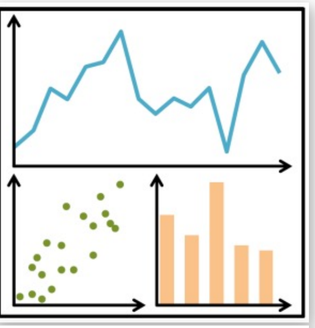
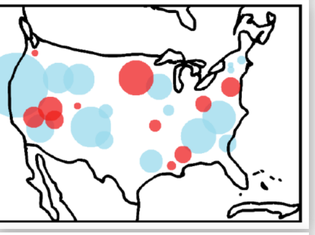
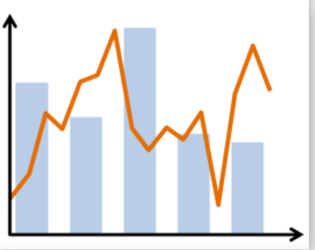
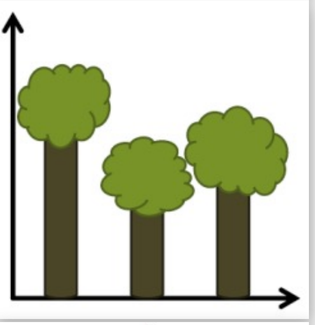
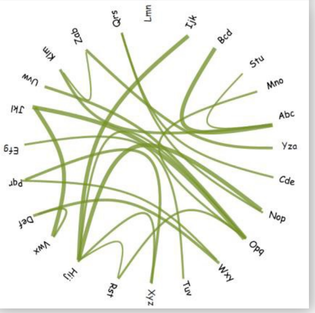






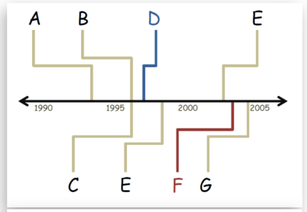
Lecture 14: Evaluation Analysis

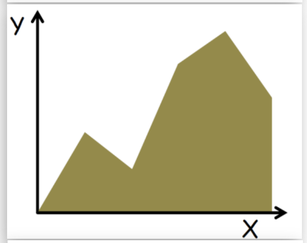
DS 4200

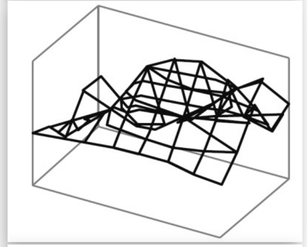
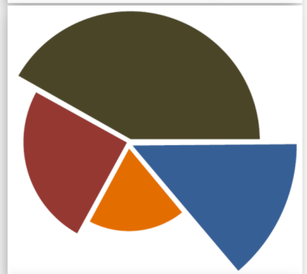
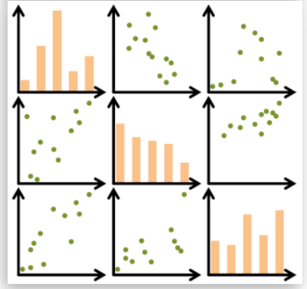
SPRING 2023

Prof. Ab Mosca (they/them)

NORTHEASTERN UNIVERSITY





Slides and inspiration from Cody Dunne, Michelle Borkin, Remco Chang, Dylan Cashman, Krzysztof Gajos, Hanspeter Pfister, Miriah Meyer, Jonathan Schwabish, and David Sprague

Last Class

We:

- Reviewed Validation and Evaluation

Any Questions?

Evaluation Analysis

Evaluation Design

1) Goal

→ Munzner's nested model

2) Method

→ Select an evaluation technique

3) Who

→ Who are your participants (and how many)

4) Tasks

→ What will the participant do

5) Metrics

→ Decide what data to collect

6) Analysis

→ How will you analyze collected data

Evaluation Design

1) Goal

→ Munzner's nested model

2) Method

→ Select an evaluation technique

3) Who

→ Who are your participants (and how many)

