Intro to Quantitative Methods in HCI (or how to not lie with a t-test)

Prof. Benji Xie (“she”+”eh”), he|they
Asst. Professor of Computer Science, DU
benji.xie@du.edu benjixie.com

I’m Dr. Benji. I design human-data interactions for equity

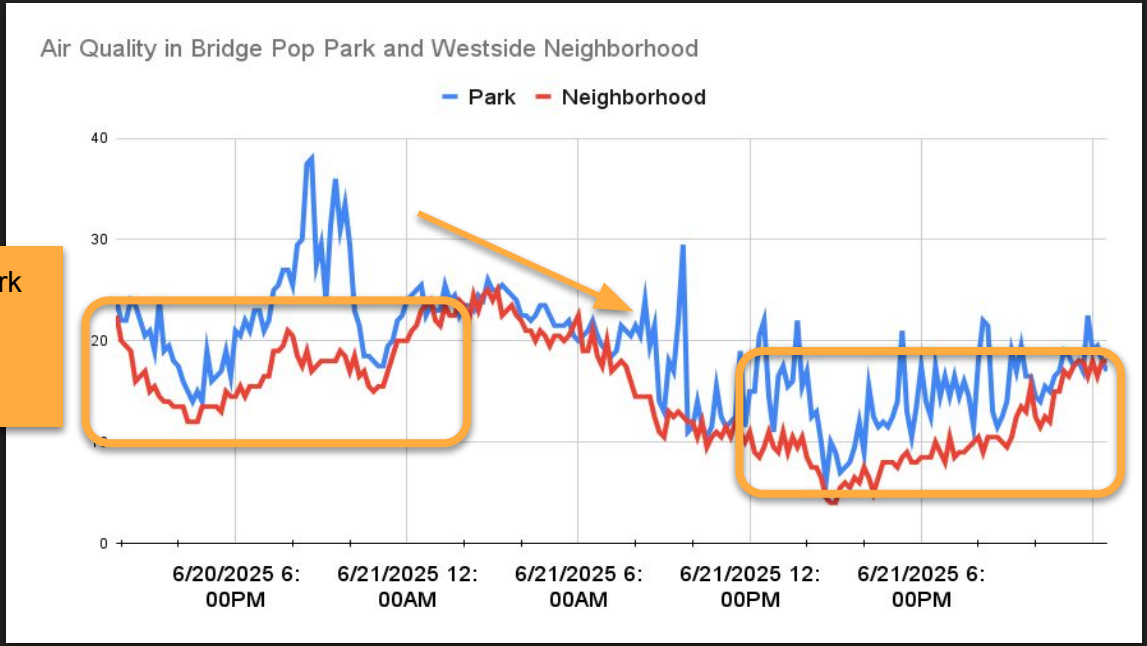
- Research: AI evaluation, AI-assisted programming
- Previously…
  - Postdoc @ Stanford
  - PhD @ University of Washington Information School (w/ Prof. Al!)
  - “4+1” at MIT
  - Ebay, Code.org, a few tech start-ups

How have I used quantitative methods?

- Objective measures: Comparing to agreed upon standard
  - Measuring performance w/ atests
  - Measuring temperature, sound, light
- Subjective measures: individual or relative evaluation
  - Surveys of participants
- Behavioral data: measuring people’s behavior
  - Keystroke logs
  - Eye tracking data
  - Logs of how users navigate app or website

There is a difference in air quality in the park vs the neighborhood. Is it “significant?”

