Participants

Research Methods for Human-Centered Computing

Participants

Today's goal:

Explain how to recruit participants for your study

Outline:

- Recruitment
- Power analysis

Recruitment

Population and sampling

Recruitment

"We are testing our system
on our colleagues/students."

-or-

"We posted the study link
on Facebook/Twitter."

Sampling

Are your connections, colleagues, or students **typical** users of your system?

- They may have more knowledge of the field of study
- They may feel more excited about the system
- They may know what the experiment is about
- They probably want to please you

You should sample from your **target population**

An unbiased sample of users of your system

Limiting scope

"We only use data from frequent users."

Limiting scope

What are the consequences of **limiting** your scope?

- You run the risk of catering to that subset of users only
- You cannot make generalizable claims about users

For scientific experiments, the target population may be **unrestricted**

- Especially when your study is more about human nature than about a specific system

Sometimes you want to limit the scope afterwards

- e.g. when participants should not be aware of the scope

WEIRD participants

The problem that, due to our recruitment strategies, most study participants are:

> Western, Educated, and from Industrialized Rich and Democratic countries

Studies cannot be generalized to non-WEIRD participants

> Leads to unfair distribution of the benefits of our science

Most problematic with surveys, less with experiments

> Because of ceteris paribus… unless demographics are a moderator!

Skewed samples

Ideally a sample is randomly drawn from a population, but commonly we use a "convenience sample"

- Like with WEIRD participants, more problematic with surveys, less with experiments

What if your sample is skewed?

- Towards a certain gender, age, education level, region, etc.

Skewed samples

Steps:

- Identify the issue
- Use the skewed parameter as a covariate to test the effect of the skewness on your outcome
- Stratify your sample

Demographics

Always check (and report on) basic demographics!

- Most common: gender, age
- Also common: education/employment, country/state, income
- Common in HCI: familiarity with and use of technology under investigation

Use the demographics as a **covariate** to test their impact

- This is a good practice anyway, but especially when your sample is skewed

Stratification

Covariate approach invalid for small groups / heavy skew

E.g. if a sample has very few people > 45yrs of age, it is difficult to test the effect of age

Solution: stratified sampling

Sample participants evenly from multiple groups (e.g. age groups, regions, etc.)

In your subsequent analyses, you should **re-weigh** your data to match the population distribution

This can also be used directly with skewed samples

Multiple samples

A variant of stratified sampling: collect multiple samples

Example: energy-saving expertise

- Recruit 50% of participants from a message board on energy-saving
- Recruit the remaining 50% from a message board on child rearing
- Tip: Measure expertise regardless!

Multiple samples

Another example: recruiting LGBTQ+, black, cis/straight/ white students:

- Tip: Recruit most difficult group first
- Think about how you want to deal with intersectionality

Multiple samples ≠ manipulation

- Other differences between the samples may exacerbate or limit your hypothesized effect!
- Solution: try to measure these things as much as possible
- You can also try to "match" participants between samples

Study location

Think carefully about where you want the study to take place

In the lab:

- More control over technology and environment (ability to study physical devices)
- Easier to explain complicated study procedures
- Advanced measurement capabilities (e.g. eye tracking)
- Availability of resources (space, computers)
- Recruitment more difficult
- Samples tend to be skewed

Study location

Online:

- Fast recruitment; parallel participation
- Generally cheaper
- No experimenter involvement = less Hawthorne effect
- Usually anonymous
- Limited control (distractions, cheating)
- Limited measurement and manipulation (whatever is possible in a browser)

Study location

Elsewhere (e.g. on location):

- Mostly for qualitative work
- Can be more convenient (instantaneous)
- Observe where the studied practice actually takes place (situated action!)

Study location

Lab studies are generally more expensive

- You have to pay people to show up; may involve parking

"Live" studies are possible online, but generally difficult

- Example: a study where multiple participants interact with each other
- Example: a Wizard-of-Oz study

In all locations, you can recruit difficult-to-reach participants through snowball sampling

- Definitely not random though!

Sources

Qualtrics:

- Expensive but high quality
- Ability to ask for certain requirements
- Easiest way to get a good international sample

Craigslist:

- Post in various cities under Jobs > Etcetera (for a fee)
- US participants only
- Create a geographically balanced sample
- Payment difficult (do a raffle, with an “act quick” incentive)

Sources

Amazon Mechanical Turk:

- Often used for very small tasks, but Turk workers appreciate more elaborate studies
- Anonymous payment facilities (+40% fee)
- High quality, but only if you select US workers with a high reputation
- Increasingly restrictive in what tasks are allowed

Facebook ads:

- Don’t bother

Crowd-sourcing

Demographics tend reflect the general Internet population

- Qualtrics: can be highly tailored upon request
- Craigslist: a bit higher educated and more wealthy
- MTurk: less likely to complain about tedious study procedures, but are also more likely to cheat

Make your study simple and usable

Use quality checks, add an open feedback item to catch unexpected problems

Advertising

State the goal of the study (without revealing too much)

E.g. “test this new recommender system”

Explain the study tasks step by step

Mention average time needed for each step

Mention payment

But also the fact that they’ll be contributing to science

Ask participants to participate only once, and to be careful

Qualitative studies

Recruitment for qualitative studies should be more targeted

- Smaller sample means you want an “even spread” of participants

Determine who you want to interview before you start interviewing

- If multiple types of people are involved in the phenomenon, try to interview all of them!
- E.g. if you are studying the online job market, study both job seekers and recruiters

