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

laylaguyot

Overview | Repositories 8 | Projects

Layla Guyot
laylaguyot

Statistics and Data Science enthusiast: teacher and researcher in education, focusing on bridging the gap between academia and industry.

https://github.com/laylaguyot/CBRS_Intro_Statistical_Modeling

README

Introduction to Statistical Modeling

Materials

Download the slides and corresponding .Rmd files each day

Access RStudio

Go to https://gsafcomp02.ccbb.utexas.edu and select RStudio

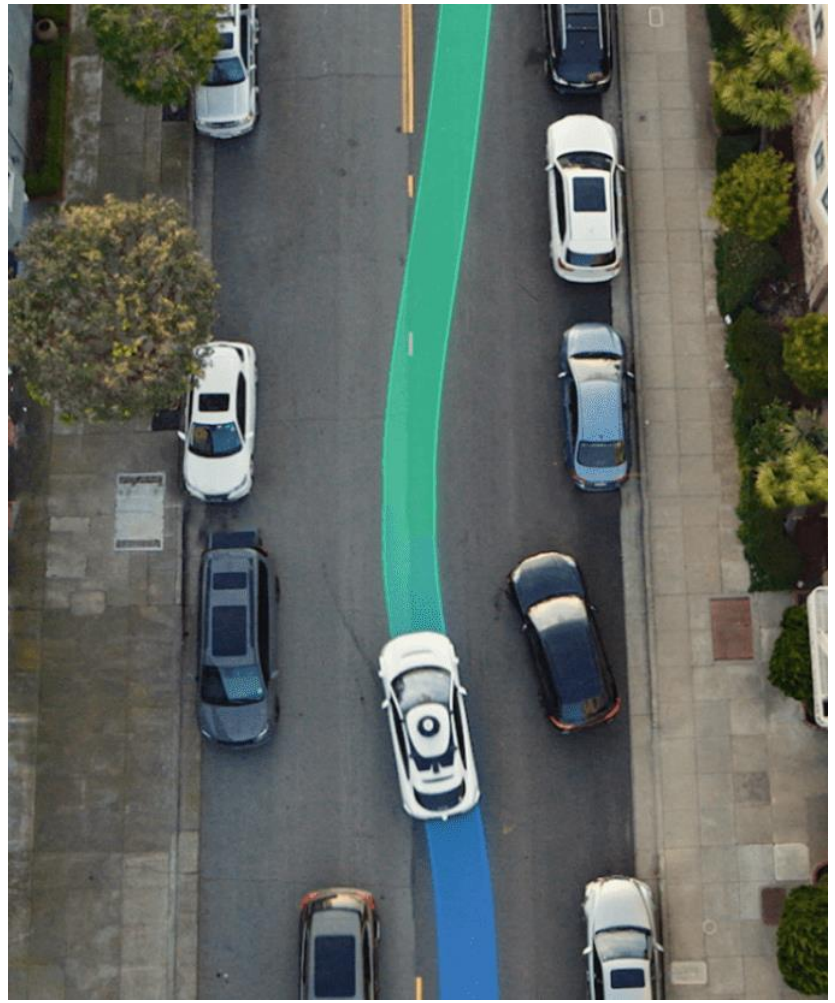
Statistical Modeling

What is a statistical model?

Consider the challenge of building a driverless car like Waymo. Statistical models are used to:

- simplify complex inputs like camera feeds, sensor data, and maps.
- help predict outcomes such as pedestrian movement, road conditions, and vehicle behavior.

These models don't capture every detail but extract the essential features and patterns needed to drive safely.



Statistical Modeling

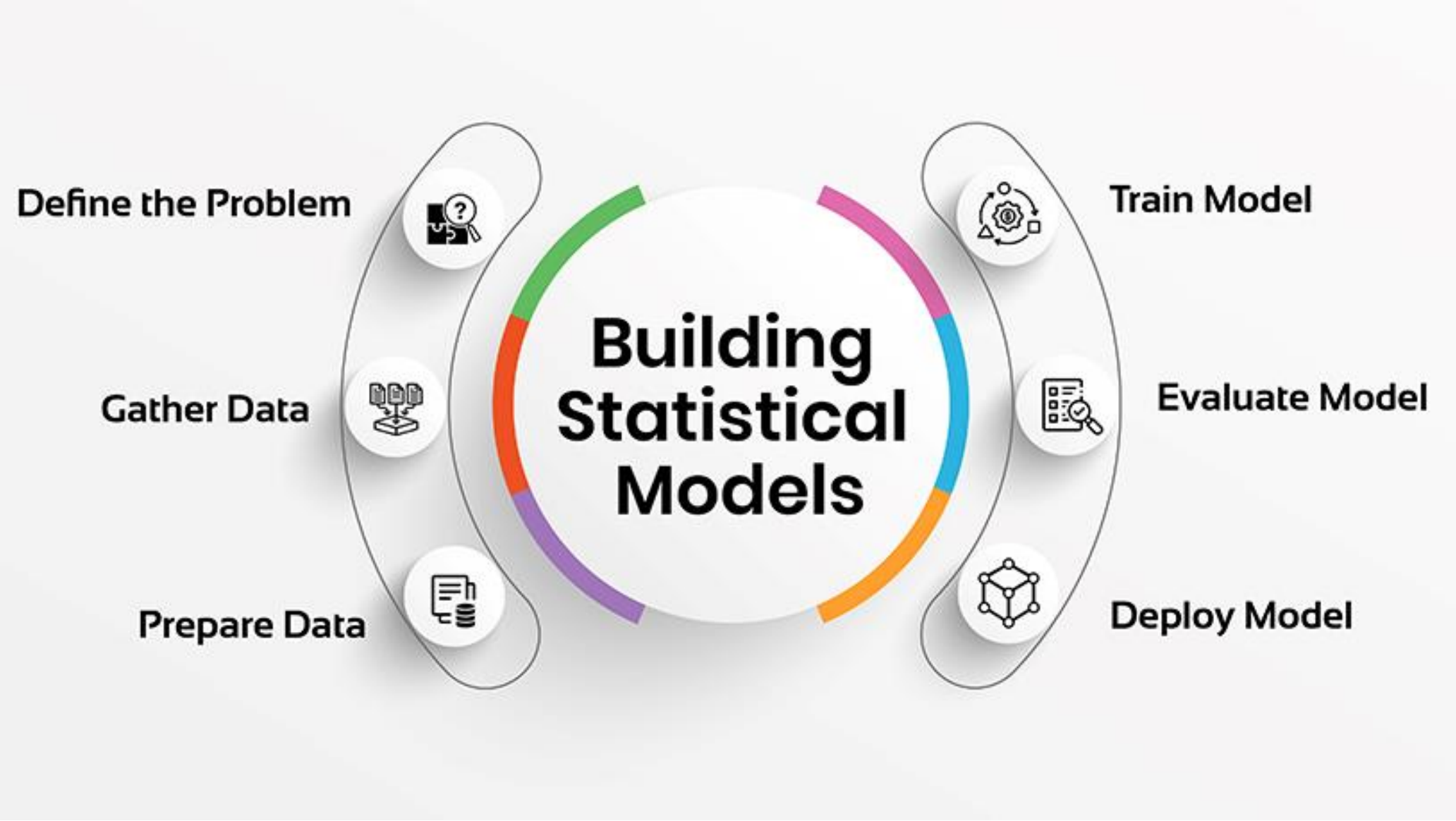
What is a statistical model?

A statistical model simplifies the real world: the real world is too complex to model perfectly. It focuses on capturing the essential relationships between variables, capturing the signal among the noise.