Air quality in park is worse at the park than the neighborhood



Air Quality is steadily higher

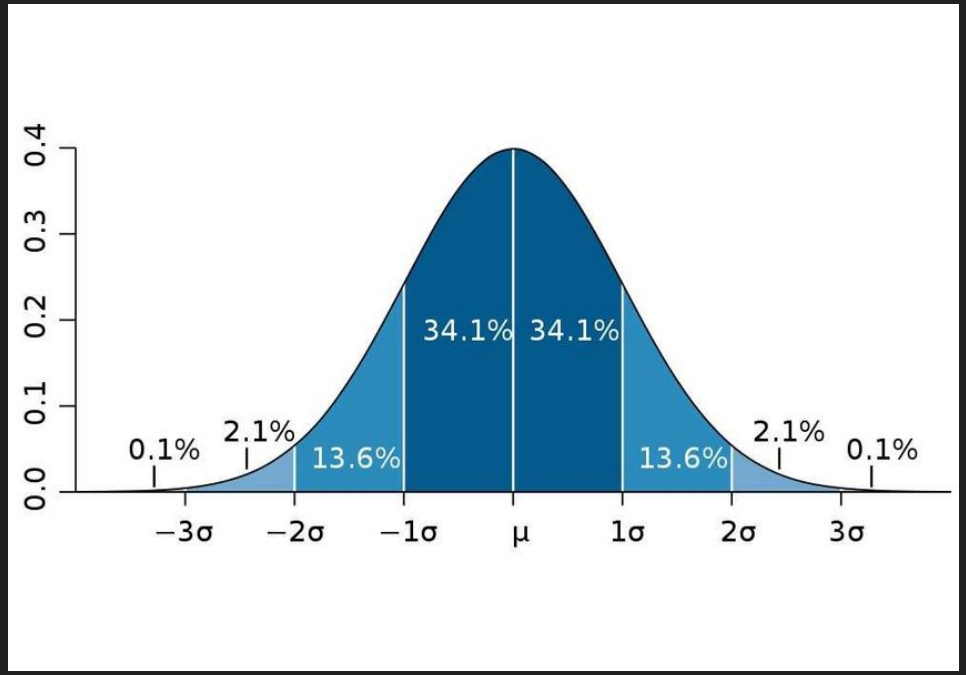
Learning Objectives

1. Understand what a p-value means
2. Understand how to conduct a t-test
    a. Check assumptions: normality, homoscedasticity (continuous, independence)
    b. Measure effect size
3. Understand how to apply a t-test in Python
4. Recognize limitations of a t-test and other approaches to mitigating these limitations

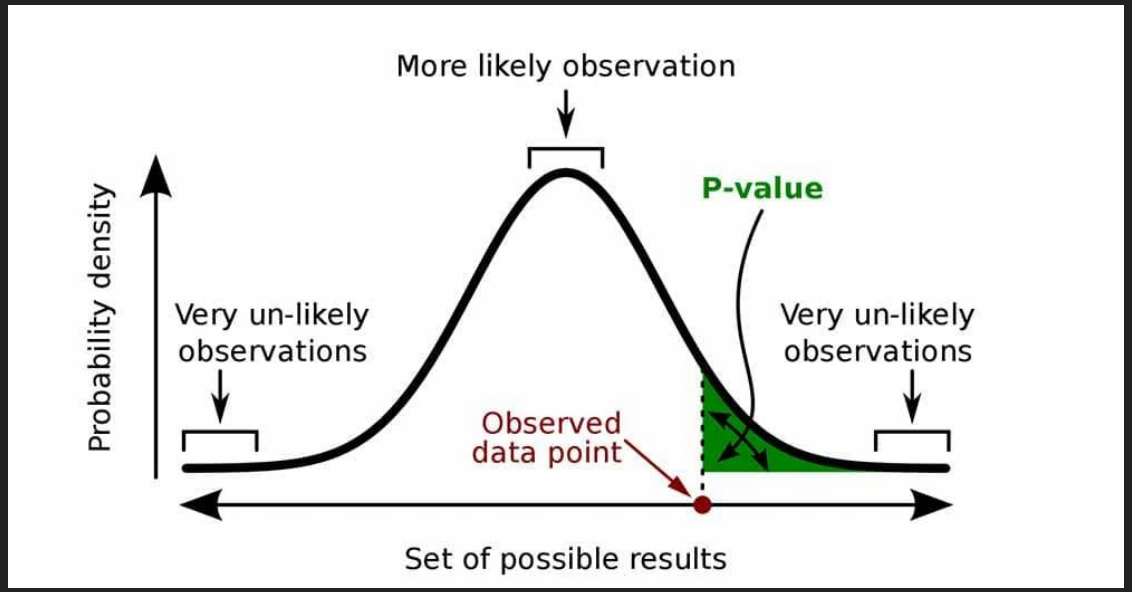
Stats review: normal distribution (“bell curve”)