4) Tasks

→ What will the participant do

5) Metrics

→ Decide what data to collect

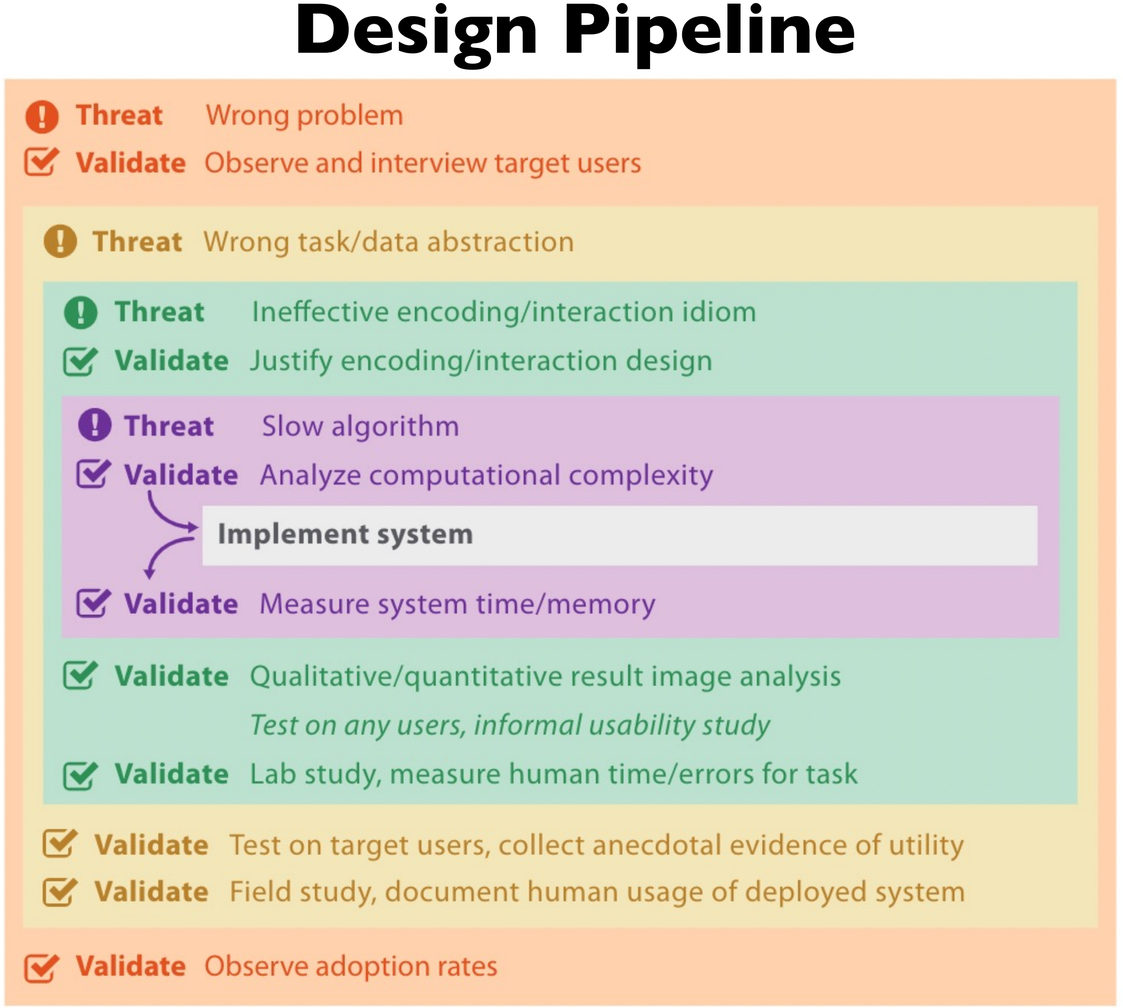
6) Analysis

→ How will you analyze collected data

Goal: Munzner's Nested Model

→Threats to validity exist at each stage of visualization design

→Validate every step using appropriate technique



Method

Theoretical / Simulation Based

→ **Inspection or Principled Rationale**

→ **Theoretical Analysis**

→ **Benchmarks**

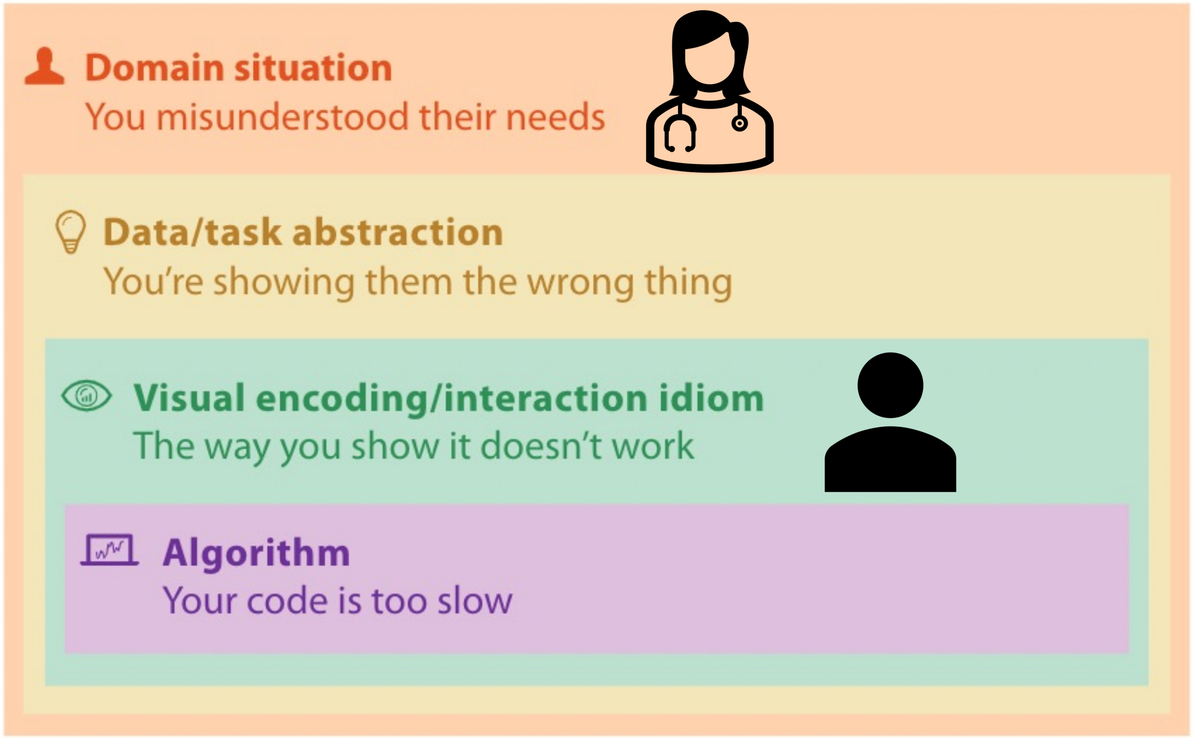
Human Subject Experiments

→ **Controlled Experiment**

→ **Observational Study**

→ **Field Deployment or Case Studies**

Who



Human Subject Experiments

→ **Controlled Experiment**

→ **Observational Study**

→ **Field Deployment or Case Studies**

Tasks

Tasks depend on goal

- Typical workflow
- Find something interesting
- Find outliers
- Report difference between X and Y
- Find/explain trend(s)

Metrics

Metrics depend on goal and task

→ **Common metric choices include:**

- Speed
- Accuracy
- Quality of finding
- Satisfaction
- Recall/Memorability
- Requiring help
- Completeness and Thoroughness of Analysis
- False Positives and False Negatives
- Immersiveness
- Deviation from optimality
- Information accessed
- Use frequency
  - → How often do users use this tool?
- Effort / Ease of Use
  - → Cognitive
  - → Physical (how many clicks did it take)
  - → NASA-TLX

Evaluation Design

1) Goal

→ Munzner's nested model

2) Method

→ Select an evaluation technique

3) Who

→ Who are your participants (and how many)

4) Tasks

→ What will the participant do

5) Metrics

→ Decide what data to collect

6) Analysis

→ How will you analyze collected data

ANALYSIS

Analysis

1) Goal

→ Munzner's nested model

2) Method

→ Select an evaluation technique

3) Who

→ Who are your participants (and how many)

4) Tasks

→ What will the participant do

5) Metrics

→ Decide what data to collect

6) Analysis

→ How will you to analyze collected data

Evaluation Methods

Theoretical / Simulation Based

→ Inspection or Principled Rationale

→ Theoretical Analysis

→ Benchmarks

Human Subject Experiments

→ Controlled Experiment

→ Observational Study

→ Field Deployment or Case Studies

Evaluation Methods

Theoretical / Simulation Based

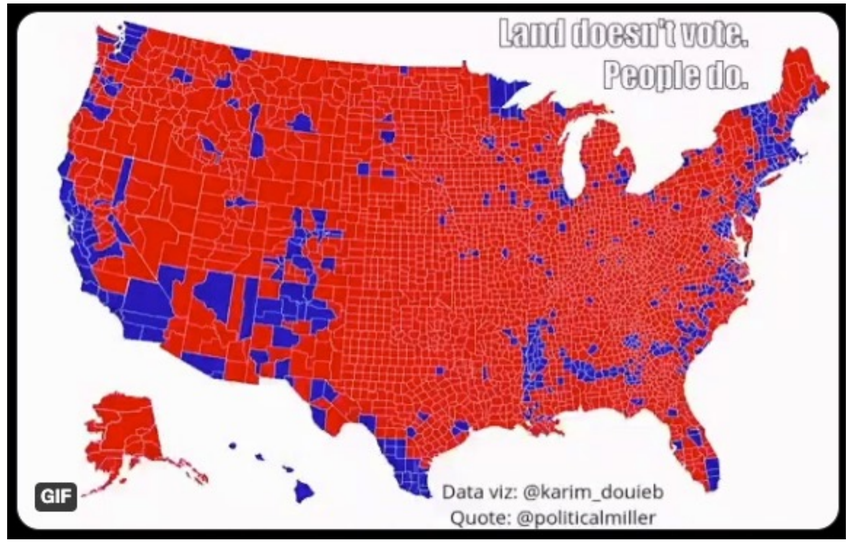
→ **Inspection or Principled Rationale**

