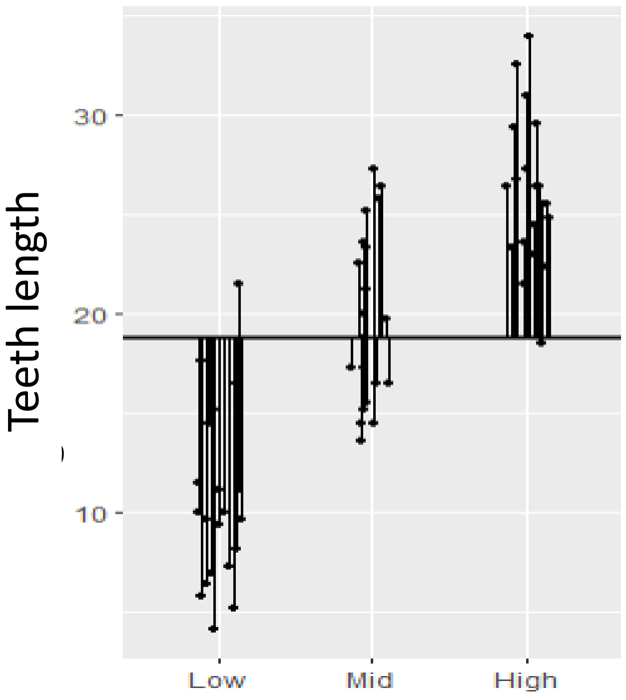


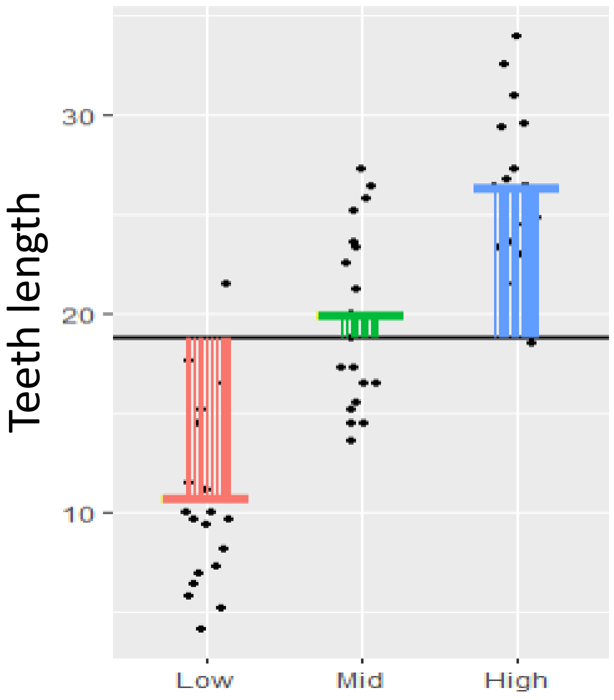
SS_{total} = SS_{group} + SS_{error}

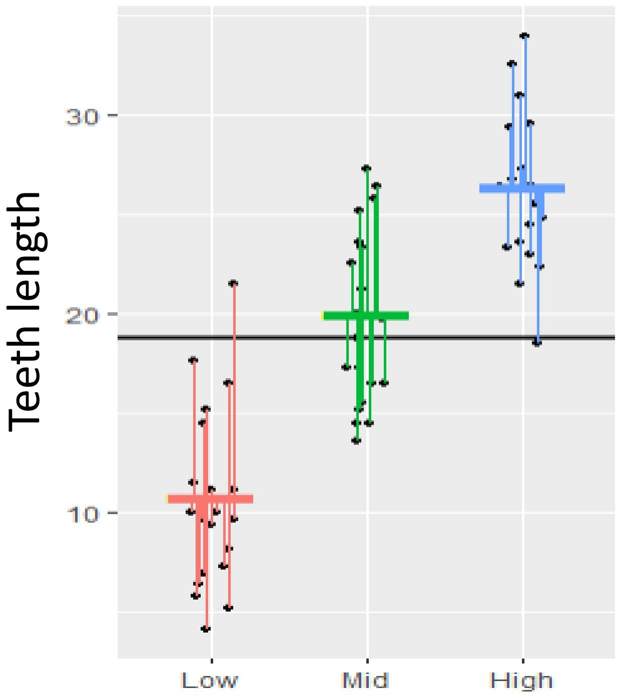
Total Variation | **Variation between groups** | **Variation within groups**