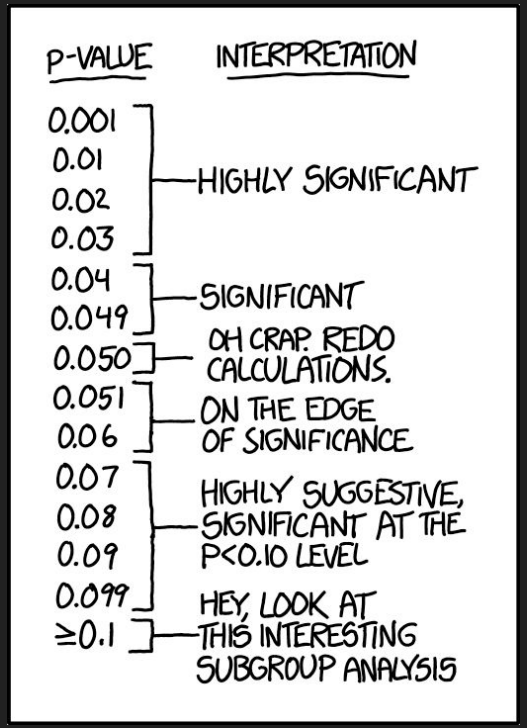
Represented by two values

- μ: mean (measure of central tendency)
- σ: standard deviation (measure of spread)

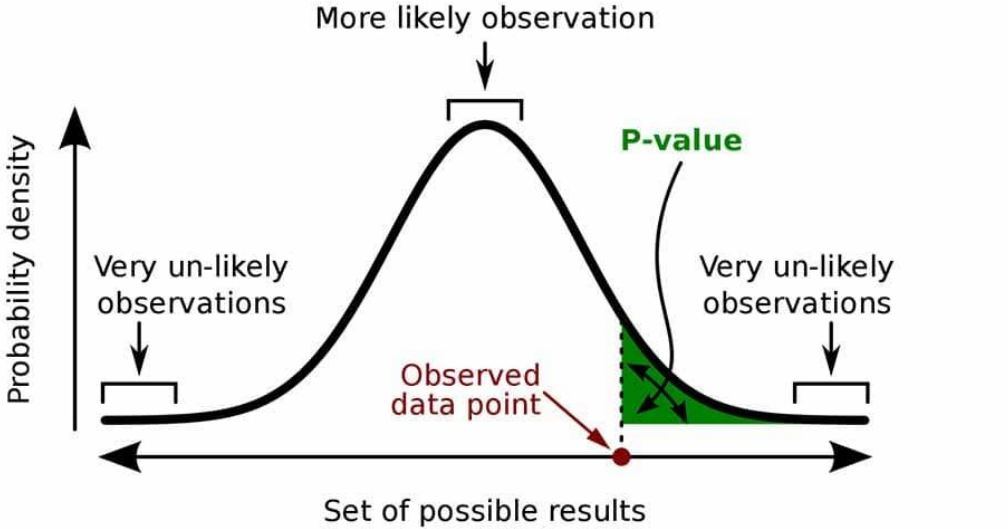


DISCUSS: What do you think a p-value means? Why does it matter?