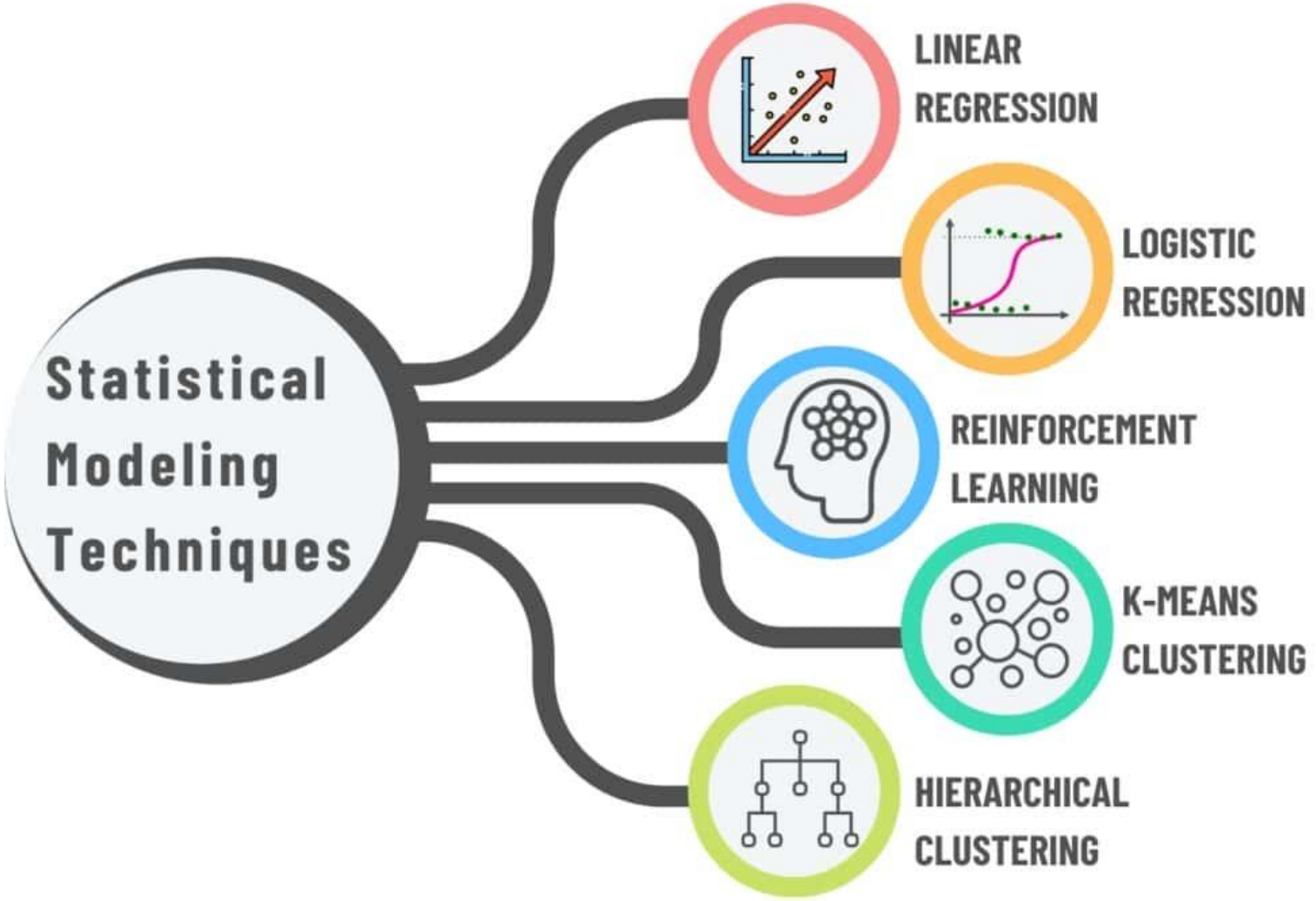
"All models are wrong, but some are useful."
George Box, 1976

Many biological and medical outcomes depend on several factors.
Example: A patient's blood pressure may depend on age, weight, stress, genetics, medication…

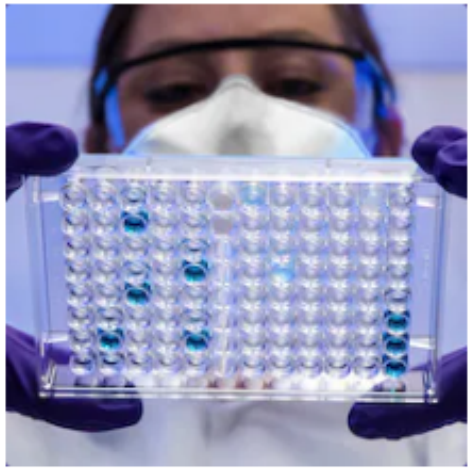
Statistical Modeling



Statistical Modeling



Statistical Modeling in the wild



September 18, 2023

Tests that diagnose diseases are less reliable than you'd expect. Here's why

Adrian Barnett, *Queensland University of Technology* and Nicole White, *Queensland University of Technology*

Many diagnostic tests are far from 100% accurate – and even in the era of big data and machine learning, they never will be.

October 13, 2023

South Africa's 2022 census missed 31% of people - big data could help in future

David Everatt, *University of the Witwatersrand*

Big data is not the answer to all the challenges that faced Census 2022, but it may be a key enabler for gathering reliable national data in the future.

Statistical Modeling in the wild

PeerJ

The implementation of rare events logistic regression to predict the distribution of mesophotic hard corals across the main Hawaiian Islands

Research Article | Ecology | Ecosystem Science | Environmental Sciences | Marine Biology

Lindsay M. Veazey[1], Erik C. Franklin[2], Christopher Kelley[3], John Rooney[†4], L. Neil Frazer[5], Robert J. Toonen[6]

Published July 6, 2016

PALEONTOLOGY AND EVOLUTIONARY SCIENCE

The oldest record of the Steller sea lion *Eumetopias jubatus* (Schreber, 1776) from the early Pleistocene of the North Pacific

Research Article | Evolutionary Studies | Marine Biology | Paleontology | Taxonomy | Zoology

Nahoko Tsuzuku[1], Naoki Kohno[1,2]

Published August 27, 2020

BRAIN, COGNITION AND MENTAL HEALTH

Clustering of fMRI data: the elusive optimal number of clusters

Research Article | Bioinformatics | Computational Biology | Neuroscience

Mohamed L. Seghier

Published October 3, 2018

ENVIRONMENTAL SCIENCE

Reference evapotranspiration estimate with missing climatic data and multiple linear regression models

Research Article | Agricultural Science | Environmental Sciences | Plant Science | Natural Resource Management