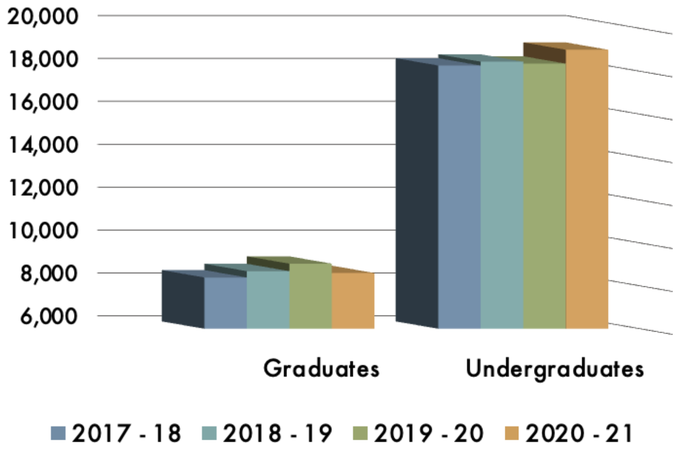
→ Apply design heuristics, perceptual principles

→ **Qualitative comparison to design guidelines**

Ex. This uses the wrong data – channel mapping

Ex. This violates a design rule of thumb





https://twitter.com/bettinaforget/status/1324139878666391555?lang=en

Evaluation Methods

Theoretical / Simulation Based

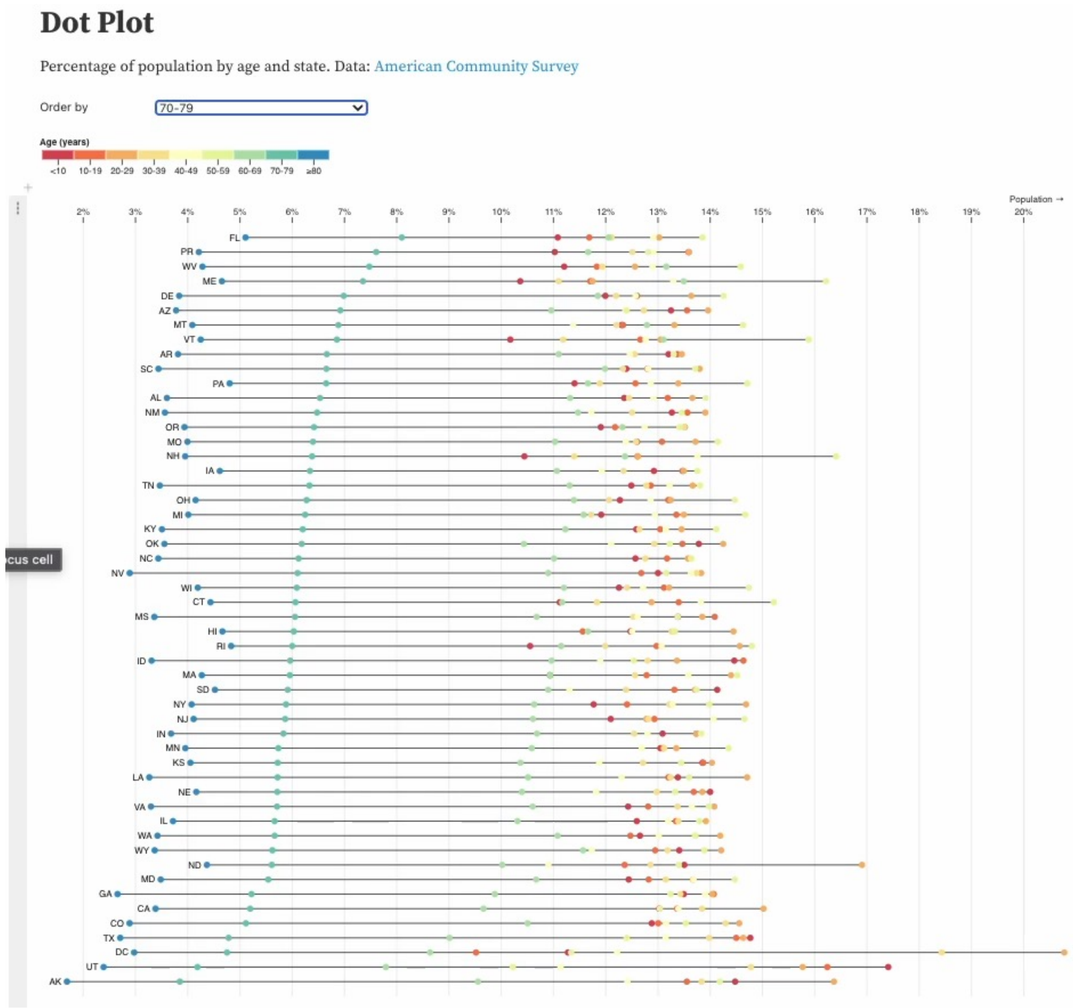
→ **Theoretical Analysis**

→ Algorithm time and space complexity

→ **O(n) analysis**

Ex. This loads in O(n), whereas this loads in $O(n^2)$

Ex. This filters in O(log n) whereas this filters in $O(n^2)$



https://observablehq.com/@d3/dot-plot

Evaluation Methods

Theoretical / Simulation Based

→ **Benchmarks**

→ Scalability to larger data sets

→ Performance (e.g., interactive frame rates)

→ **Quantitative comparison to industry standards**

Ex. This responds to interaction in less than 1 second

Time Constant	Value (in seconds)
perceptual processing	0.1
immediate response	1
brief tasks	10

https://observablehq.com/@d3/versor-dragging

Evaluation Methods

Human Subject Experiments

→ **Controlled Experiment**

→ Specific repeated task across multiple conditions

→ Collect quantitative data (ex. time, accuracy)