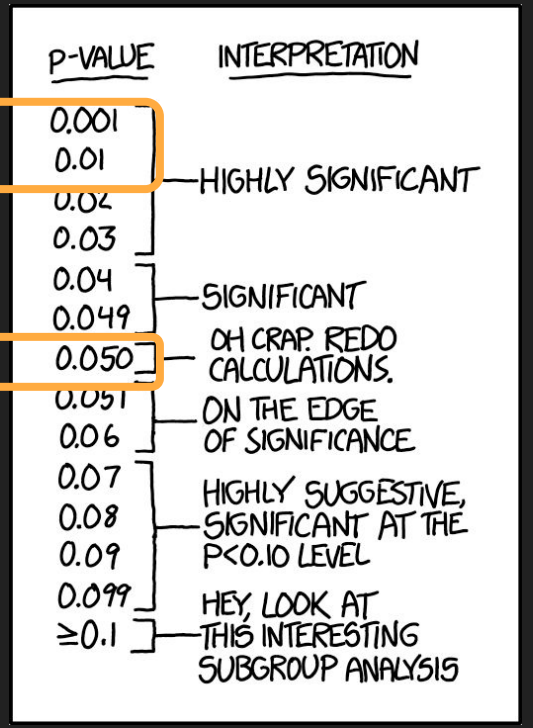
P-Value: Probability that data supports the null hypothesis



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

p=0.031 => 3.1% probability that data supports the null hypothesis

Typical thresholds (α): 0.05*, 0.01**, 0.001***



What to do with a p-value:

If $p<\alpha$ where $\alpha = \{0.05, 0.01, 0.001\}$, we reject the null hypothesis (“significance”)

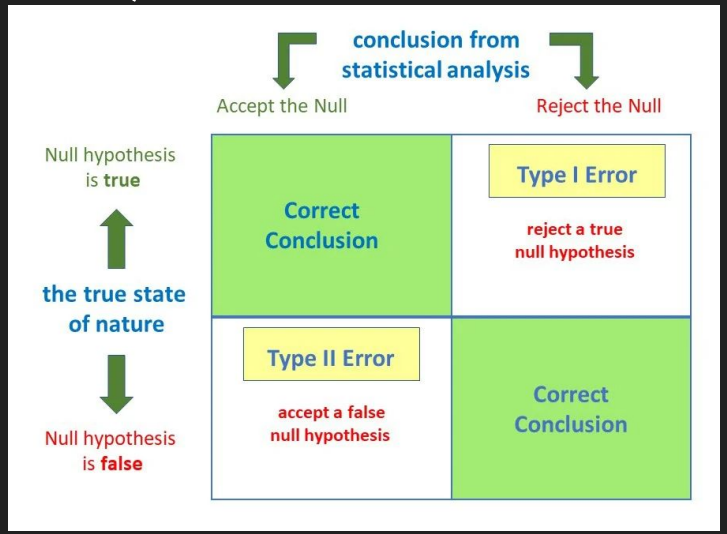
If $p \geq \alpha$, we have insufficient data/**fail to reject the null hypothesis** (“not significant”, n.s.)

- w/ frequentists approaches, we can never accept a null hypothesis. Bayesian approaches help with that

Types of errors

Type I error: false **positive** (reject null hypothesis when it's actually true)

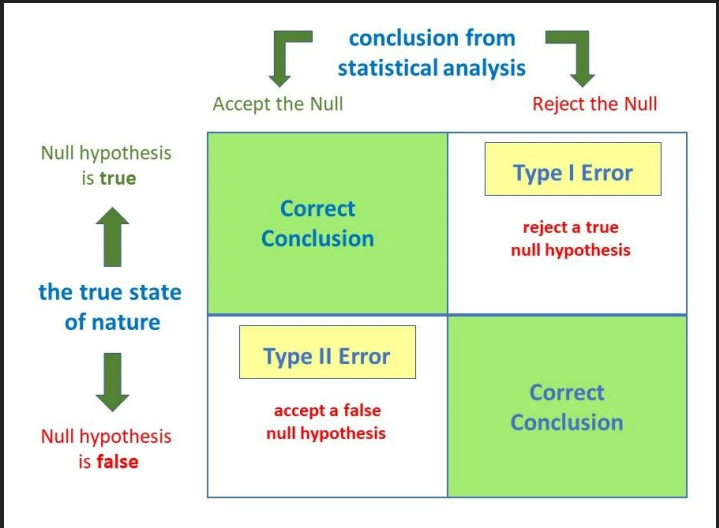
Type II error: false **negative** (fail to reject null hypothesis when it's actually false)