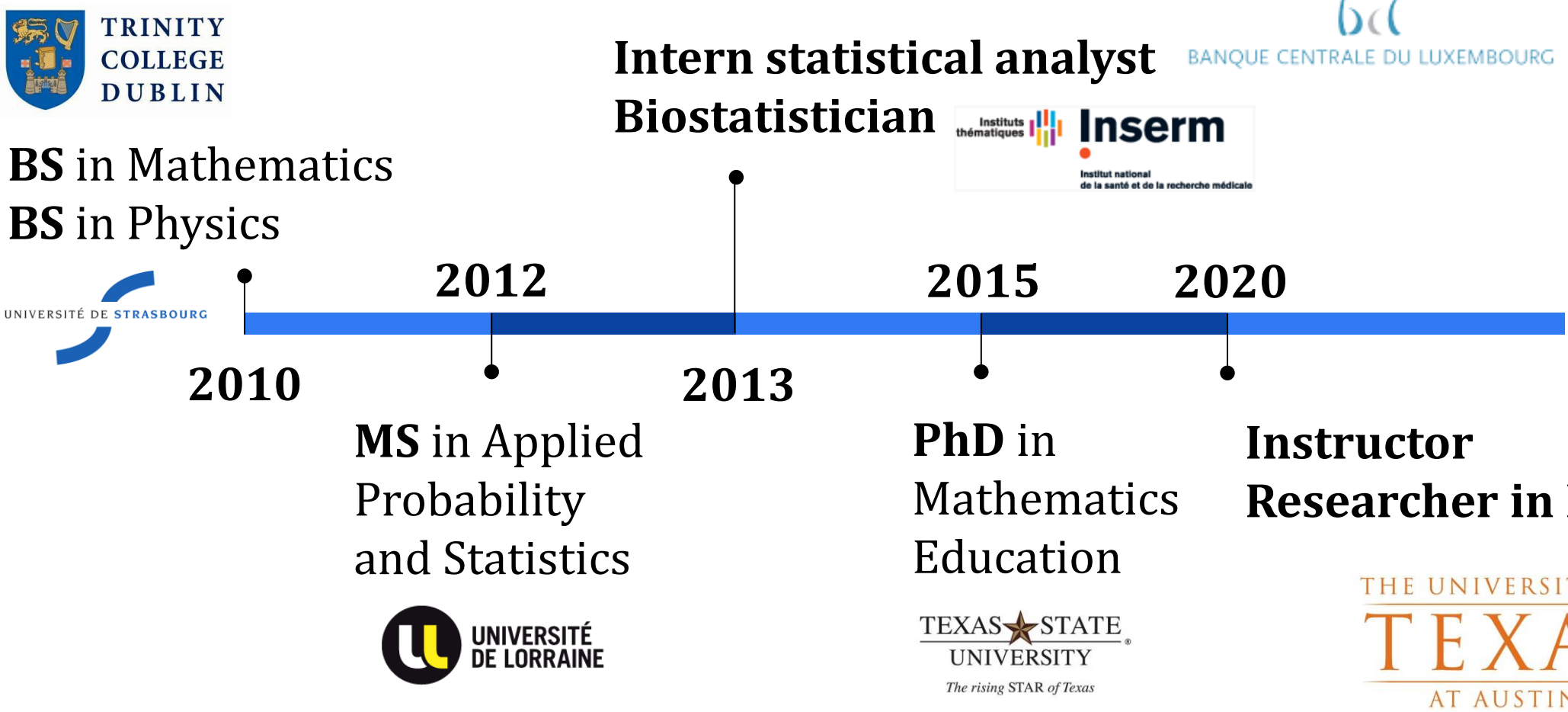
Deniz Levent Koç, Müge Erkan Can

Published April 27, 2023

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

About me


TRINITY COLLEGE DUBLIN
BS in Mathematics
BS in Physics
UNIVERSITÉ DE STRASBOURG
2010
2012
MS in Applied Probability and Statistics
UNIVERSITÉ DE LORRAINE
2013
Intern statistical analyst
Biostatistician
Instituts thématiques
Inserm
Institut national de la santé et de la recherche médicale
BANQUE CENTRALE DU LUXEMBOURG
2015
PhD in Mathematics Education
TEXAS STATE UNIVERSITY
The rising STAR of Texas
2020
Instructor
Researcher in Education
THE UNIVERSITY OF TEXAS AT AUSTIN

Main contributions

The Efficacy of Research-Based "Mathematics for All" Professional Development

Elementary school teachers' noticing of essential mathematical reasoning forms: justification and generalization

16. AI-Supported Coding Assignments (Statistics & Data Science)

ICOTS11 (2022) Invited Paper - Refereed (DOI: 10.52041/iase.icots11.T5A1) Guyot & White

PROMOTING OPPORTUNITIES TO LEARN FOR STATISTICIANS

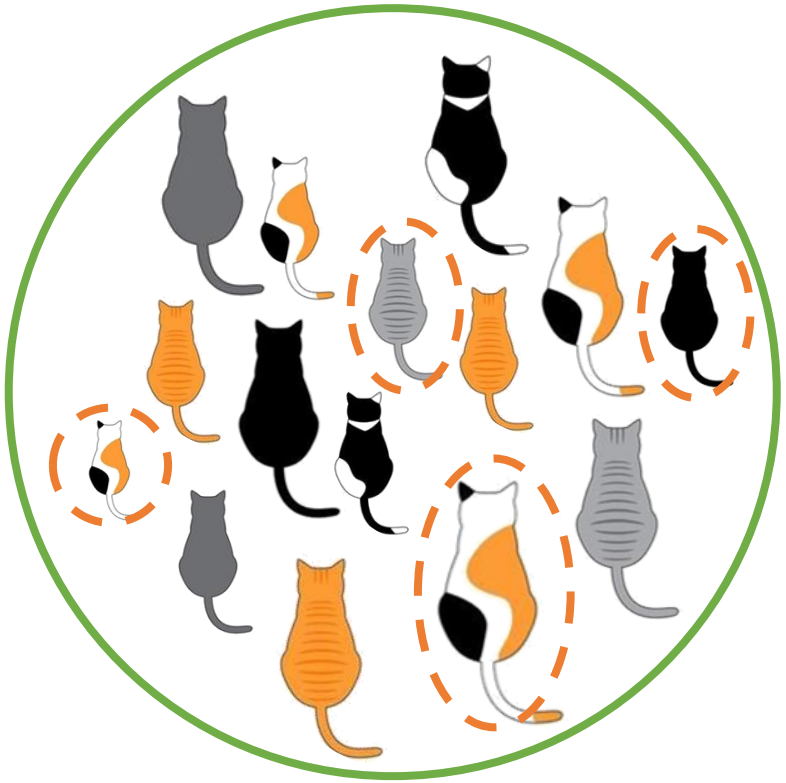
LES FONDS DE FONDS DOMICILIÉS AU LUXEMBOURG : PRÉSENTATION ET ANALYSE STATISTIQUE

About you!

Please complete this quick survey so we can better understand your interest and expectations for this workshop.

https://forms.gle/rVVx8JPcPFpaU4wz5

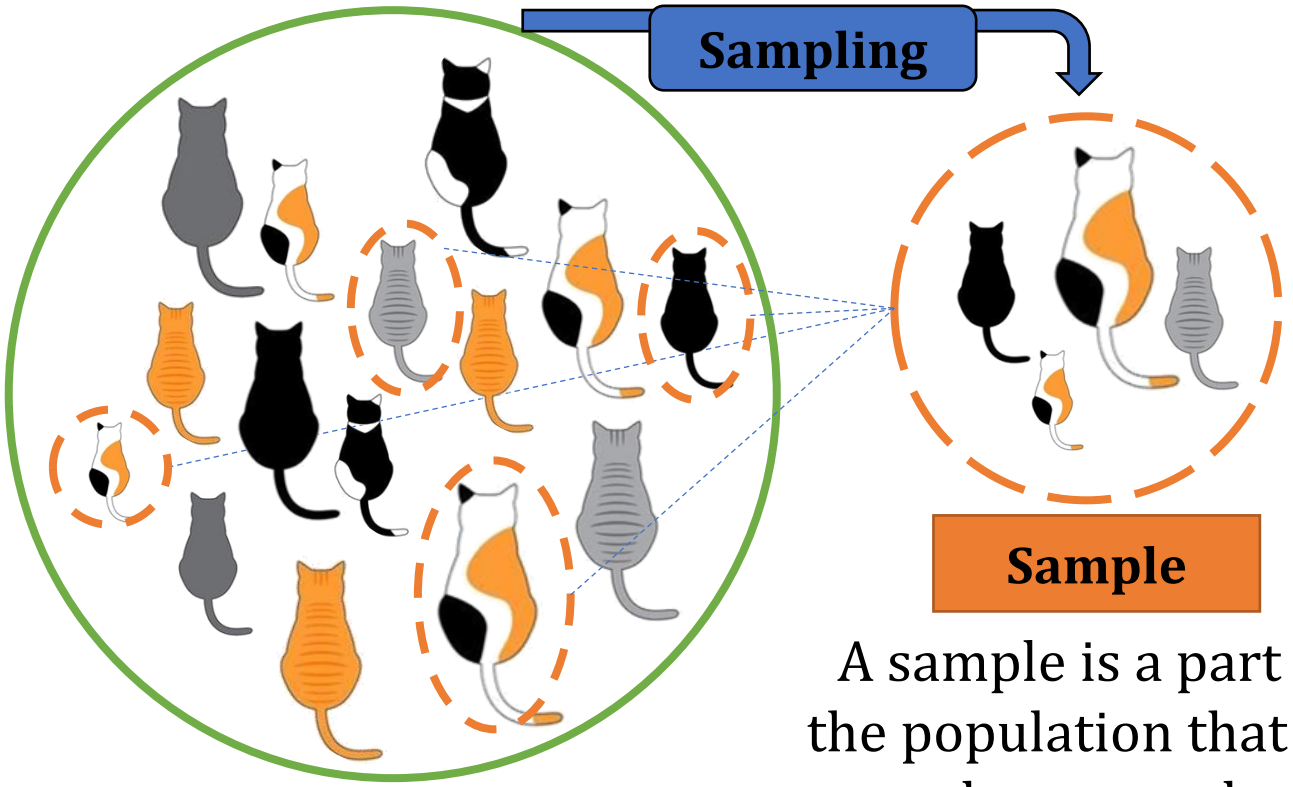
The Big Idea of doing statistics



Population

The population is the entire collection of individuals that we want to learn about.