→ **Quantitative comparison across conditions**

Ex. T-Test, ANOVA, Chi Squared

A
New Vis

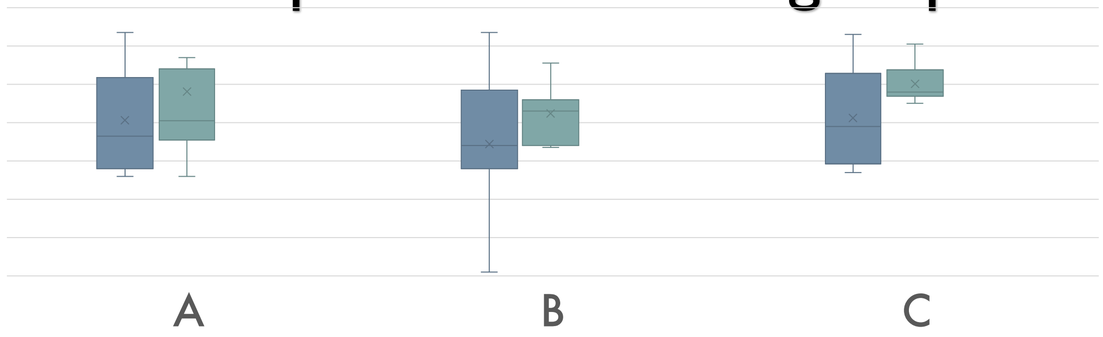
B
Comparison Vis 1

C
Comparison Vis 2

Task: Find all outliers in the data

Collect: Time, Accuracy

Compare data across groups



Evaluation Methods

Human Subject Experiments

→ **Observational Study**

→ Specific repeated task across multiple conditions (or not)

→ Collect qualitative data (ex. user findings, think-out-loud data, satisfaction) and quantitative

→ **Qualitative comparison across conditions**

Ex. Coding analysis, compare Likert scores

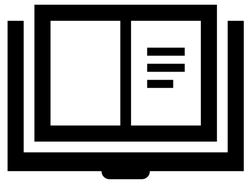
A New Vis	B Comparison Vis 1	C Comparison Vis 2

Task: Find all outliers in the data

Collect: Think out loud, Likert scales

Compare data across groups

- Code data & compare themes
- Compare # insights
- Compare Likert scale evaluations on satisfaction and ease of use



Evaluation Methods

Human Subject Experiments

→ **Observational Study**

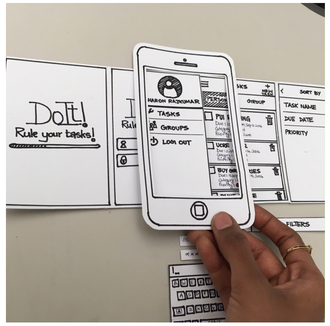
Usability Studies

→ Observational study specifically focused on usability of a visualization

→ Range from low to high fidelity testing

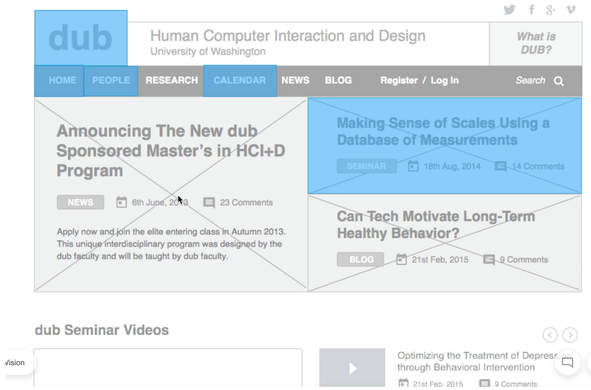
Qualitative analysis

Ex. How many errors? Are major functions missing?



Low Fidelity

→ Paper prototype

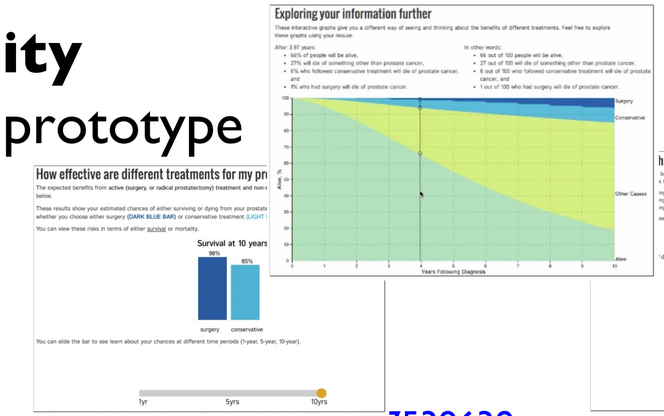


Medium Fidelity

→ Wireframe

High Fidelity

→ Working prototype



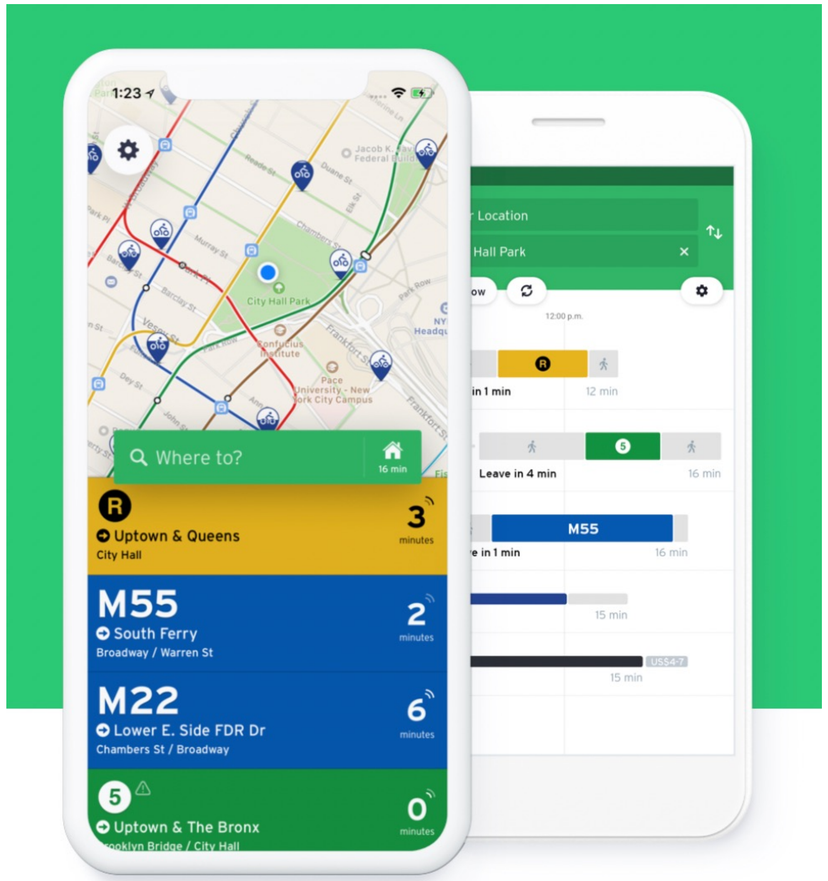
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7539638