DISCUSS: How does changing p-value threshold from 0.05 to 0.01 affect Type I or Type II errors? (if at all)

Type I error: false **positive** (reject null hypothesis when it's actually true)

Type II error: false **negative** (fail to reject null hypothesis when actually false)



DISCUSS: How does changing p-value threshold from 0.05 to 0.01 affect Type I or Type II errors? (if at all)

Type I error: false **positive** (reject null hypothesis when it's actually true)

- P-value directly related to Type I error. => changing α from 0.05 to 0.01 decreases Type I error

Type II error: false **negative** (fail to reject null hypothesis when actually false)

- P-value inversely related to Type II error. => changing α from 0.05 to 0.01 increases Type II error

t-tests: how we use p-values to determine *significant* difference

t-test: determines difference in mean between 2 groups

- 1 factor (categorical independent variable)
- 2 levels (values or groups within factor)
  - e.g. factor: input device. Levels: mouse, touchpad
- Two kinds of t-test
  - **Independent/two-sample/student's**: groups from 2 different populations (*between* subjects
    - e.g. surveying people who went to Taylor Swift concert and people who did not
  - Paired t-test: same population (*within* subjects)
    - pre & post test, repeated samples
    - e.g. measuring people's motor functions before, during, and after attending Taylor Swift concert

DISCUSS: Which t-test for each setup (or neither)

1. Understand typing speed of two bird species: mallard vs stellar jay
2. Determine number of native bird species UGs of different years (freshman, sophomore, junior, senior) can name
3. Measure time it takes neurodiverse programmers to complete a given task with and without AI assistance.
4. Determine the relationship between age and number of Taylor Swift songs someone can recite from heart

DISCUSS: Can we use a t-test?

1. Understand typing speed of two bird species: mallard vs stellar jay
   a. ✅ 1 factor: bird species. 2 levels: {mallard, stellar jay}
2. Determine number of native bird species UGs of different years (freshman, sophomore, junior, senior) can name
   a. ❌ >2 levels (frosh, soph, …). Could change to underclassmen & upperclassmen
3. Measure time it takes neurodiverse programmers to complete a given task with and without AI assistance.
   a. ✅ 1 factor: AI assistance. 2 levels: {with, without}.
4. Determine the relationship between age and number of Taylor Swift songs someone can recite from heart
   a. ❌ > 2 levels (age). Could change to {child (<18 yrs old), adult (18+ yrs old)}

Example of t-test

```python
import numpy as np
from scipy.stats import ttest_ind, levene, shapiro
import pandas as pd
import matplotlib.pyplot as plt # for visualization
```

```python
# read data
df = pd.read_csv('air_quality.csv')
```

Is there a difference in air quality between East Palo Alto and Palo Alto?

factor: city levels: {East Palo Alto ("EPA or BH"), Palo Alto ("PA or MP")}

```python
aqi_epa = df[df['city'] == "EPA or BH"]['aqi']
aqi_paly = df[df['city'] == "PA or MP"]['aqi']
```

```python
t_statistic, p_value = ttest_ind(aqi_epa, aqi_paly)
p_value
```

```
1.8768108496139391e-23
```

```python
# AQI: higher means worse air quality
print(aqi_epa.mean())
print(aqi_paly.mean())
```

```
36.02056277056277
30.97697922515441
```

=> the air quality in East Palo Alto is significantly worse than the air quality in Palo Alto