ANOVA

Decomposing variation across *k* groups







SS_{total} = SS_{group} + SS_{error}

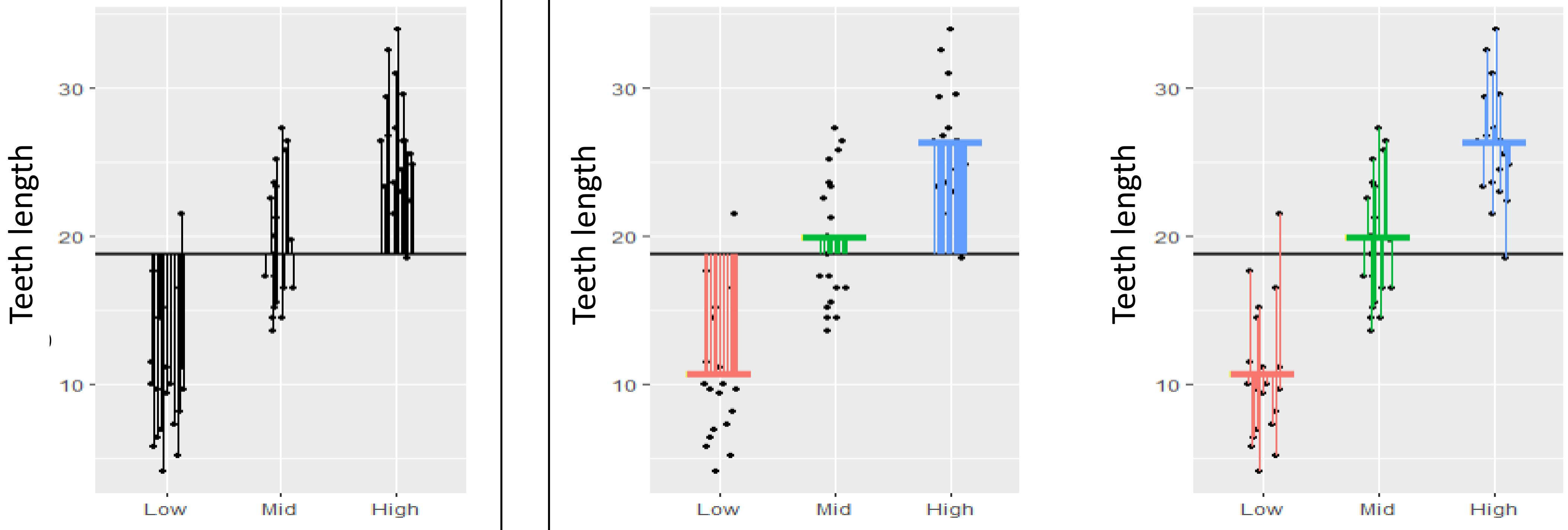
Total Variation
Compare each observation compared to grand mean

Variation between groups
Compare each group mean to grand mean

Variation within groups
Compare each observation to its group mean

ANOVA

Decomposing variation across *k* groups



SS_{total} = SS_{group} + SS_{error}

$$\sum (X_{ik} - \bar{X})^2 = \sum n_k (\bar{X}_k - \bar{X})^2 + \sum\sum (X_{ik} - \bar{X}_k)^2$$

ANOVA (ANalysis Of VAriance)

Comparing population means across *k* groups

1. State your hypotheses

H_0: The means are all equal, $\mu_1 = \mu_2 = \ldots = \mu_k$

H_A: Not all the means are equal

ANOVA

Comparing population means across *k* groups

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{F}$ (based on sample data)

$$\boldsymbol{F = \frac{MS_{group}}{MS_{error}}}$$ with $$\boldsymbol{MS_{group} = \frac{SS_{group}}{df_{group}}} \longleftarrow k-1$$

$$\boldsymbol{MS_{error} = \frac{SS_{error}}{df_{error}}} \longleftarrow n-k$$

If the null hypothesis is true, do we expect the *F* value to be large?

ANOVA

Comparing population means across *k* groups

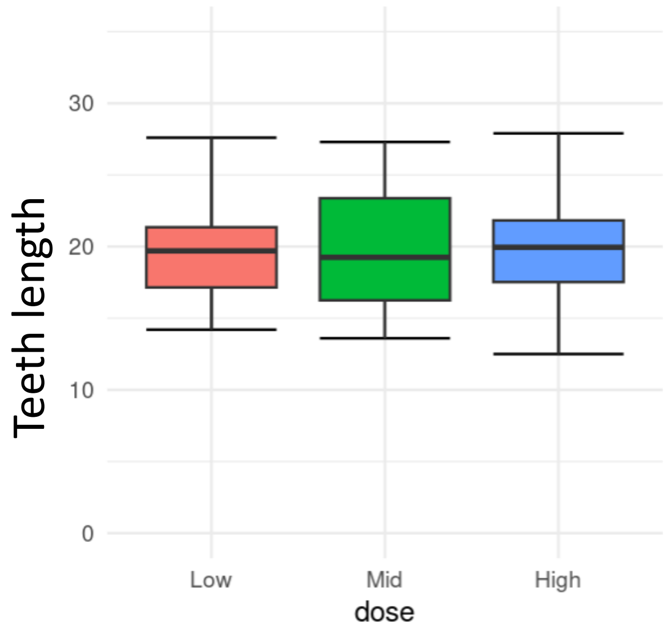
1. State your hypotheses

2. Calculate the test statistic ***F*** (based on sample data)

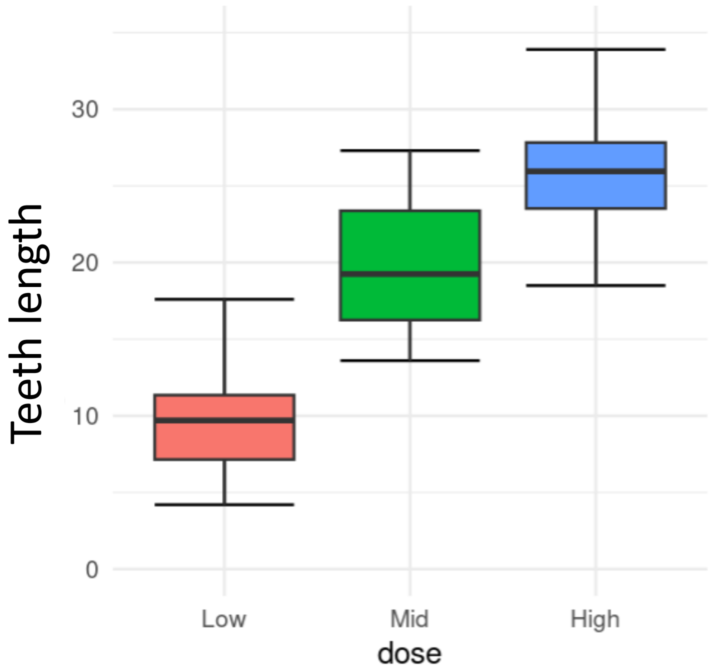
$$F = \frac{\dfrac{SS_{group}}{df_{group}}}{\dfrac{SS_{error}}{df_{error}}}$$