Evaluation Methods

Human Subject Experiments

→ **Field Deployment or Case Studies**
- → Deploy tool in realistic settings
- → Document effects on work practices
- → Observe usage in the "real world"
- → Collect qualitative or quantitative data (ex. satisfaction, number of times used in a week, etc.)
- → **Qualitative analysis**

Ex. How often is it used? Did workflow change?



https://transitapp.com/

Evaluation Methods

Theoretical / Simulation Based

→ **Inspection or Principled Rationale –** Qualitative comparison to design guidelines

→ **Theoretical Analysis –** O(n) analysis

→ **Benchmarks –** Quantitative comparison to industry standards

Human Subject Experiments

→ **Controlled Experiment –** Quantitative comparison across controlled conditions

→ **Observational Study –** Quantitative or qualitative comparison

→ **Field Deployment or Case Studies –** Qualitative or quantitative analysis "in the wild"

Controlled Experiment Design

Human Subject Experiments

→ **Controlled Experiment**
- → Specific repeated task across multiple conditions
- → Collect quantitative data (ex. time, accuracy)

→ There are many ways to design a controlled experiment (i.e. many experimental designs). We'll touch on some more later.

A
New Vis

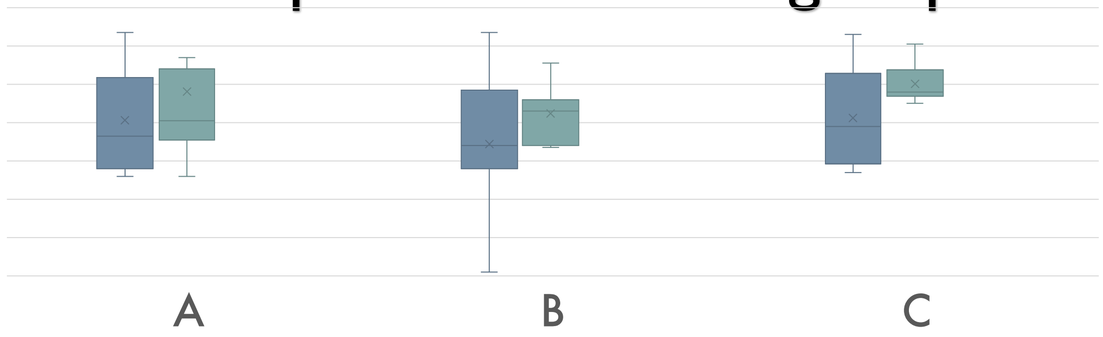
B
Comparison Vis 1

C
Comparison Vis 2

Task: Find all outliers in the data

Collect: Time, Accuracy

Compare data across groups

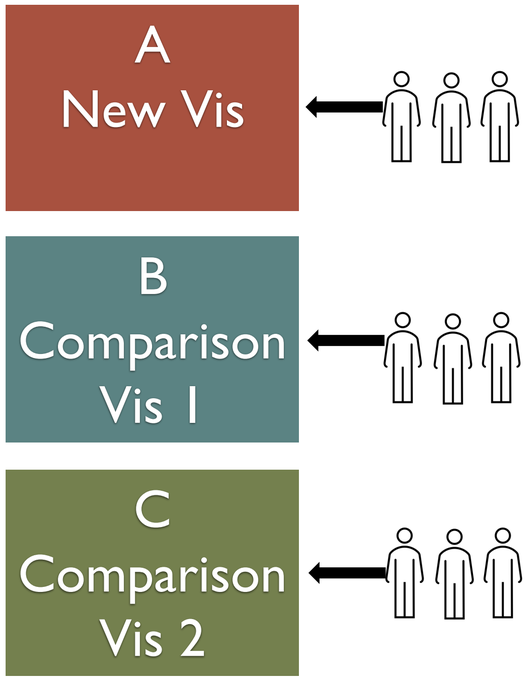


Controlled Experiment Design

Between Subject Design

→Each participant sees one stimuli

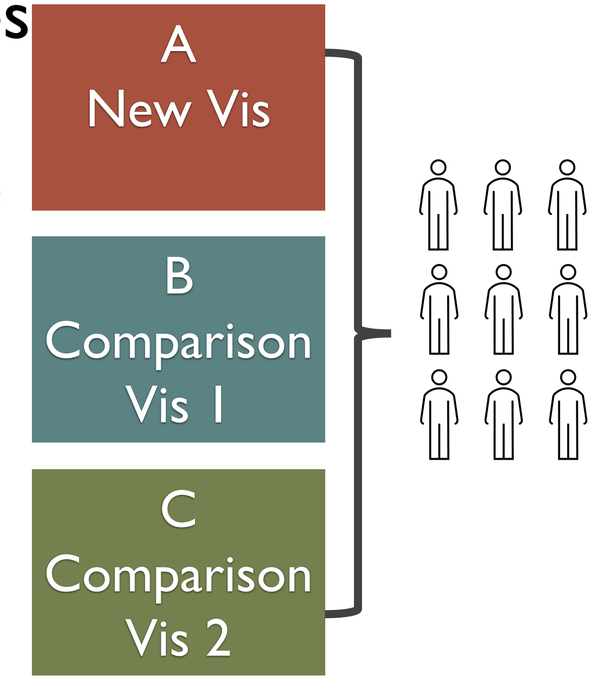
→Typically randomize stimuli assignment



Within Subject Design

→Each participant sees all stimuli

→Typically randomize stimuli order



Controlled Experiment Design

Between Subject Design

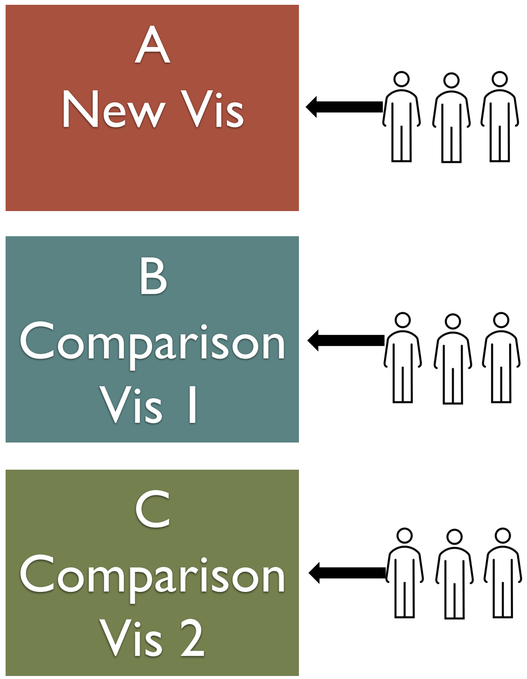
→Pros