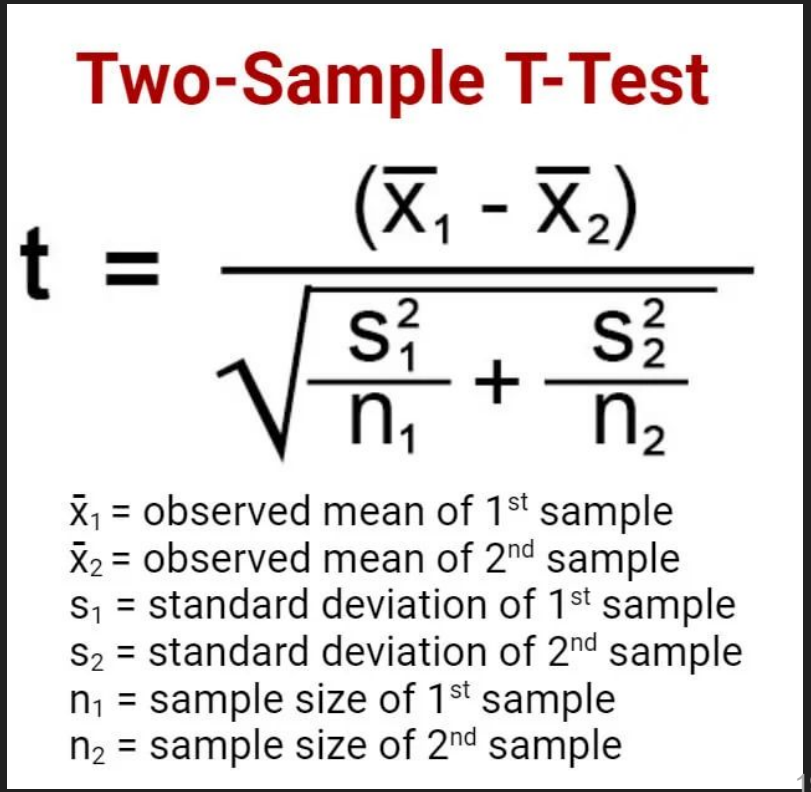
Assumptions of t-test…

Assumptions we justify by argument:

- Data is continuous
- Independent samples

Assumptions we can test for:

1. Normality: data approximately fits normal distribution
2. Homoscedasticity (same variance)



Testing for assumptions

Normality: Shapiro-WIlks

```
# Test for normality
test_statistic, pvalue = shapiro(df["aqi"])
pvalue # if <0.05, then fails Shapiro-Wilks test
```

```
4.207258509288831e-41
```

Homoscedasticity: Levene's Test

```
# test for homoscedacity https://docs.scipy.org/doc
statistic, pvalue = levene(aqi_epa, aqi_paly)
pvalue # if <0.05, then fails homoscedacity
```

```
0.02242654010431343
```

Effect Size: The important thing everyone forgets

Value measuring strength/magnitude of relationship between two variables

Calculate only if significant ($p<\alpha$)

For t-tests: Cohen's *d*

- 0.2: small
- 0.5: medium
- 0.8: large

Cohen's d formula:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^3 + s_2^3}{2}}}$$

Conclusion: Beyond t-tests

Other ANOVA tests

data not normal, variance not same

Analyses of Variance

Factors	Levels	Between or Within	Parametric Tests	Nonparametric Tests	
			Linear Models	Generalized Models	
1	2	B	Independent-samples *t*-test	Median test	Mann-Whitney *U* test
1	2	W	Paired-samples *t*-test	Sign test	Wilcoxon signed-rank test
1	≥2	B	One-way ANOVA	Kruskal-Wallis test	
1	≥2	W	One-way repeated measures ANOVA	Friedman test	
≥2	≥2	B	Factorial ANOVA Linear Model (LM)	Aligned Rank Transform (ART) Generalized Linear Model (GLM)	
≥2	≥2	W	Factorial repeated measures ANOVA Linear Mixed Model (LMM)	Aligned Rank Transform (ART) Generalized Linear Mixed Model (GLMM)	

