Experiments and Quantitative Analysis
Experiments

- Why now?
  - Some of you have comparative questions
  - I want to talk about some of the crisper results we have
    - "what's the bottom line?"
- Today: a crash course in experimental design and statistics
Observational vs. Interventional Studies

- Observational: watch something (without controlling it)
  - Can establish *correlation*
  - Do C programs have more security vulnerabilities than Haskell programs?
    - Look at a bunch of programs, count vulnerabilities, and compare
- Interventional: change something and watch what happens
  - Can establish *causation*
  - Recruit programmers, **assign them** to use C or Haskell, count bugs in resulting programs.
Causation Vs. Correlation

• I observe: it is raining. People are using umbrellas.

• Possibilities:
  • Rain causes people to use umbrellas.
  • Umbrellas cause rain.
  • Both rain and umbrellas are caused by a third, unseen factor (e.g. wind)
Causation

• Are people more likely to click ads if I put them at the top of the page or the bottom of the page?

• If I randomly choose whether the ad is at the top or bottom

• AND THEN people are more likely to click the ad in one condition

• THEN it is likely that the choice of top or bottom caused the difference in behavior.
Variables

• Independent variables: ones the experimenter controls

• Dependent variables: ones the experimenter measures

• Want to know if red squiggly underlines in IDEs help people finish tasks faster:
  • IV?
    • whether underlines appear
  • DV?
    • task completion time

• Confounding variable?
  • Color-blindness
Correlation

- What if it's impossible, too expensive, or unethical to manipulate an IV?
- Does smoking cause cancer?
- Is Rust better than C in projects with > 1M LOC?
Vocabulary

- Randomized controlled studies (RCTs)
- A/B tests: RCTs with only two conditions
Dealing With Confounding Variables

- Two options:
  - Control them (mitigate their effects or restrict participant population)
  - Record them
Multiple Factors

- What if there are two IVs?
- *Factorial* design: try every combination
- Example: factors influencing exam scores:
  - IV 1: test time (morning or afternoon)
  - IV 2: coffee (0 or 1 cups)
  - DV: exam scores
- 2 x 2 design

<table>
<thead>
<tr>
<th></th>
<th>morning</th>
<th>afternoon</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No coffee</strong></td>
<td>(scores)</td>
<td>(scores)</td>
</tr>
<tr>
<td><strong>1 cup of coffee</strong></td>
<td>(scores)</td>
<td>(scores)</td>
</tr>
</tbody>
</table>
Within Vs. Between

• Q: can people answer code understanding questions faster with if statements or with the ternary operator?

• Within-subjects: every participant gets both kinds of code problems

• Between-subjects: some participants get only if; others only get ternary operators

```java
int result;
if (x > 42) {
    result = 37;
}
else {
    result = 95;
}
```

```java
int result = (x > 42) ? 37 : 95;
```
Within Vs. Between

• Within: have to worry about learning effects. But otherwise more statistical power.

• Between: greater variance; might accidentally have demographic differences between groups

  • Randomly assign to conditions. Suppose participants in one group accidentally have more programming experience?

  • Need to balance groups.
Within Vs. Between

• Task: fix a bug in codebase X.

• Conditions: llmdb vs. `println` debugging

• Surely once you've fixed the bug once, it's much easier to fix it again!
Within Vs. Between

- Within also known as repeated measures
- Used with paired tests
## SIMPSON'S PARADOX

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th></th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
<td>Applicants</td>
</tr>
<tr>
<td>Total</td>
<td>8442</td>
<td>44%</td>
<td>4321</td>
</tr>
</tbody>
</table>

**UC Berkeley, Fall 1973**

Conclusion: discrimination against women?

Credit: Wikipedia contributors
### ADMISSIONS BIAS?

<table>
<thead>
<tr>
<th>Department</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
</tr>
<tr>
<td>F</td>
<td>373</td>
<td>6%</td>
</tr>
</tbody>
</table>

Bickel et al.: women tended to apply to competitive departments with low rates of admission even among qualified applicants (such as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants (such as in engineering and chemistry).

Credit: Wikipedia contributors
## KIDNEY STONES

<table>
<thead>
<tr>
<th></th>
<th>Treatment A</th>
<th>Treatment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small stones</td>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td></td>
<td><strong>93% (81/87)</strong></td>
<td>87% (234/270)</td>
</tr>
<tr>
<td>Large stones</td>
<td>Group 3</td>
<td>Group 4</td>
</tr>
<tr>
<td></td>
<td><strong>73% (192/263)</strong></td>
<td>69% (55/80)</td>
</tr>
<tr>
<td>Both</td>
<td>78% (273/350)</td>
<td><strong>83% (289/350)</strong></td>
</tr>
</tbody>
</table>

When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.

Credit: Wikipedia contributors
HYPOTHESIS TESTING

• Context: drawing from two populations.

• Null hypothesis: $\mu_1 = \mu_2$

• Alternative hypothesis: $\mu_1 \neq \mu_2$

• Question: what is the probability the null hypothesis is true?

• This is what $p$-value captures.
ERRORS

• Testing: $\mu_1 = \mu_2$

• Type 1 error: conclude $\mu_1 \neq \mu_2$ when $\mu_1 = \mu_2$

• Type 2 error: find no significant difference when $\mu_1 \neq \mu_2$

• $\alpha$: $P$ (type 1 error)

• $\beta$: $P$ (type 2 error)
• Recall: $\beta$: $P$(type 2 error)

• Power: $1 - \beta$

• Probability of rejecting null hypothesis if it is false

• Want more power?
  
  • Increase $N$
  
  • Decrease variance ($\sigma^2$)
  
  • Increase $|\mu_1 - \mu_2|$
EFFECT SIZE

• Small p-value does not imply a large effect!
• Instead, calculate effect size (Cohen's $d$)

$$d = \frac{\mu_1 - \mu_2}{s}$$

• $s$: pooled standard deviation

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very small</td>
<td>0.01</td>
</tr>
<tr>
<td>Small</td>
<td>0.02</td>
</tr>
<tr>
<td>Medium</td>
<td>0.5</td>
</tr>
<tr>
<td>Large</td>
<td>0.8</td>
</tr>
<tr>
<td>Very large</td>
<td>1.2</td>
</tr>
<tr>
<td>Huge</td>
<td>2</td>
</tr>
</tbody>
</table>