ANOVA table	***SS***	***df***	***MS***	***F***	***p*-value**
Group					
Error					
Total					

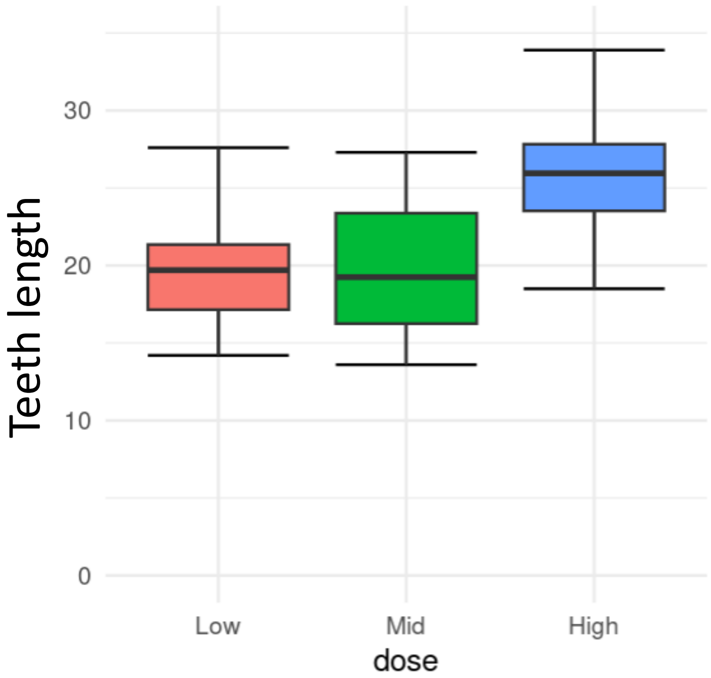
ANOVA



Graph A



Graph B



Graph C

Match each plot with corresponding *F*-stat

F = 26.2

F = 0.5

F = 67.4

ANOVA

Comparing population means across *k* groups

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{F}$ (based on sample data)

3. Compare test statistic to null distribution (calculate ***p*-value**)

4. Make a conclusion in context, reporting the appropriate statistics (F, df, p-value).

$$\boldsymbol{df_{group} = k - 1,\ df_{error} = n - k}$$

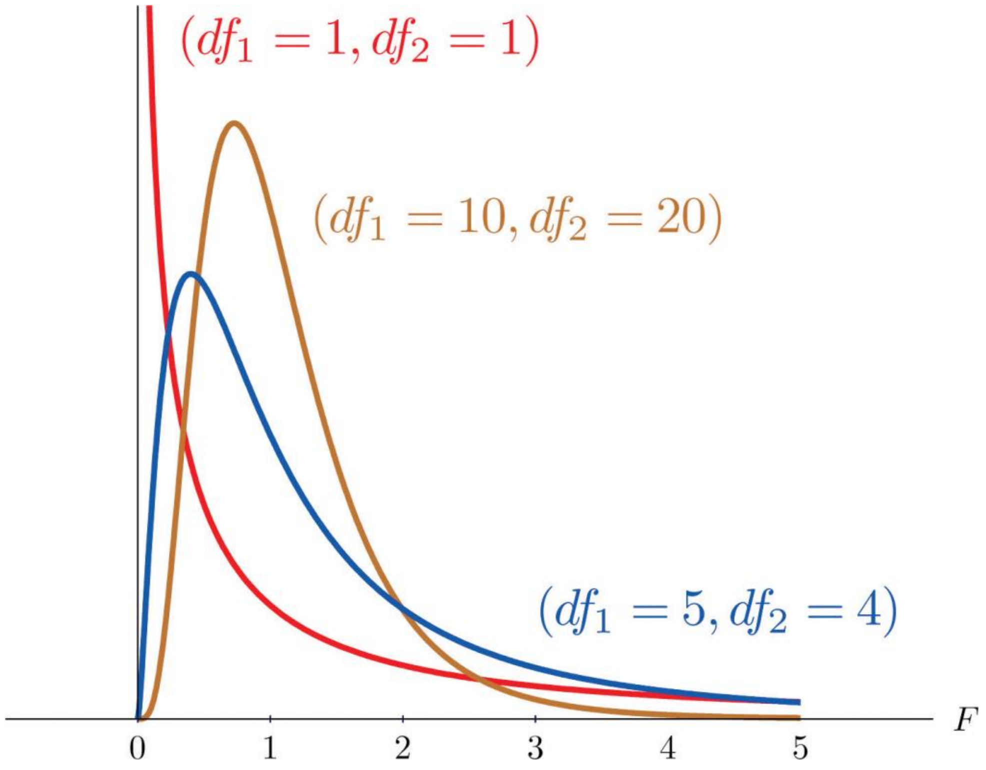
F distribution

A probability distribution that:

- is always positive
- is skewed to the right
- depends on **two** degrees of freedom

$$df_{group} = k - 1$$

$$df_{error} = n - k$$



ANOVA

Comparing population means across *k* groups

When reporting results of a significant test, also report a measure of the effect size with a **proportion of the variation explained by the differences between groups**.

$$\boldsymbol{R^2 = \frac{SS_{group}}{SS_{total}}}$$

ANOVA

Comparing population means across *k* groups

Check assumptions:

✓ Random sample

✓ Independent observations

✓ The distribution in each group is normally distributed

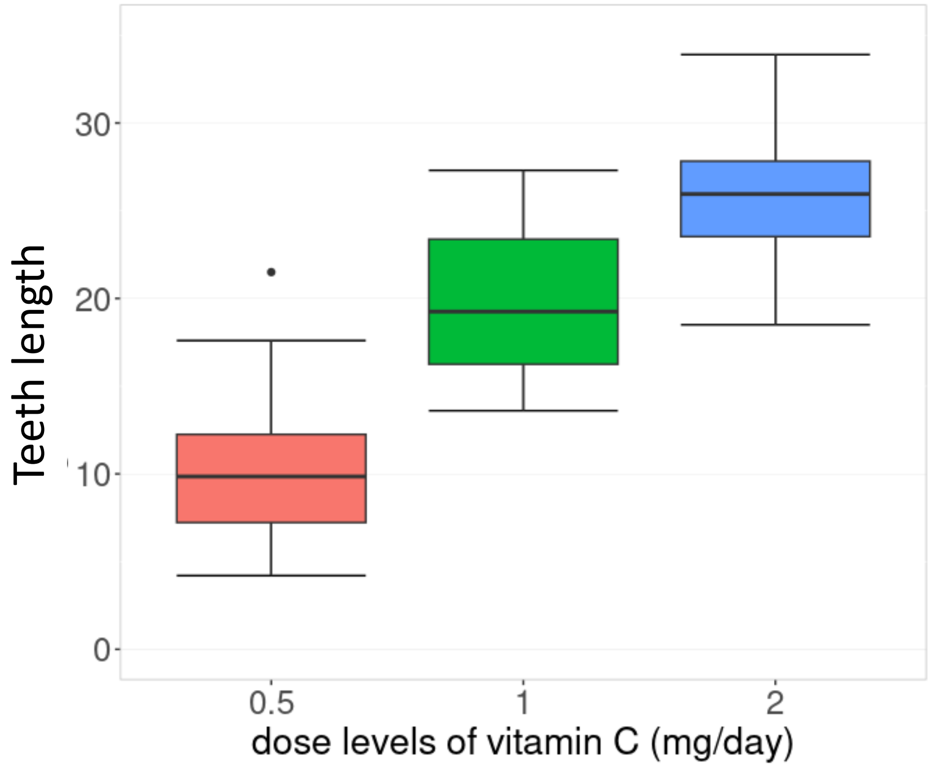
✓ All groups have equal variance

ANOVA

Post-hoc Analysis

We can conduct multiple independent t-tests to find any pairwise differences.

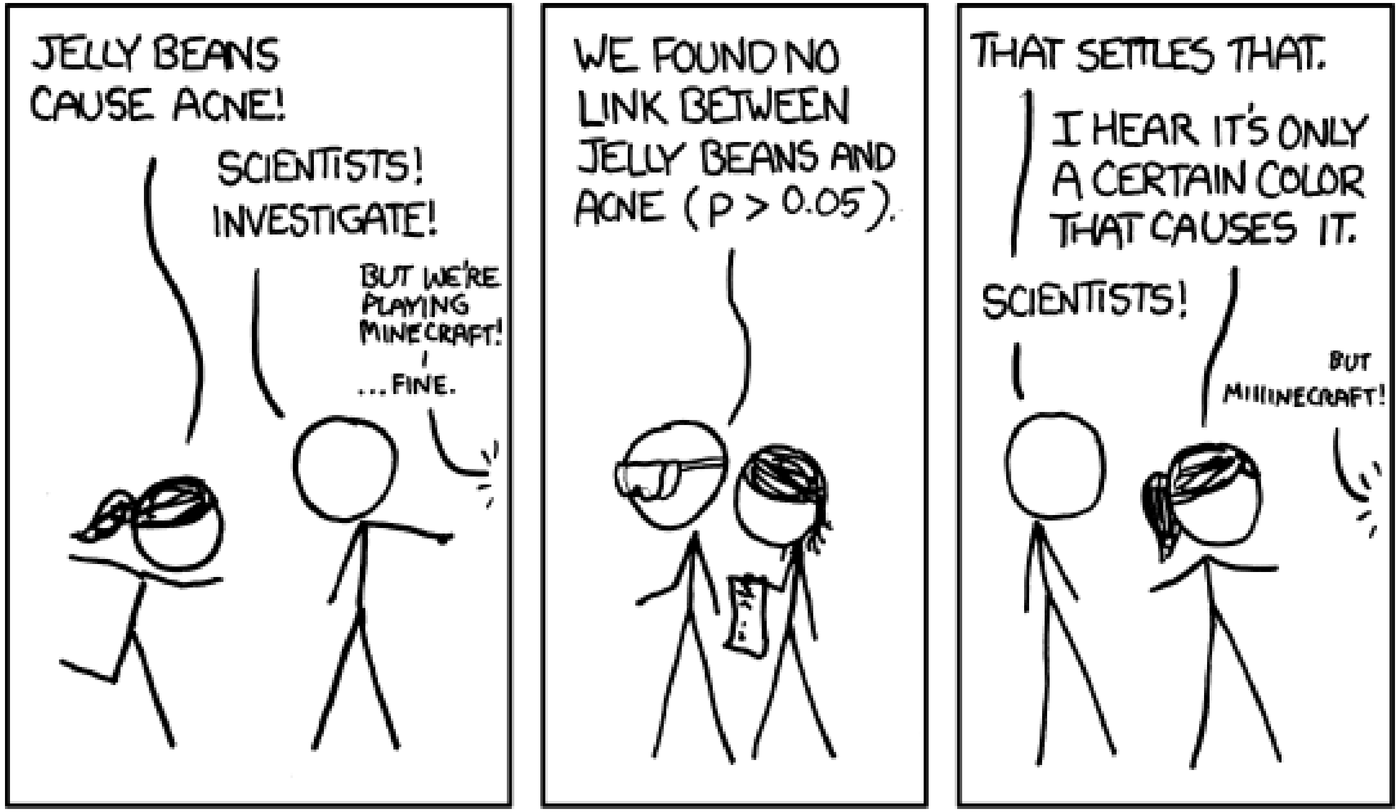
How many independent t-tests should we conduct?