→ No risk of learning effects

→Cons

→ Requires larger sample

→ No direct comparisons (same person) between conditions



Within Subject Design

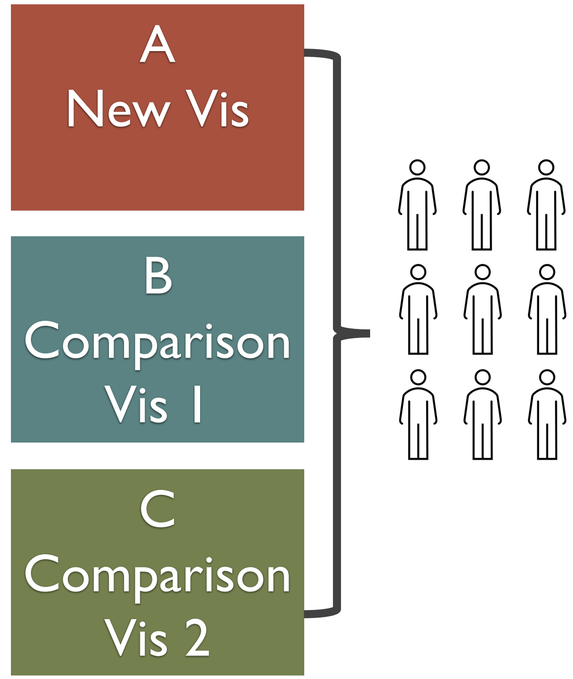
→Pros

→ Direct comparison (same person) between conditions

→ More data with fewer participants

→Cons

→ Learning effect

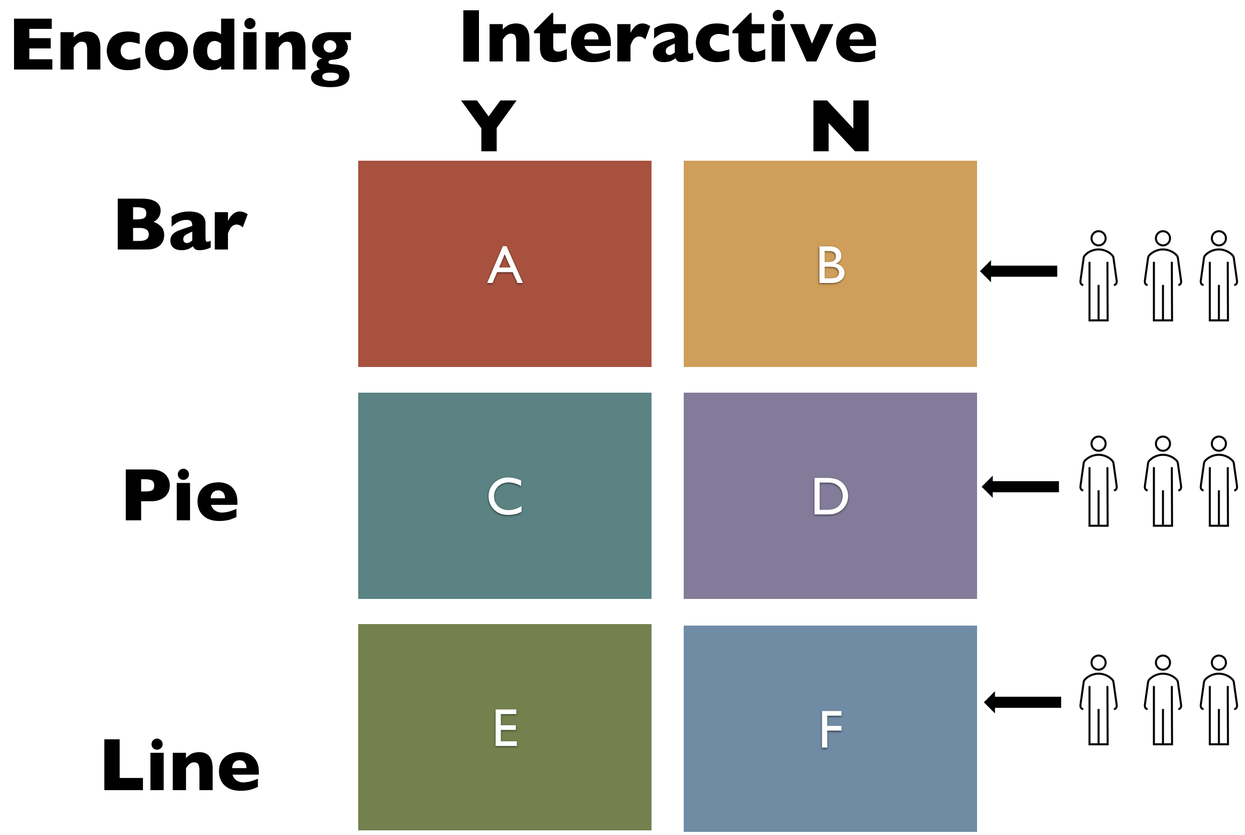


Controlled Experiment Design

Mixed

→ Experiment with a **combination of between and within** subject factors

→ **Factors** = variable aspects of an experiment



Statistical Analyses

- Results of statistical analyses allow us to statistically say whether two (or more) groups of subjects behave differently given different treatments (with all other differences controlled)

- Disclaimer: statistics is a deep field. Our goal today is to give you a “way to think” about statistics with respect to controlled user study evaluations

Statistical Analyses

Analyses start with a research question

→ This question relates to the **goal** of your evaluation
→ Ex. Does my visualization reduce the time it takes to complete Task A compared to the current state-of-the-art?

That research question is translated into a testable null hypothesis

→ A hypothesis is a statement that can be true or false
→ A null hypothesis is the statement that nothing new or interesting is happening
→ Ex. There is no difference between the amount of time it takes people to complete Task A with my visualization versus the current state-of-the-art.

Hypotheses

Research Question: Does my new visualization (V1) help people more accurately identify outliers in data than my old visualization (V2)?

Null Hypothesis

→ H_0:

Alternative Hypothesis

→ H_A:

Hypotheses

Research Question: Does my new visualization (V1) help people more accurately identify outliers in data than my old visualization (V2)?