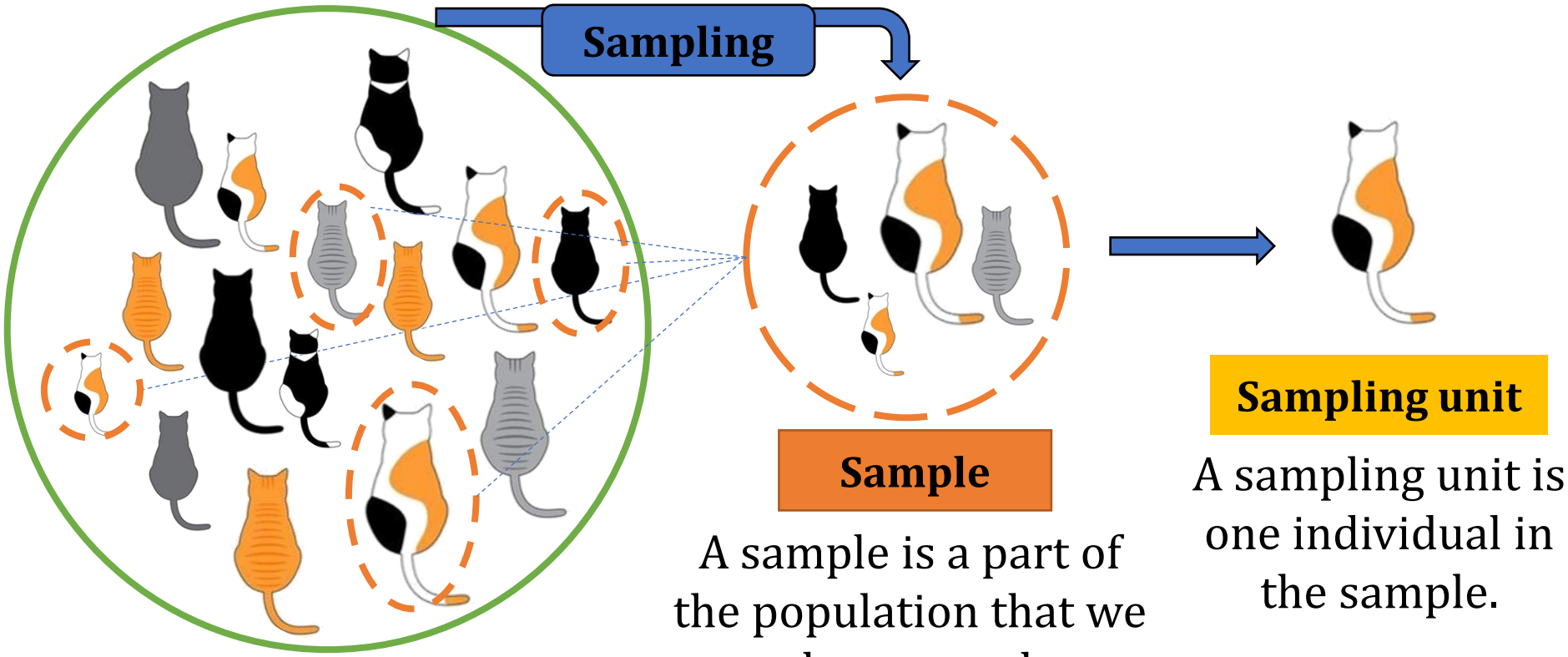
The Big Idea of doing statistics



A sample is a part of the population that we select to study.

The population is the entire collection of individuals that we want to learn about.

The Big Idea of doing statistics



Population

The population is the entire collection of individuals that we want to learn about.

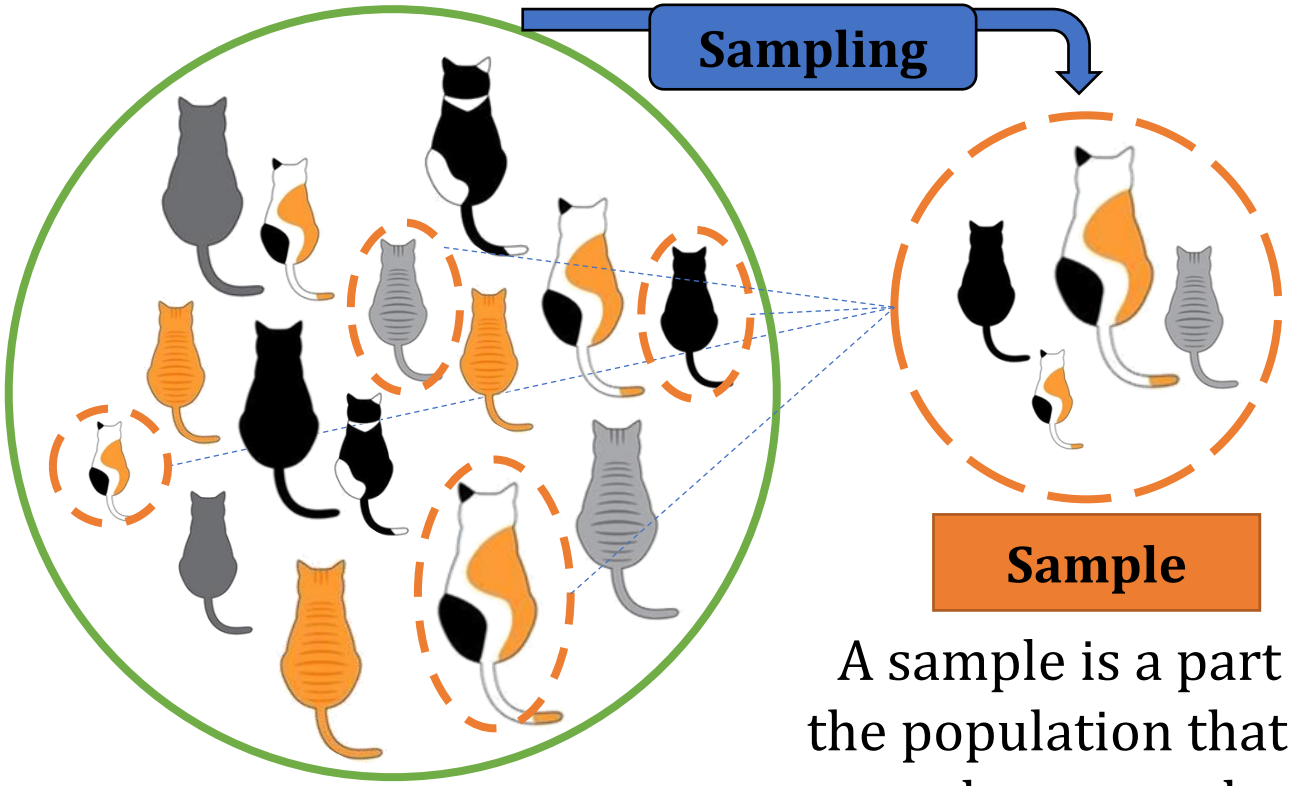
Sample

A sample is a part of the population that we select to study.

Sampling unit

A sampling unit is one individual in the sample.

The Big Idea of doing statistics



Population

The population is the entire collection of individuals that we want to learn about.

Sample

A sample is a part of the population that we select to study.

Sampling unit

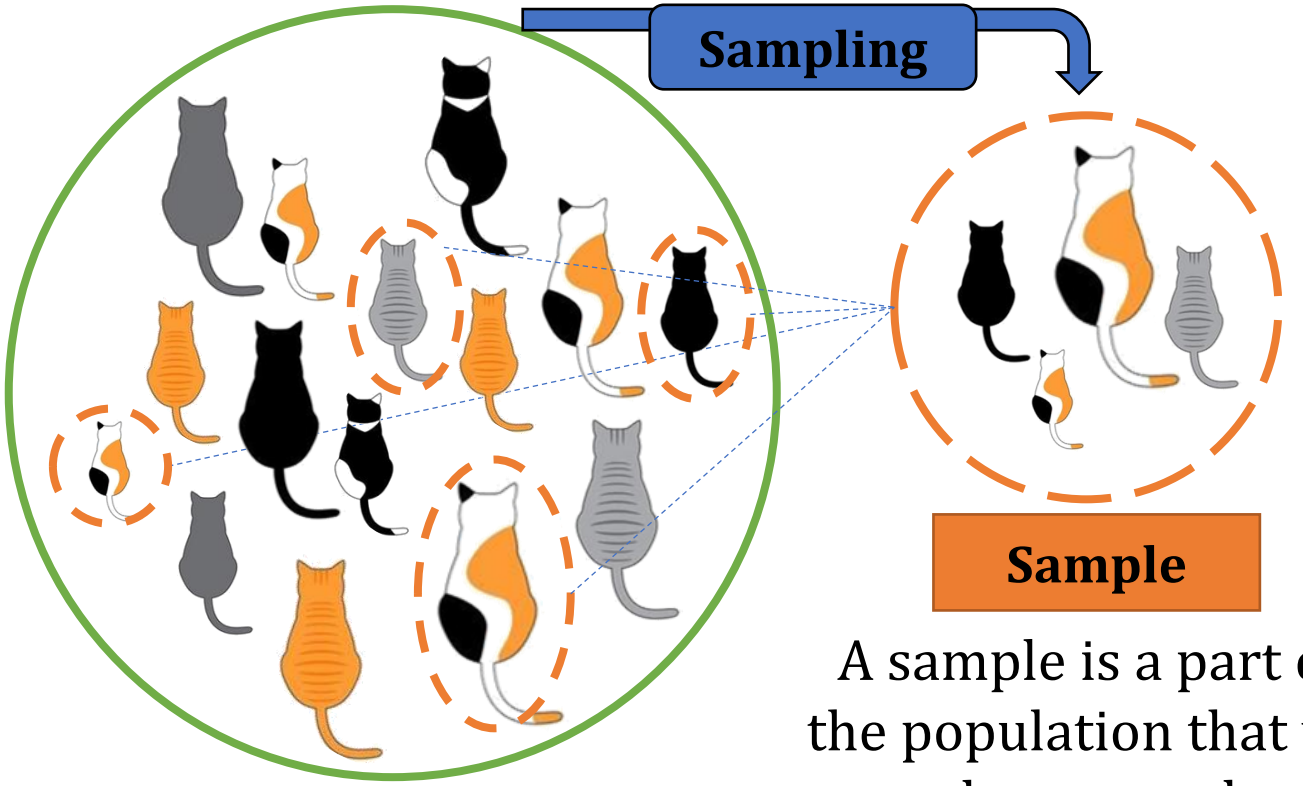
A sampling unit is one individual in the sample.