Ethics Check

If 100 studies repeatedly investigated a phenomenon for which the **null hypothesis is actually true**, about how many of them we can expect to still reject the null hypothesis?

Ethics Check



Jelly Beans http://xkcd.com/882/

What color jelly bean causes (or prevents!) acne?

Purple	NS :-(	NS :-(	Brown
Pink	NS :-(	NS :-(	Blue
Teal	Significant!!! PUBLISH!!!	NS :-(	Salmon
Red	NS :-(	NS :-(	Turquoise
Magenta	NS :-(	NS :-(	Yellow
Grey	NS :-(	NS :-(	Tan
Cyan	NS :-(	NS :-(	Green
Mauve	NS :-(	NS :-(	Beige
Lilac	NS :-(	NS :-(	Black
Peach	NS :-(	NS :-(	Orange

Test! / Clear

The chance that nothing is significant is only 0.3585,
so don't give up hope!

http://www.jerrydallal.com/LHSP/jellybean.htm

ANOVA

Post-hoc Analysis

Applying adjustments for conducting multiple comparison tests:

Bonferroni

compare p-values to adjusted α

$$\boldsymbol{\alpha' = \frac{\alpha}{\#\ of\ tests}}$$

Tukey

adjust calculations of p-value
and still compare to α

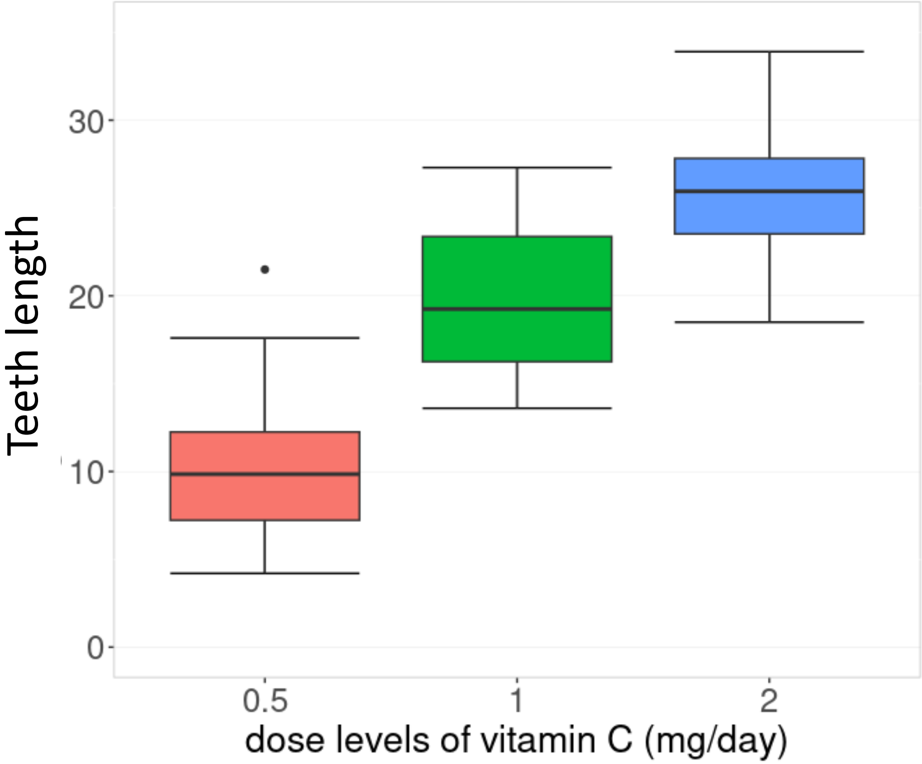
$$\boldsymbol{\alpha = 0.05}$$

ANOVA

Does the length of odontoblasts depend on the dose levels of vitamin C (0.5, 1, and 2 mg/day) received?