Null Hypothesis

→ H_0: People using V1 identify outliers with the same accuracy as people using V2.

Alternative Hypothesis

→ H_A: People using V1 identify outliers with different accuracy than people using V2.

Hypotheses

Null Hypothesis

→ H_0: People using V1 identify outliers with the same accuracy as people using V2.

Alternative Hypotheses

→ H_A: People using V1 identify outliers with different accuracy than people using V2. **(2 tailed)**

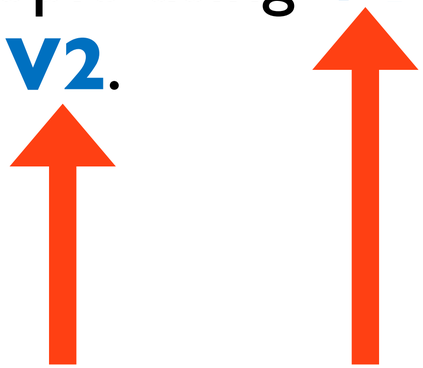
→ H_{A1}: People using V1 identify outliers with the higher accuracy than people using V2. **(1 tailed)**

→ H_{A2}: People using V1 identify outliers with the lower accuracy than people using V2. **(1 tailed)**

Hypotheses

Null Hypothesis

→ H_0: People using **V1** identify outliers with the same **accuracy** as people using **V2**.



Conditions / Groups

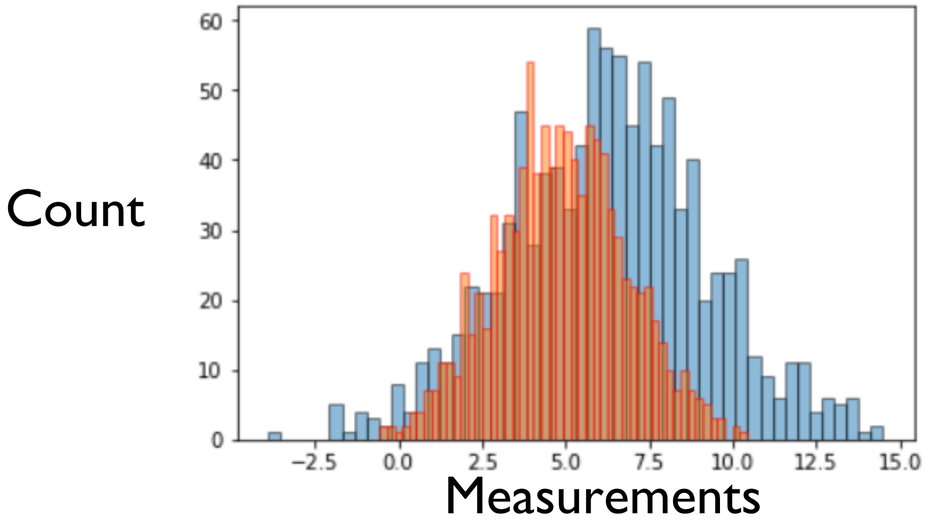
Metric / Response Variable

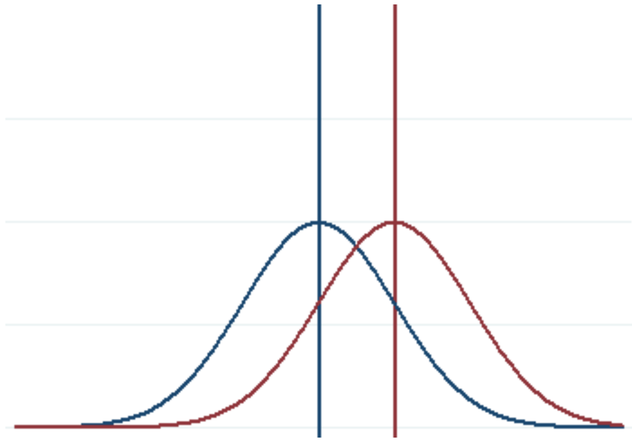
Modeling Data

- We collect data, and think about that data collectively as a distribution

- Most statistical tests compare distributions

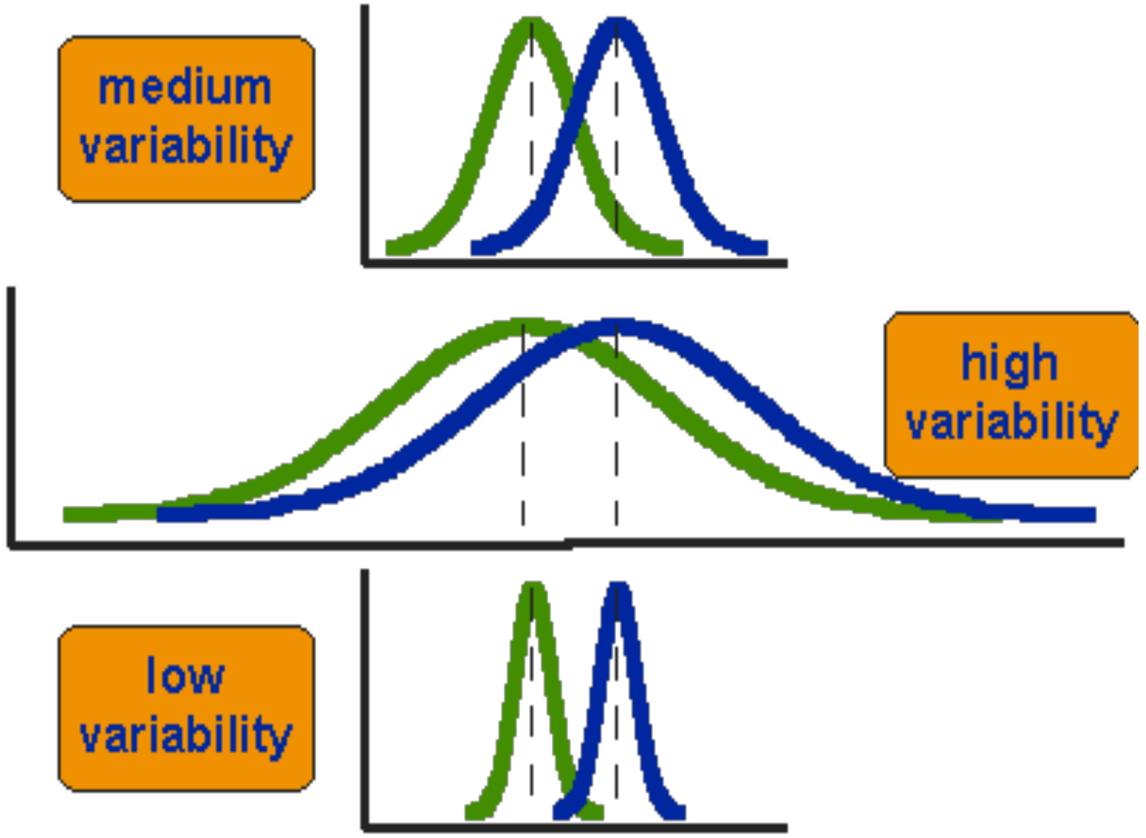
- **Parametric tests** compare normal distributions
- **Non-parametric tests** compare non-normal distributions