Weight, Color, …

Variables

A variable describes some characteristic of each sampling unit.

The Big Idea of doing statistics



Population

The population is the entire collection of individuals that we want to learn about.

Sample

A sample is a part of the population that we select to study.

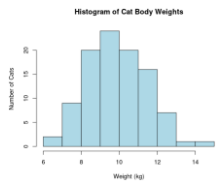
Sampling unit

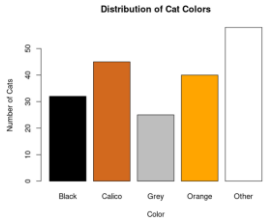
A sampling unit is one individual in the sample.

Weight, Color, …

Variables

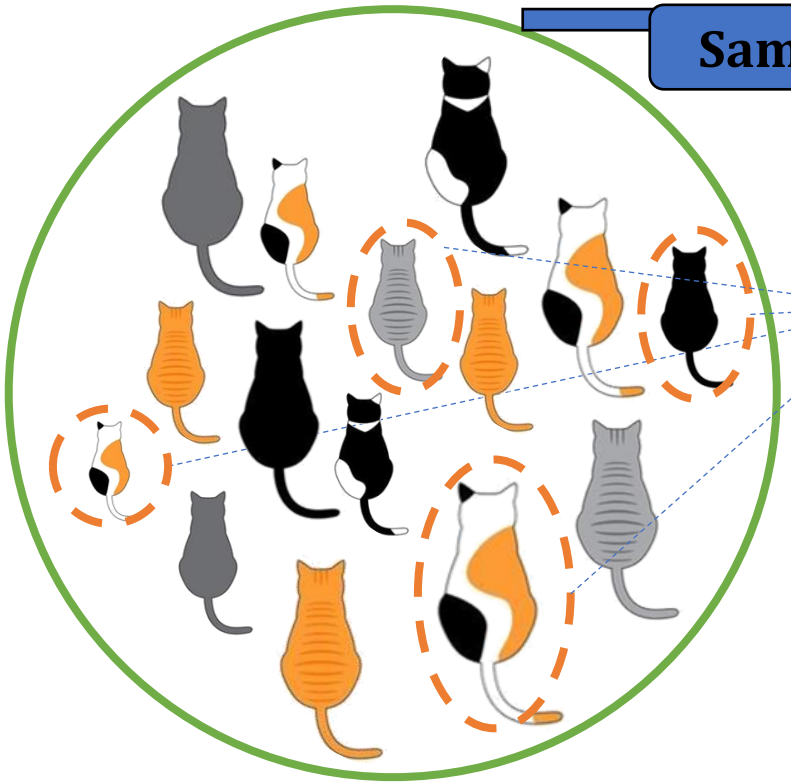
A variable describes some characteristic of each sampling unit.





Calculate **statistics** on sample data:
12 lbs on average, 27% calico

The Big Idea of doing statistics



Population

The population is the entire collection of individuals that we want to learn about.

Sampling

Sample

A sample is a part of the population that we select to study.

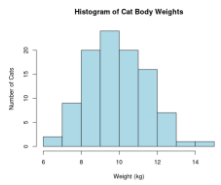
Sampling unit

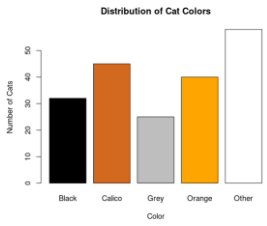
A sampling unit is one individual in the sample.

Weight, Color, ...

Variables

A variable describes some characteristic of each sampling unit.





Calculate **statistics** on sample data: 12 lbs on average, 27% calico

Inference

We want to estimate a **parameter**: mean weight, percentage of calico

The Big Idea of doing statistics

Important notations:

Sample	**P**opulation
What we calculate:	What we estimate:
Statistics	**P**arameters
$\bar{X}$ s	μ σ
r b_1	ρ β_1
$\bar{X} = 12$ lbs on average in sample	$\mu = ?$ mean weight in population

Example

Scenario:

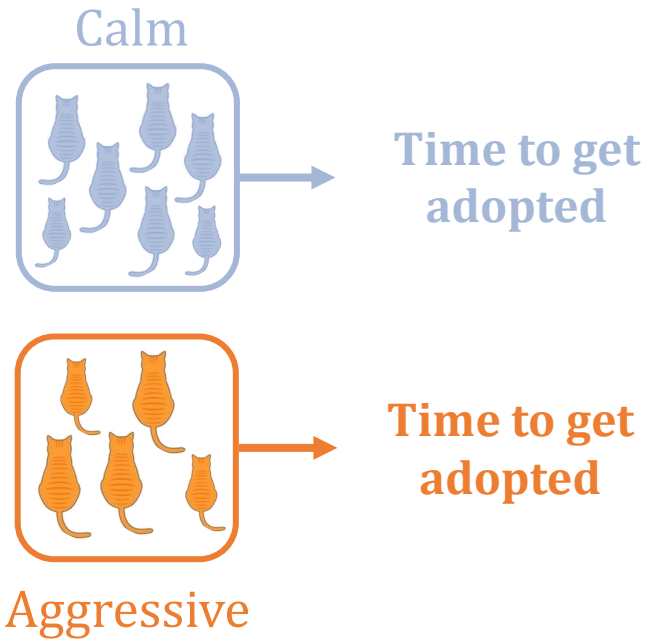
We would like to research how effective the summer workshops offered by CBRS are for the participants. We send a survey at the end of the workshop to all participants.

- population of interest?
- sample?
- variables to collect?
- parameter of interest?

Study design

Observational studies

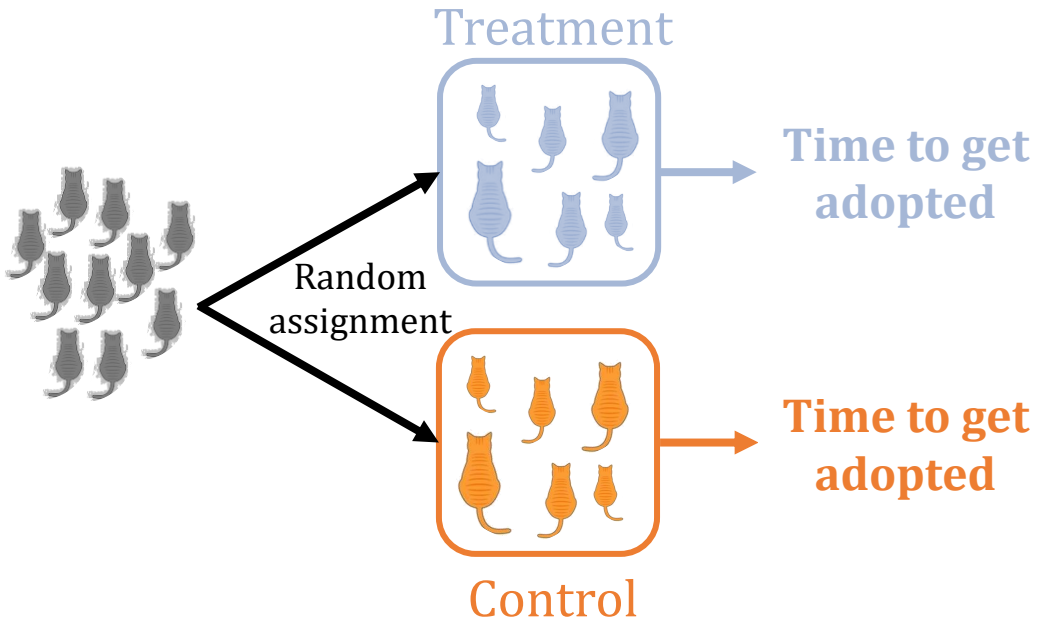
Comparing how long it takes for cats to get adopted depending on their general behavior.



vs

Experiment

Offering some behavioral enrichment program to see if cats get adopted faster.