Bonferroni adjustment

Dose	Results of independent t-test
0.5 vs 1 mg/day	`t = -6.4766, df = 38, p-value = 0.0000001266`
1 vs 2 mg/day	`t = -4.9005, df = 38, p-value = 0.00001811`
0.5 vs 2 mg/day	`t = -11.799, df = 38, p-value = 0.00000000000002838`



Which groups are different based on the Bonferroni adjustment?

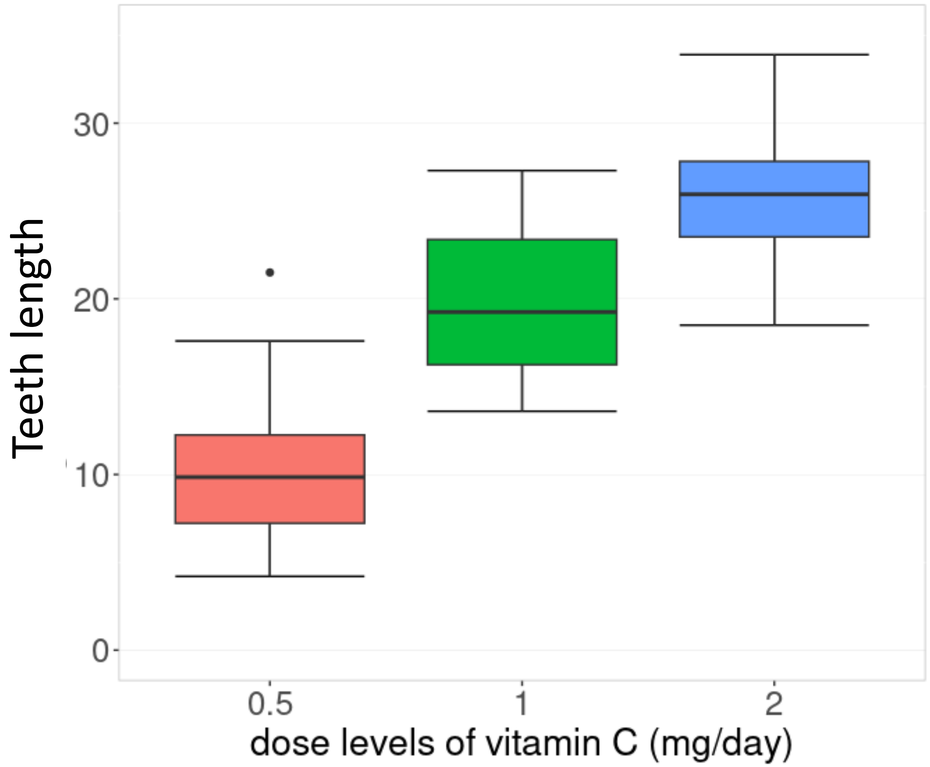
ANOVA

Does the length of odontoblasts depend on the dose levels of vitamin C (0.5, 1, and 2 mg/day) received?

Tukey adjustment

```
$contrasts
 contrast          estimate   SE df t.ratio p.value
 dose0.5 - dose1     -9.13 1.34 57   -6.806  <.0001
 dose0.5 - dose2    -15.49 1.34 57  -11.551  <.0001
 dose1 - dose2       -6.37 1.34 57   -4.745  <.0001

P value adjustment: tukey method for comparing a family of 3 estimates
```



Which groups are different based on the Tukey adjustment?

USING R AND RSTUDIO

χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

1. State your hypotheses

H_0: The distribution of the categories **is** [*specify distribution of each category*]

H_A: The distribution of the categories **is not** [*specify distribution of each category*]

χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{\chi^2}$ (based on sample data)

$$\chi^2 = \sum \frac{(\boldsymbol{Observed - Expected\ Count})^2}{\boldsymbol{Expected}}$$

What does a large value of χ^2 indicate about the null hypothesis?

with $\boldsymbol{Expected\ Count} = (\boldsymbol{Expected\ percentage}) \cdot \boldsymbol{n}$

χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

1. State your hypotheses
2. Calculate the test statistic $\boldsymbol{\chi^2}$ (based on sample data)
3. Compare test statistic to null distribution (calculate ***p*-value**)
4. Make a conclusion in context, reporting the appropriate statistics (χ^2, *df*, *p*-value).

$\boldsymbol{df = number\ of\ categories - 1}$

χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

Check assumptions:

✓ Random sample

✓ Independent observations

✓ Must have sufficient sample size for:

- All expected counts to be greater than 1
- At least 80% of expected counts are ≥ 5

χ^2 Test of Independence

Comparing population counts for different groups

1. State your hypotheses

H_0: The two variables **are** independent

H_A: The two variables **are not** independent

χ^2 Test of Independence

Comparing population counts for different groups

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{\chi^2}$ (based on sample data)

$$\boldsymbol{Expected\ Count_{ij} = \frac{(row\ total)_i \cdot (column\ total)_j}{grand\ total}}$$

$$\boldsymbol{\chi^2 = \sum \frac{\left(Observed_{ij} - Expected_{ij}\right)^2}{Expected_{ij}}}$$