Hypothesis Testing

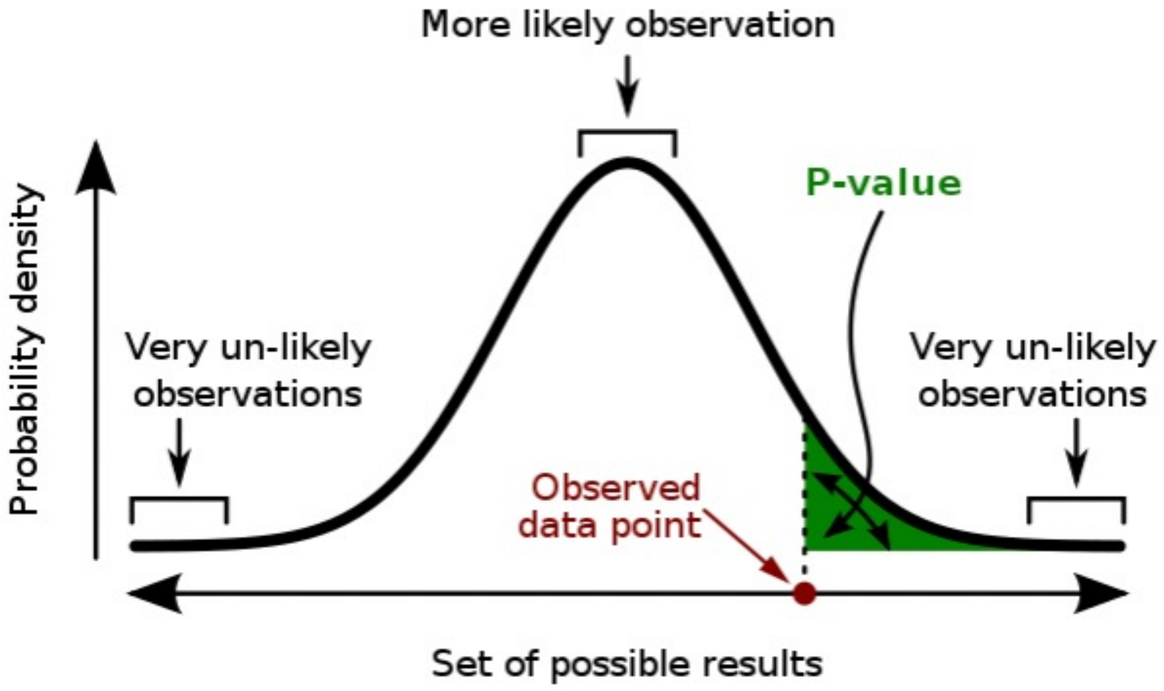
- Parametric statistical tests compare normal distributions
- They consider differences in means and amount of overlapping values (variability) when comparing distributions



Hypothesis Testing

- Statistical tests **calculate a test statistic**, based on experimental data and several parameters we set before hand (alpha, beta, etc.)

- We compare that test statistic to some pre-determined critical threshold to determine if we **reject** or **fail to reject** the null hypothesis

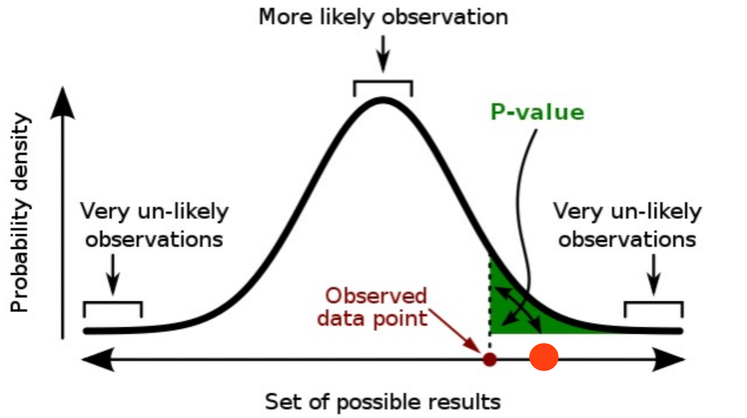


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

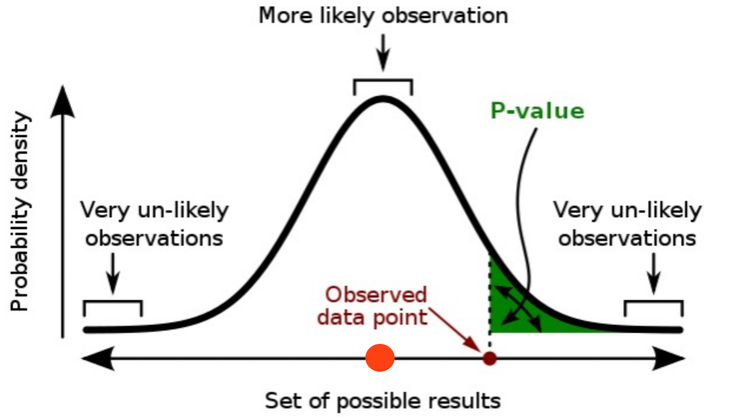
Hypothesis Testing

H_0: People using V1 identify outliers with the same accuracy as people using V2.

- If our test statistic is **sufficiently unlikely** to be seen by chance, we **reject** the null hypothesis
  - It is sufficiently unlikely that our experimental data differs from the null hypothesis only due to chance (i.e. **H_0** is most likely not true).
- If our test statistic is **likely** to be seen by chance, we **fail to reject** the null hypothesis
  - It is likely that our experimental data differs from the null hypothesis only due to chance (i.e. **H_0** is most likely true)



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Hypothesis Testing

Selecting the appropriate statistical test depends on your data and experimental design.

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

Source: http://yatani.jp/teaching/doku.php?id=hcistats:start

Hypothesis Testing

Selecting the appropriate statistical test depends on your data and experimental design.

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

Source: http://yatani.jp/teaching/doku.php?id=hcistats:start

Google is your friend! Always confirm your data meets assumptions before proceeding!

Summary

Today we:

- Reviewed analysis of evaluations

ic-14 is DUE today.
pm-04 is DUE before next class.