Why do you think experiments are usually more reliable compared to observational studies?

Sampling

Ideally, we wish to have samples that are:

Random Every unit in the population has an equal chance to be included in the sample. Drawn using a random-number generator.

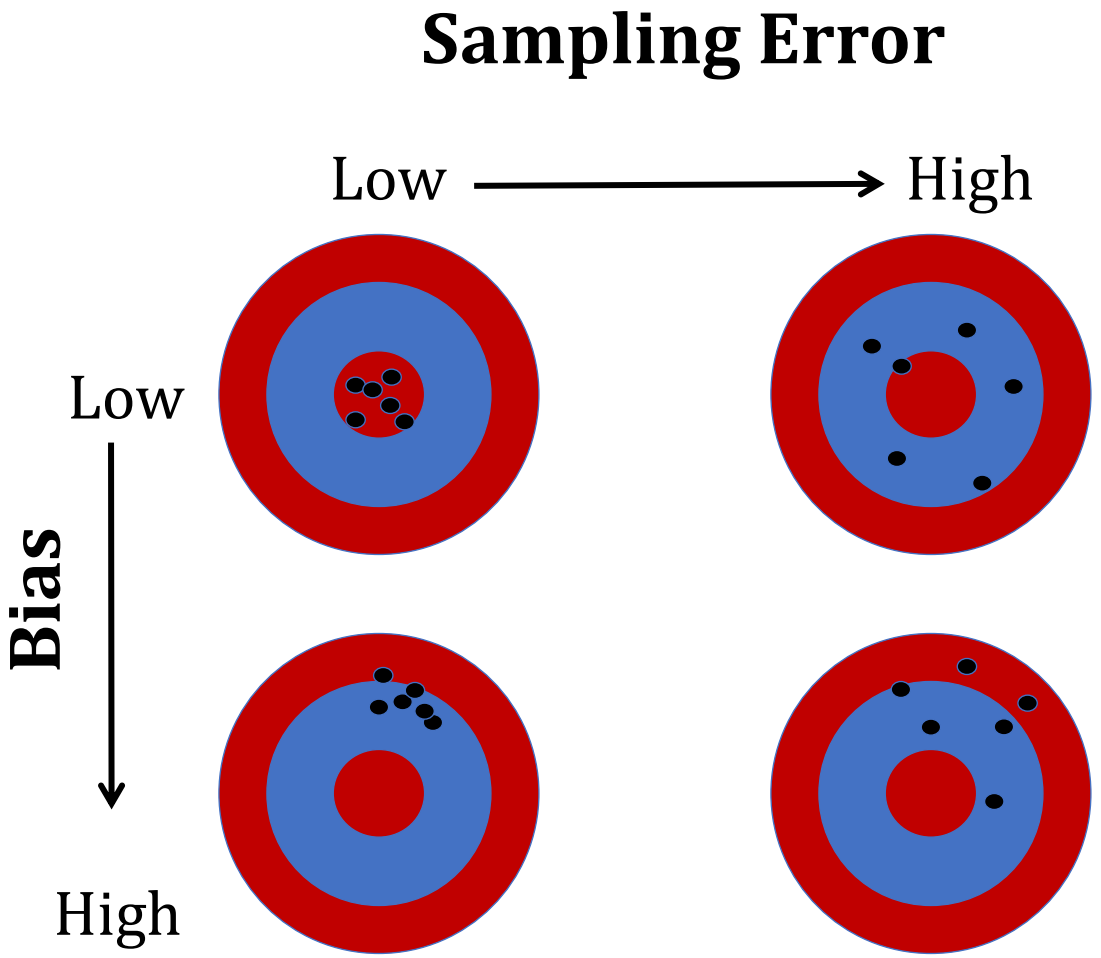
Independent The selection of one sampling unit does not influence the selection of any other sampling unit.

Large Let's set a minimum sample size of 30 sampling units (per group).

Sampling

In our sample, there could be some:

- **Bias:** Systematic tendency for samples to differ from the corresponding population.
- **Sampling Error:** The sample is never a perfect representation of the population and statistics vary from sample to sample. It can be taken into account by adjusting our estimates with what we call the standard error.



Ethical Considerations

Always consider the following for a statistical study:

- **Confounding variables** are variables that could have an impact on the outcome or relationships, but were not considered in the analysis.
- **Stakeholders** are those who are interested in or would be affected by the research and findings.
- **Consequences** are what might result following the presentation of the research and findings.
- **Implications** are how findings might be used for policy, practice, or future research.
- **Limitations** are considerations that might impact the applicability of the research and findings (based on design, assumptions, …).

Variables

A **variable** describes some characteristic of each sampling unit.

We differentiate between:

- A **response** variable: measuring the phenomenon/outcome.
- Some **predictor** variables: trying to explain the phenomenon/outcome.

Identify one response variable and some predictor variables that should be collected in the following scenario

A local animal shelter wants to understand the factors that influence how long it takes for cats to get adopted in their facility.

Variables

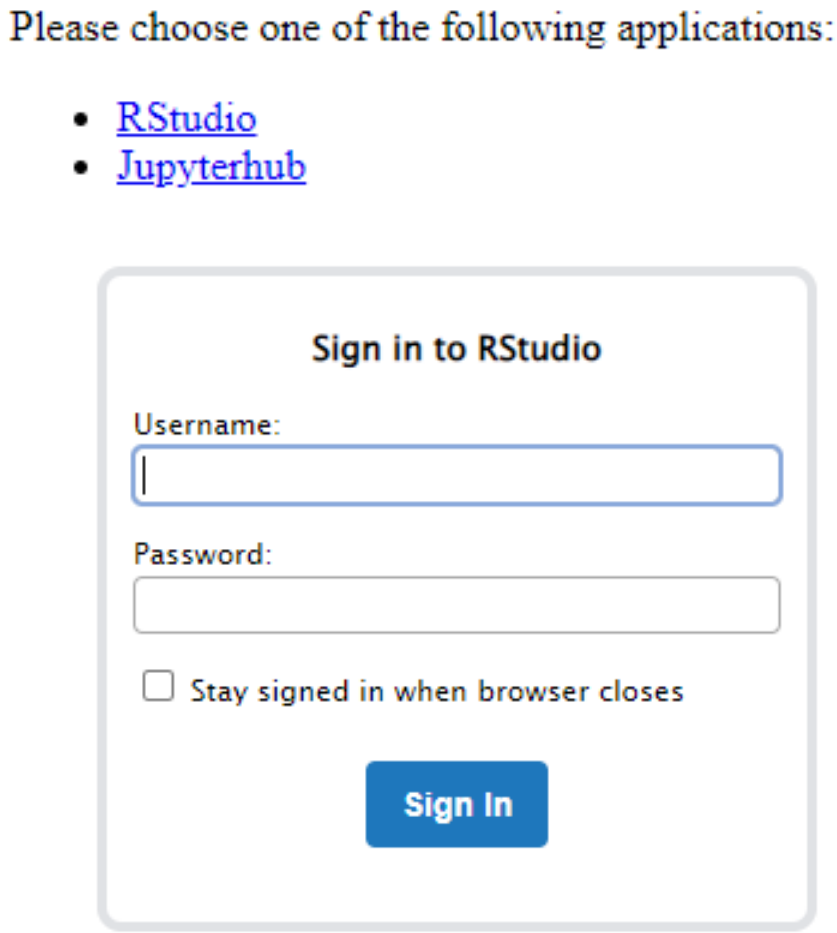
Type	Scale	
Categorical	**Ordinal**	**Nominal**
	Groups with a natural order	Groups with no natural order
		
Numeric	**Discrete**	**Continuous**
	Only few possible values 	Infinitely many possible values 