χ^2 Test of Independence

Comparing population counts for different groups

1. State your hypotheses
2. Calculate the test statistic $\boldsymbol{\chi^2}$ (based on sample data)
3. Compare test statistic to null distribution (calculate ***p*-value**)
4. Make a conclusion in context, reporting the appropriate statistics (χ^2, *df*, *p*-value).

$$\boldsymbol{df = (number\ of\ categories - 1)(number\ of\ categories - 1)}$$

χ^2 Test of Independence

Comparing population counts for different groups

Check assumptions:

✓ Random sample

✓ Independent observations

✓ Must have sufficient sample size for:

- All expected counts to be greater than 1
- At least 80% of expected counts are ≥ 5

Summary:

Name of test	Variables involved	Hypotheses	Test statistic	df	Assumptions*	Effect size
One-sample t-test	numeric response	$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$t = \frac{\bar{X} - \mu_0}{SE}$	$n - 1$	✓ normal	$\bar{X} \pm t^* \cdot SE$
Independent t-test	numeric response binary predictor	$H_0: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$	$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$	$n_1 + n_2 - 1$	✓ normal ✓ equal variance	$\bar{X}_1 - \bar{X}_2 \pm t^* \cdot SE$
ANOVA	numeric response categorical predictor	$H_0: \mu_1 = \mu_2 = \cdots$ $H_A: not\ all\ equal$	$F = \frac{MS_{group}}{MS_{error}}$	$df_{group} = k - 1$ $df_{error} = n - k$	✓ normal ✓ equal variance	Post Hoc Model fit R^2
Chi2 Goodness-of-Fit	categorical response	$H_0: distrib\ is\ \ldots$ $H_A: distrib\ is\ not\ \ldots$	$\chi^2 = \sum \frac{(obs - exp)^2}{exp}$	$\#\ cat - 1$	✓ sample size	percentages
Chi2 Test of Independence	categorical response categorical predictor	$H_0: independent$ $H_A: not\ independent$	$\chi^2 = \sum \frac{(obs - exp)^2}{exp}$	$(\#\ cat_1 - 1)$ $\cdot (\#\ cat_2 - 1)$	✓ sample size	percentages

***Random sample and Independent observations are common assumptions to all these tests**

USING R AND RSTUDIO

Failing Assumptions

Some assumptions we have discussed :

Comment on limitations or consider alternatives

Based on the **study design**:

- ✓ Random sample
- ✓ Independent observations

- Is the sample still representative of the population? Address any potential bias.
- Were observations collected independently?

Based on **statistics**:

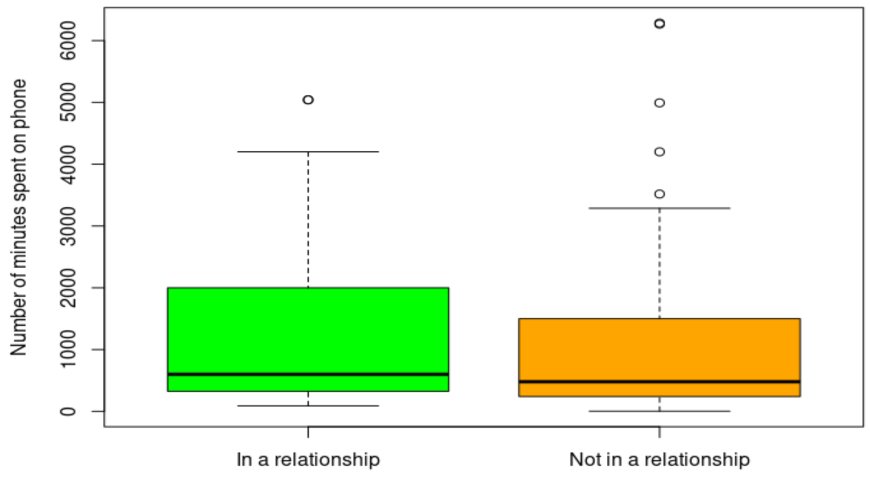
- ✓ Normality
- ✓ Equal variance

- Transform the response variable
- Perform Welch's t-test for unequal variance

Failing Assumptions: Equal Variance

Testing for equal variance: Levene's test

H_0: the variances of the two populations **are equal**
H_A: the variances of the two populations **are not equal**




```
> leveneTest(phone~romance, data)
Levene's Test for Homogeneity of Variance
       Df F value Pr(>F)
group   1  0.7239 0.3959
      199
```

Failing Assumptions: Equal Variance

What to do if the equal variance assumption is <u>not met</u>?

Conduct a **Welch's test** (similar to independent t-tests)

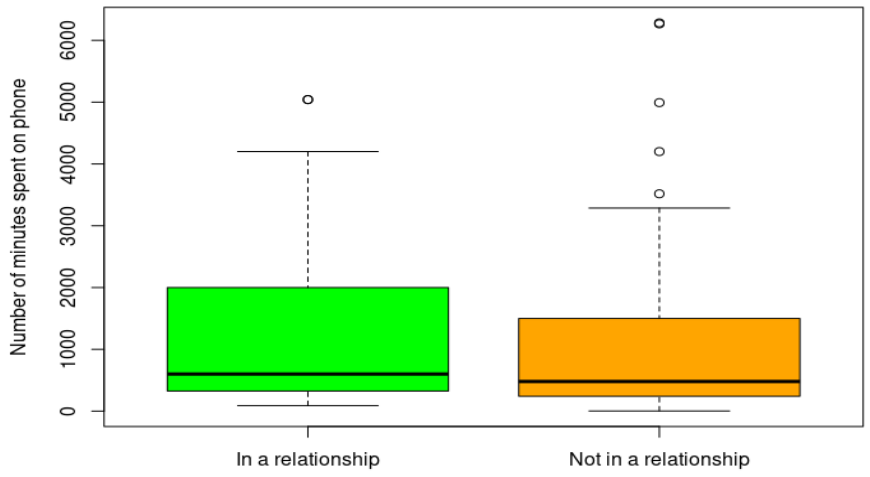
$$t = \frac{\bar{X}_1 - \bar{X}_2 - 0}{SE_{\bar{X}_1 - \bar{X}_2}}$$