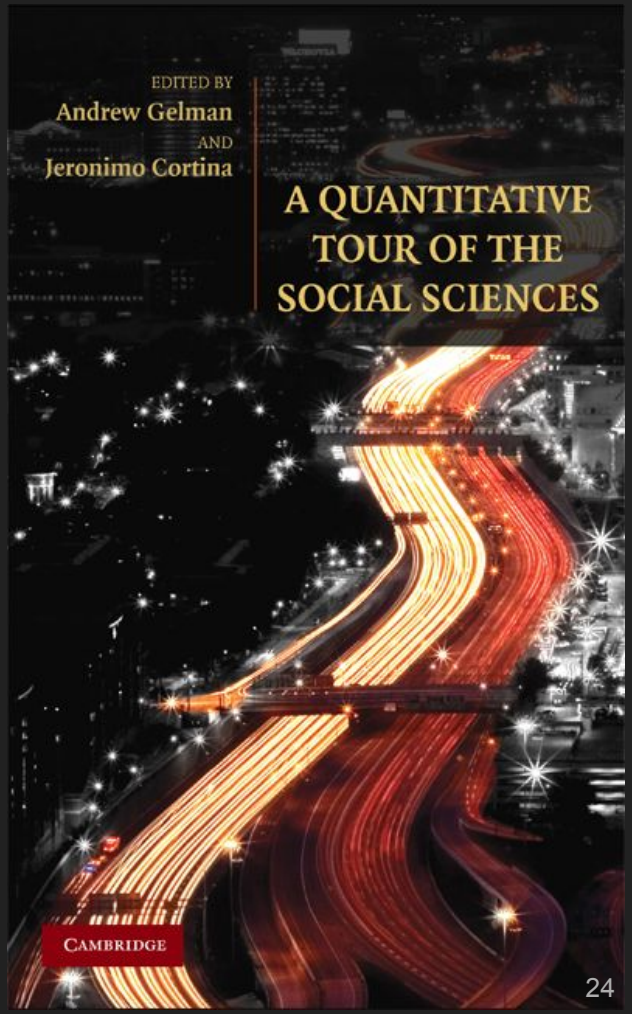
more than 1 factor or level

https://depts.washington.edu/acelab/proj/Rstats/index.html

Bayesian Approaches

Can provide evidence to accept null hypothesis

More robust approaches, but less understood

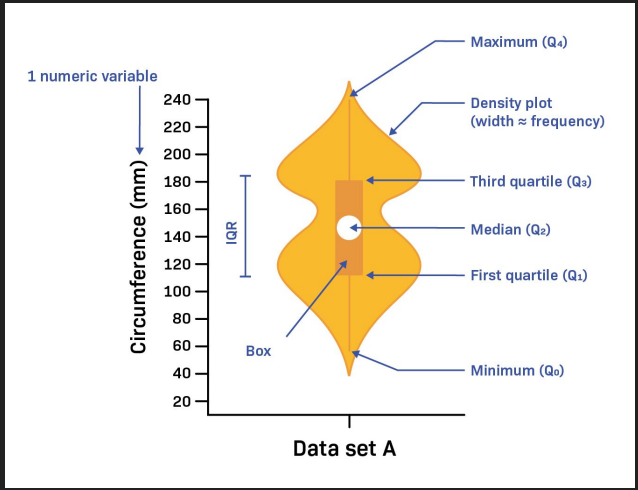


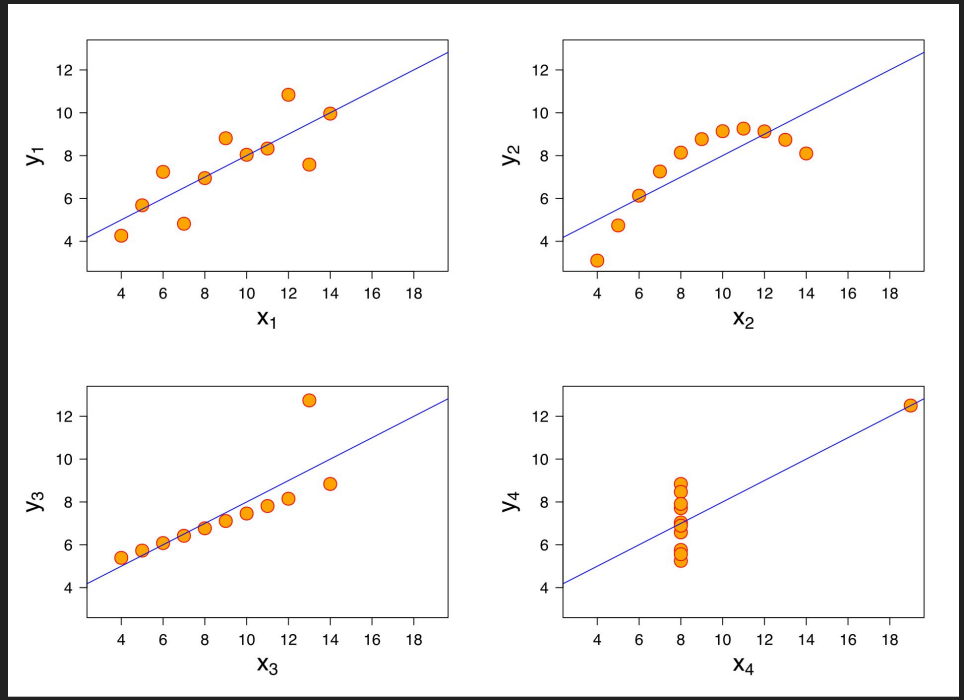
VISUALIZE your distribution, check ASSUMPTIONS