USING R AND RSTUDIO

Using R and RStudio

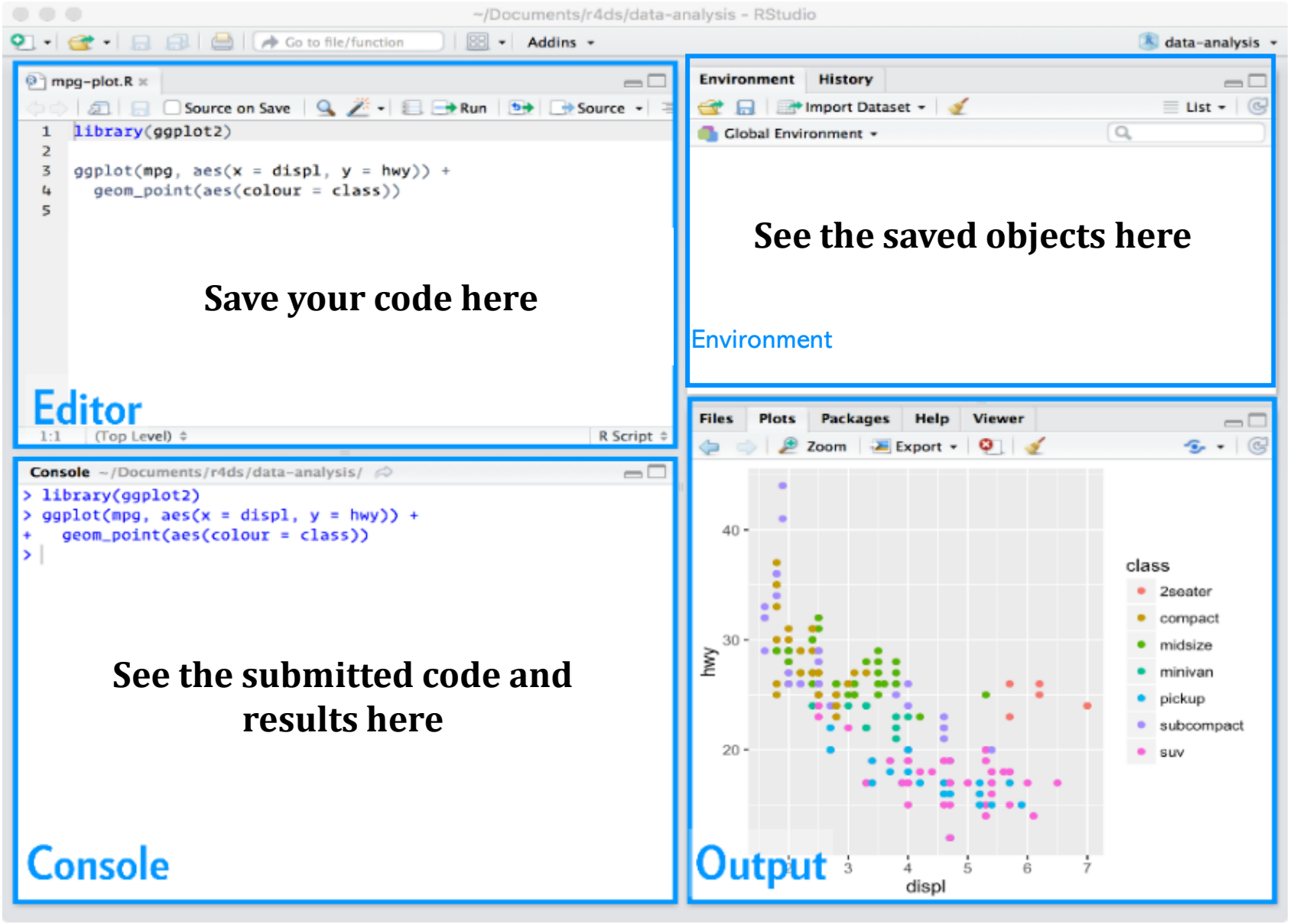
You will use RStudio from your web browser:

- ✓ Open your favorite browser (Chrome, Firefox, Safari, …)
- ✓ Go to https://gsafcomp02.ccbb.utexas.edu and select RStudio
- ✓ Log in with **student46-student70**
- ✓ Password is **CbrsSummer25**



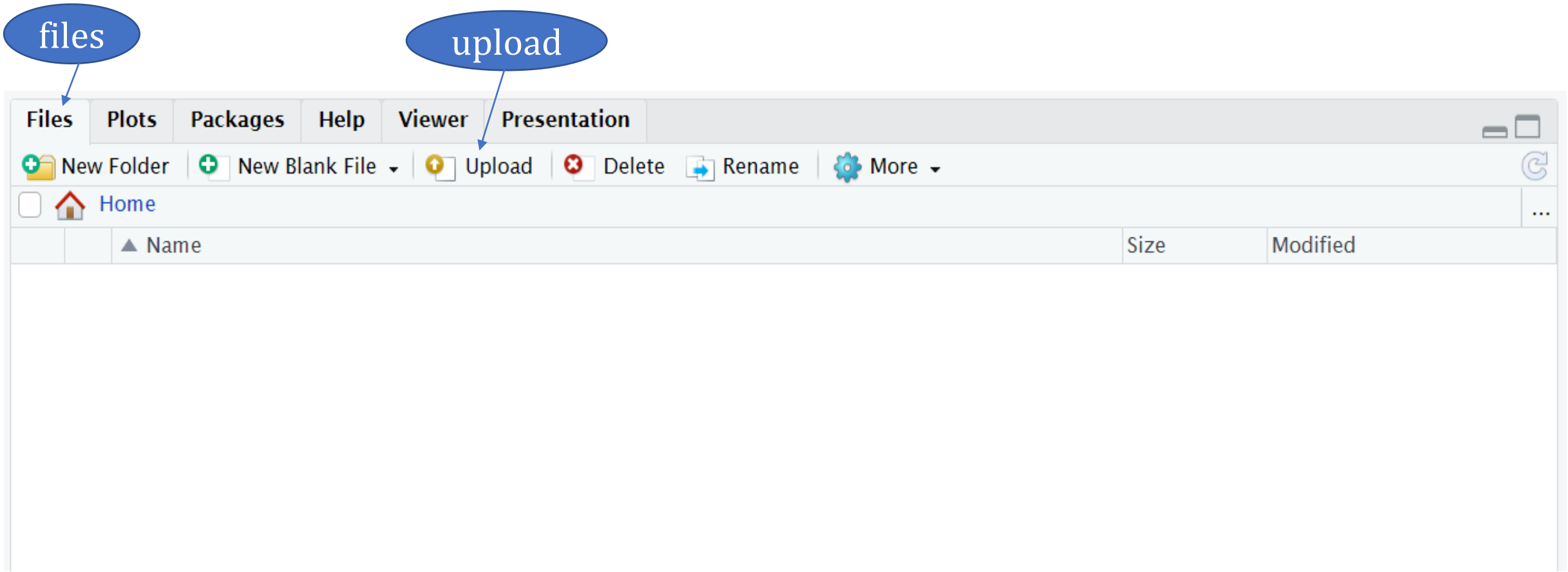
Using RStudio

When you open RStudio, you will see 4 panes:



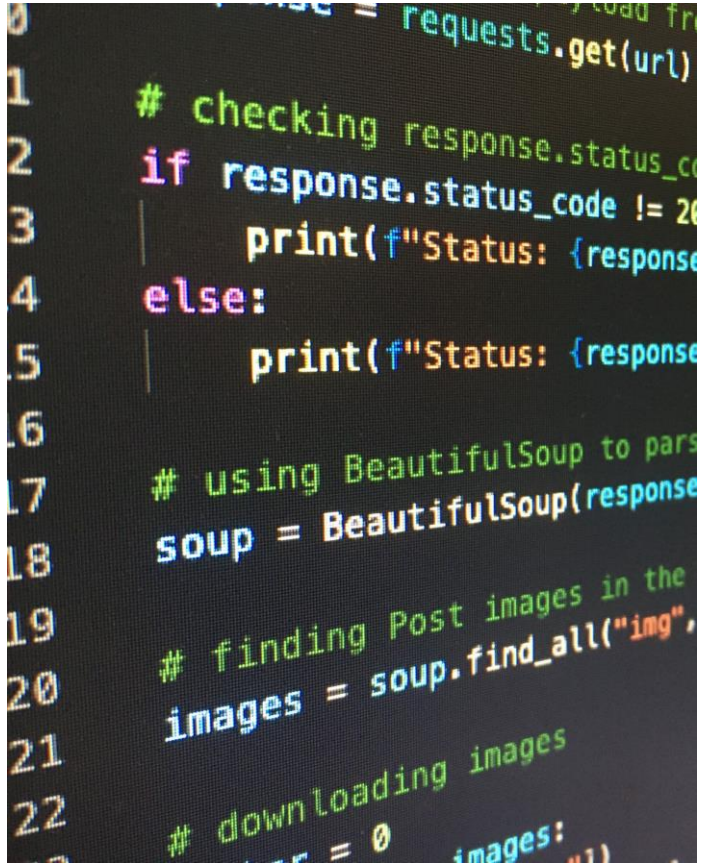
Using RStudio

Your files will be saved on the server, but you can upload and export files (datasets, outputs, …):