Only the calculation for $SE_{\bar{X}_1 - \bar{X}_2}$ will change (but always provided)

Failing Assumptions: Normality

Testing normality: Shapiro-Wilk Test

H_0: the sample values **come** from a normal distribution
H_A: the sample values **do not come** from a normal distribution




```
> shapiro.test(data[data$romance == TRUE,]$phone)

        Shapiro-Wilk normality test

data:  data[data$romance == TRUE, ]$phone
W = 0.81362, p-value = 0.0000003

> shapiro.test(data[data$romance == FALSE,]$phone)

        Shapiro-Wilk normality test

data:  data[data$romance == FALSE, ]$phone
W = 0.77155, p-value = 0.000000000001519
```

The assumption of normality is not met for either group.

Failing Assumptions: Normality

What to do if the normality assumption is <u>not met</u>?

We can try to transform our response variable

$$\mathbf{X}' = \boldsymbol{f}(\mathbf{X})$$

- Apply a (one-to-one) **function** to every single observation to transform data.

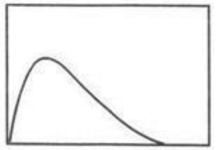
- The **interpretation** of the variable being analyzed must change according to the transformation used.

- It's **valid** to try multiple transformations to find one that makes your data normal (<u>BUT NOT</u> to run multiple H_0 tests to find the lowest p-value).

Failing Assumptions: Normality

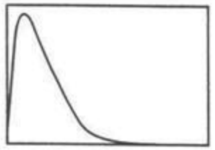
What to do if the normality assumption is <u>not met</u>?

We can try to transform our response variable



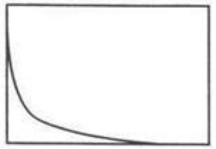
Square root

$$X' = \sqrt{X}$$



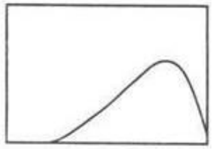
Logarithm

$$X' = \ln(X)$$



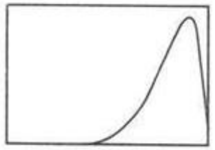
Inverse

$$X' = 1/X$$



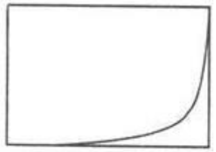
Reflect and square root

$$X' = \sqrt{\max(X) + 1 - X}$$



Reflect and logarithm

$$X' = \ln\left(\frac{X}{1 - X}\right)$$



Reflect and inverse

$$X' = \frac{1}{\max(X) + 1 - X}$$

https://www.statisticssolutions.com/transforming-data-for-normality/

USING R AND RSTUDIO

Nonparametric tests

Resampling methods

- Randomization/Permutation test

  Shuffling the relationship between variables in our sample to generate a null distribution against which to compare an observed test statistic.

- Bootstrapping

  Taking samples with replacement to generate an empirical sampling distribution of an estimate for precision (e.g., standard error, 95% CI).

No assumed distribution, no assumptions besides random sample and independent observations

✓ Most used when sample size is small or some assumptions were violated
✓ If assumptions are met, a parametric test will have more power

Nonparametric tests

Randomization test

1. Calculate the observed statistic
2. Randomly mix up the association by permuting one variable
3. Recalculate the test statistic on mixed-up data
4. Repeat the two last steps many times to generate a sampling distribution under the null hypothesis of no association
5. Compare the observed tests statistic to the sampling distribution

Nonparametric tests

Bootstrapping

1. Calculate the observed statistic
2. Randomly resample from the data with replacement
3. Recalculate the statistic on resampled data
4. Repeat the two last steps many times to generate a sampling distribution
5. Calculate the standard error (standard deviation of the sampling distribution) and construct confidence intervals

Nonparametric tests

Mann-Whitney U Test: comparing the distribution of a numeric variable across two independent groups

Most used when sample size is small or normality was violated

1. State your hypotheses

2. Calculate the rank sums of each group and the test statistic ***U***

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 - U_1 \qquad \boldsymbol{U} = \mathbf{Max}(\boldsymbol{U_1}, \boldsymbol{U_2})$$

Nonparametric tests

Mann-Whitney U Test: comparing the distribution of a numeric variable across two independent groups

Most used when sample size is small or normality was violated

1. State your hypotheses
2. Calculate the rank sums of each group and the test statistic $\boldsymbol{U}$
3. Compare test statistic to null distribution (calculate ***p*-value**)
4. Make a conclusion in context, reporting the appropriate statistics (U, n_1, n_2, p-value).

Next

Day 3

Linear Regression

- Simple Linear Regression
- Multiple Regression with different types of predictors
- Model assumptions, evaluation, and comparisons

Any questions? comments?