Explore data yourself, show data to others

Example: Anscombe's quartet (Same mean & variance)

Show each data point >> violin plot >> box plot





Anscombe's quartet

SHAMELESS PLUG: Join the DUHCI Group! 🦆

- If you are work-study eligible, apply to be our lab manager!
- Interested in any of the projects I'm working on? Reach out!
  - Usability of AI-assisted programming tools for neurodiverse programmers
  - Understand how people with different kinds of sexism use AI differently
  - Evaluating the impact of AI on CO communities (quantitatively, qualitatively)

@benji xie on Slack | benji.xie@du.edu | ECS 355

Learning Objectives

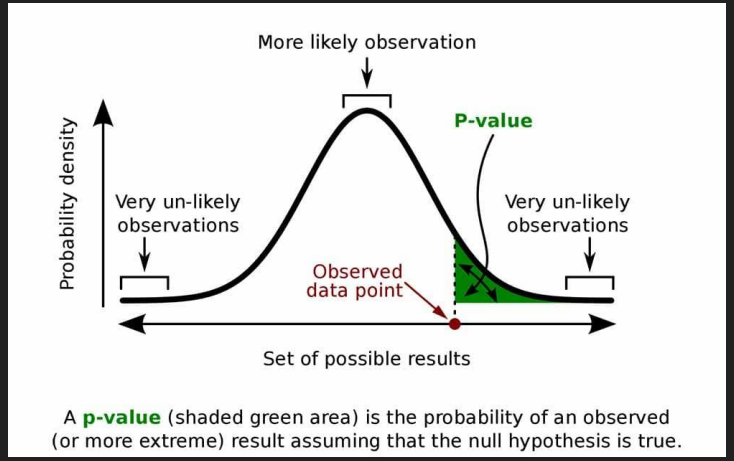
1. Understand what a p-value means
   a. p(data|hypothesis)
2. Understand how to conduct a t-test
   a. 1 factor, 2 levels. Between or within subjects
   b. Check assumptions: normality, homoscedasticity (continuous, independence)
   c. Measure effect size
3. Understand how to apply a t-test in Python
   a. R is easier
4. Recognize limitations of a t-test and other approaches to mitigating these limitations
   a. Assumptions, 1 factor & 2 levels, reliance on arbitrary thresholds (α={0.05, 0.01, 0.001})

Intro to Quantitative Methods in HCI (or how to not lie with a t-test)

Prof. Benji Xie (“she”+”eh”), he|they. Asst. Professor of Computer Science, DU
benji.xie@du.edu benjixie.com

Conducting a t-test

- 1 factor, 2 levels
- Visualize distribution (for self, audience)
- Check assumptions (normality, homoscedasticity)
- Report effect size



Two-Sample T-Test

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$ = observed mean of 1st sample
$\bar{x}_2$ = observed mean of 2nd sample
s_1 = standard deviation of 1st sample
s_2 = standard deviation of 2nd sample
n_1 = sample size of 1st sample
n_2 = sample size of 2nd sample