Coding

- ❑ No need to memorize lines of code: organize your code so that you can quickly find an example of “code that works” when you need it.
- ❑ Write an R file to save your code. Leave comments around your code to note what each piece of code does.
- ❑ When you get stuck, ask us! Or refer to the community. For example, stackoverflow.com



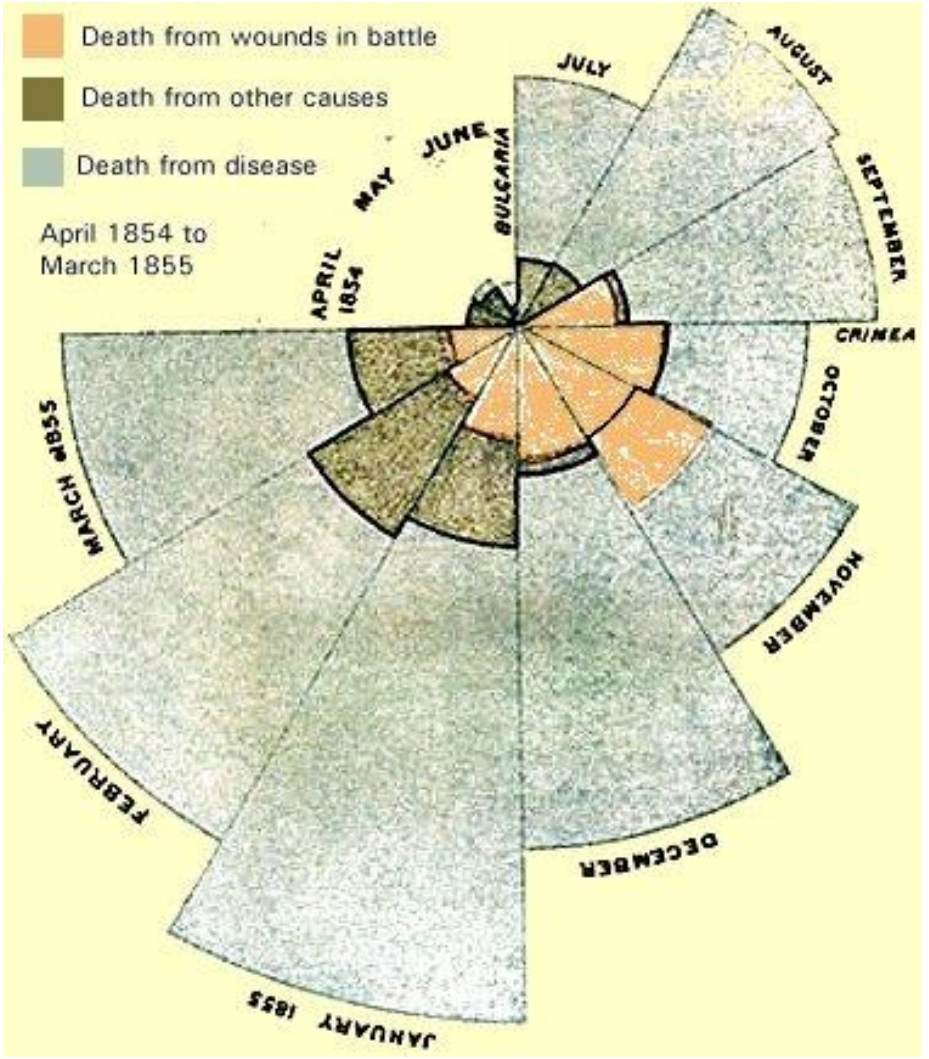
TRY IT!

Describe and Display

What do you notice in this graph?
Anything you wonder/you are curious about?

Visualization pioneer: Florence Nightingale

- Collected data and visualized the causes of death of British troops during the Crimean War (1853-1856)
- With the graphs, she campaigned for military and public health measures



Describing 1 variable

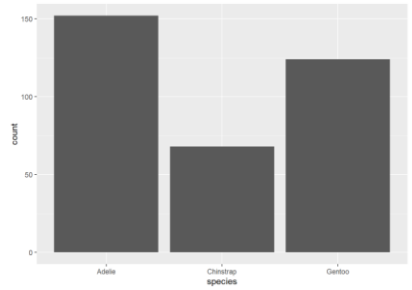
Univariate

visualizations and summary statistics

- depend on the **types** of variables
- help us identify typical or rare values

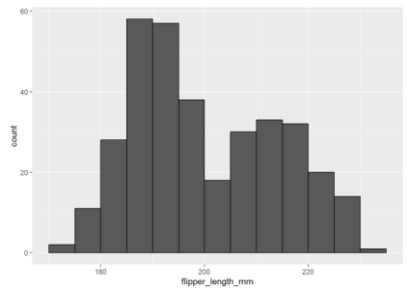
Categorical

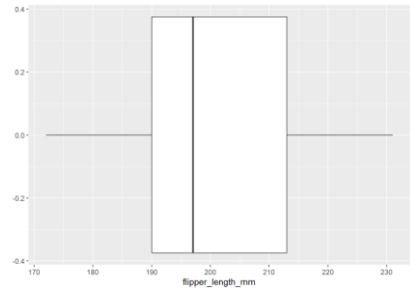
- Bar plot
- Frequency
- Relative frequency



Numeric

- Histogram or Boxplot
- Center: Mean, Median
- Spread: SD, IQR





Describing relationships

Multivariate

visualizations and summary statistics

- depend on the **types** of variables
- help us identify patterns

We usually identify:

- A **response** variable which measures an outcome.
- Some **predictor** variables which might explain why the outcome varies.

RQ: Does the predictor variable affect the response variable?

Describing relationships

Multivariate

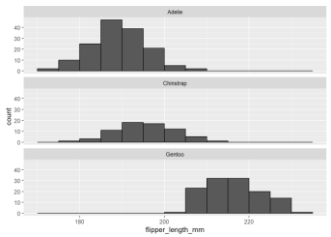
visualizations and summary statistics

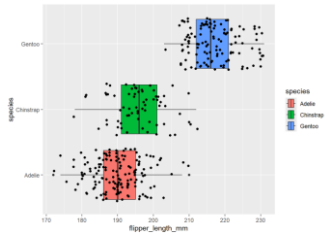
- depend on the **types** of variables
- help us identify patterns

Categorical

Numeric

- Grouped Histogram or Grouped Boxplot
- Center and Spread for each group

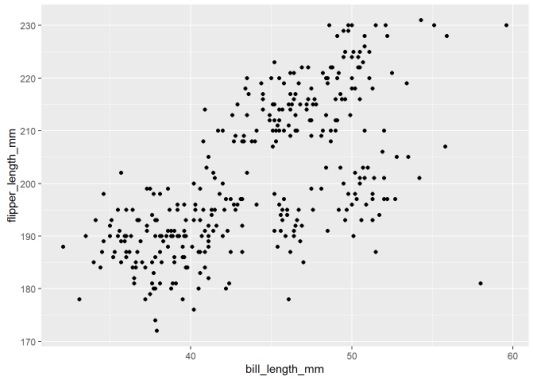




Numeric

Numeric

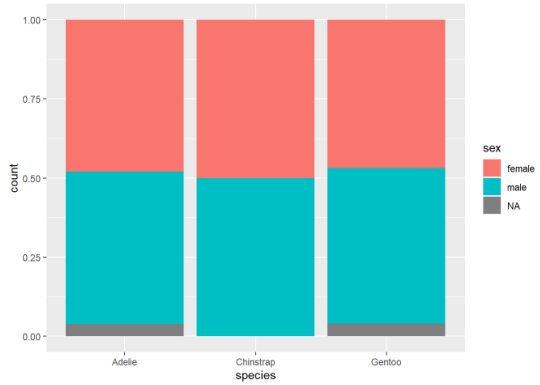
- Scatterplot
- Correlation



Categorical

Categorical

- Segmented Bar plot
- Contingency table



TRY IT!

Introduction to Hypothesis Testing

What is hypothesis testing?

It uses data from **a sample** to draw conclusions about a larger **population**. It helps us decide whether patterns in the data reflect **real** relationships, or just random variation.

- What would my sample look like if the true mean was something else?
- Does a relationship observed in the sample reflect a real relationship in the population?

This process relies on **counterfactual reasoning**: imagining what the data might look like if a null hypothesis were true. Does my sample look like that? If not, reject the null hypothesis.

Introduction to 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.

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

Find descriptive **statistics**

Suppose the **null hypothesis is TRUE**. If we had many random samples from the population, how does our sample correspond to the null distribution?

How **likely** were we to observe what we observed under the null hypothesis?

Introduction to 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

Decisions can be correct or incorrect…

Test Decision \ Reality	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"?

Type I error (false positive)

Type II error (false negative)

Next

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

Any questions? comments?