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

	morning	afternoon
No coffee	(scores)	(scores)
1 cup of coffee	(scores)	(scores)

Within Vs. Between

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

```
int result = (x > 42) ? 37 : 95;
```

- 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

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: lldb 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

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

UC Berkeley, Fall 1973

Conclusion: discrimination against women?

Credit: Wikipedia contributors

ADMISSIONS BIAS?

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

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

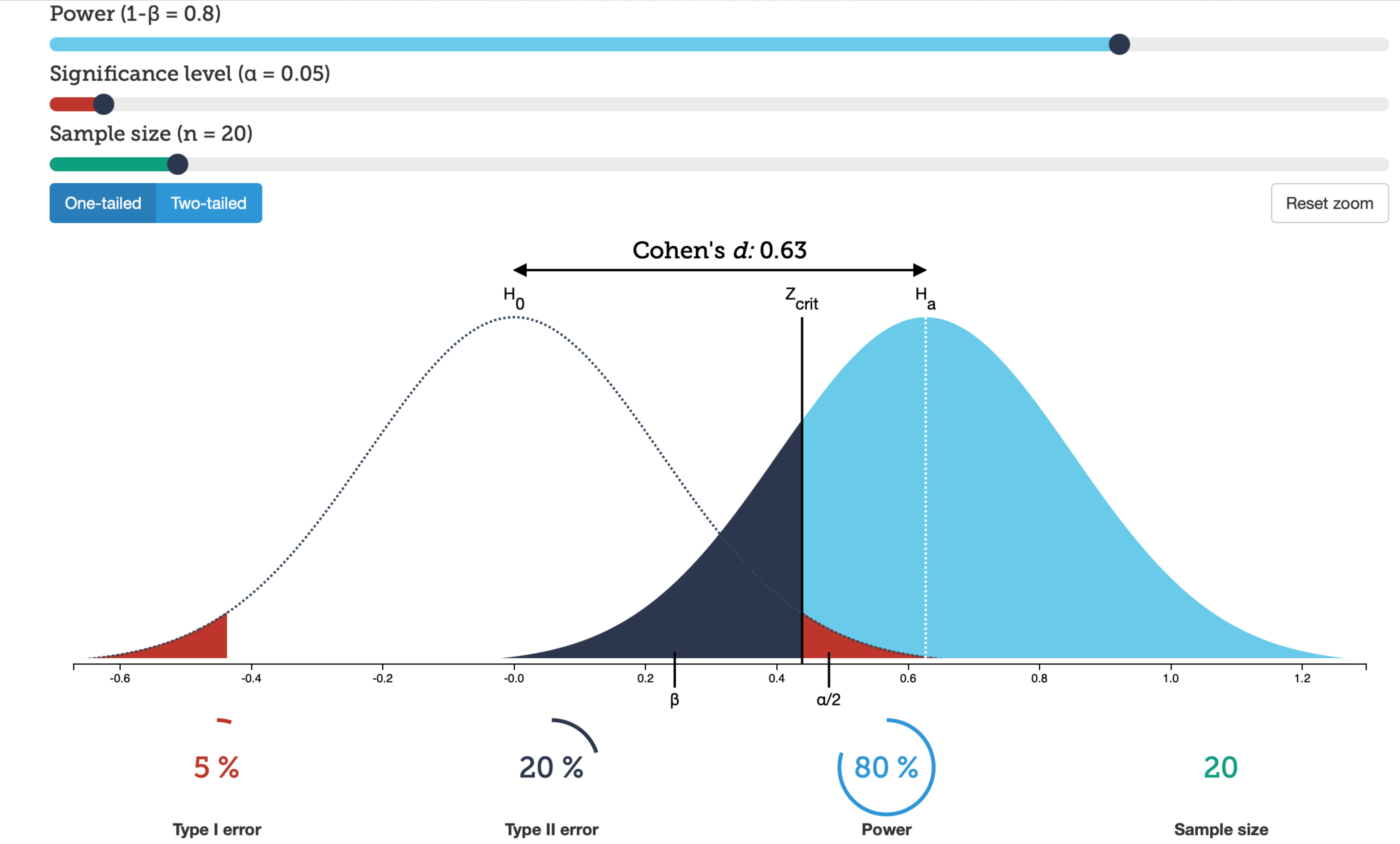
	Treatment A	Treatment B
Small stones	Group 1 **93% (81/87)**	Group 2 87% (234/270)
Large stones	Group 3 **73% (192/263)**	Group 4 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

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.



https://rpsychologist.com/d3/nhst/

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 $\mu_1 \neq \mu_2$
- α: P(type 1 error)
- β: P(type 2 error)

POWER

- Recall: β: P(type 2 error)
- Power: 1 - β
- Probability of rejecting null hypothesis if it is false
- Want more power?
  - Increase N
  - Decrease variance (σ^2)
  - Increase $|\mu_1 - \mu_2|$

EFFECT SIZE

- Small p-value does not imply a large effect!
- Instead, calculate effect size (Cohen's *d*)
- $d = \dfrac{\mu_1 - \mu_2}{s}$
- *s*: pooled standard deviation

Interpretation	d
Very small	0.01
Small	0.02
Medium	0.5
Large	0.8
Very large	1.2
Huge	2