Qualitative studies

Aim to get a wide variety of participants

- Select people with first-hand experience
- Interview people in a wide variety of contexts
  - E.g. both job seekers and people who recently found a job

Analyze after each 2-3 participants, and let your results guide recruitment

- E.g. if one participants was different in an interesting way, recruit more of those

Power analysis

for user experiments

Power analysis

A means to scientifically decide whether a sample size is sufficient for a certain study

Main takeaway: sample size depends on effect size, p-values, and power

- Let's review those concepts

We'll do a demo using G*Power

- I'll help you with your study in my feedback on your methods outline

A quick review

Is my new system (version B) better than version A?

Experimental hypothesis: H1: Mb > Ma

Calculate the means. Do they differ a lot?

Given no effect, we expect the means to be roughly equal

H0: Mb = Ma

To test H1, we try to **reject H0**

...if the difference between Mb and Ma is so large that H0 is unlikely

P-value

P-value: the likelihood that the effect was due to chance

- given that there is no difference between Mb and Ma

Weighed by the standard error (SE)

- Why? Because if the SE is large, we expect larger differences under H0, but if the SE is small, we expect smaller differences under H0
- If the difference is larger than expected based on the SE, we reject H0 (and thus, H1 is supported)

P-value depends on sample size (Why? SE depends on N!)

Effect size

Effect size: the strength of a result

- difference between Mb and Ma
- does not depend on sample size

Example

Do married men weigh more than single men?

Find 4 married men: $N_m = 4$, $Mean_m = 182$, $SD_m = 15$

Find 4 single men: $N_s = 4$, $Mean_s = 170$, $SD_s = 15$

Effect size: 12 lbs

Is this a large effect? —> Need to standardize it!

Cohen's d = $(Mean_m - Mean_s)$/pooled SD

(182–170)/15 = 0.8… this is indeed a large effect

Is it significant? No! $p = .301$

Example 2

Do married men weigh more than single men?

Find 4000 married men: $N_m = 4000$, $M_m = 177.5$, $SD_m = 15$

Find 4000 single men: $N_s = 4000$, $M_s = 176.5$, $SD_s = 15$

Effect size: 1 lb

Is this a large effect?

(177.5–176.5)/15 = 0.067... this is a very small effect

Is it significant? Yes! $p = .0014$

Reflection

Small studies (N << 100) may find medium or large effects that are not significant

Waste of resources! (unless they are pilot studies)

Large studies (N >> 100) may find very small effects that are significant

Also a waste of resources! (could have done with fewer)

How can we prevent wasting resources?

Do a power analysis!

Power analysis

We reject H0 when $p < .05$

May still be due to chance! (e.g. sample 10 men's heights repeatedly... mean will differ due to random variation)

5% of the time, two samples will be different with $p < .05$, even if they are sampled from the same population!

So, what about the 5% of the times that we reject the null hypothesis, but we got it wrong?

And what about the cases where there is a real effect but we didn't find it?

Getting it wrong

So, what about the 5% of the times that we reject the null hypothesis, but we got it wrong?

> This is a Type I error; 5% is the alpha-level

And what about the cases where there is a real effect but we didn't find it?

> This is a Type II error; we want this error to be smaller than 20%... the beta-level

Alpha and power

	There is a real effect	There is no real effect
Found an effect	**Power**	alpha (false positive)
Found no effect	beta (false negative)	1–alpha (true negative)

Power

1-beta = power

The probability of finding an effect that is really there

How high is our power? Power depends on…

…alpha (if we use p < .01, our power is lower)

…effect size (if the effect is smaller, power is lower)

…N (if we use a larger sample, we increase our power)

Given alpha = 0.05, and a certain expected effect size, how large should our N be to find a true effect 80% of the time?

Power analysis

A calculation involving the following 4 parameters:

- Alpha (cut-off p-value, often .05)
- Power (probability of finding a true effect, often .80 or .85)
- N (sample size, usually the thing we are trying to calculate)
- Effect size (usually the “expected effect”)

Types

A priori: compute N, given other variables

Conducted before you run your study

Post-hoc: compute power, given other variables

Conducted afterwards to find out if you had enough participants to detect the found effect*

Sensitivity: compute effect size, given other variables

Find out the minimum effect size you can detect, given the number of participants

Expected effect

An “educated guess” based on:

- Pilot study results
- Findings from similar studies
- Whatever is considered “meaningful”
- Educated guess

Examples per test

Statistic	Small	Medium	Large
Means - Cohen's d	0.2	0.5	0.8
ANOVA - Cohen's f	0.1	0.25	0.4
ANOVA - eta squared	0.01	0.06	0.14
Regression - f-squared	0.02	0.15	0.35
Correlation - r or point biserial	0.1	0.3	0.5
Correlation - R-squared	0.01	0.06	0.14
Association - 2x2 odds ratio	1.5	3.5	9
Association - w or Phi	0.1	0.3	0.5

G*Power demo

An existing study found that a new TurboTax interface reduced tax filing time from 3.0 hours (SD: 0.5 hours) to 2.7 hours (SD: 0.5 hours).

You created a new interface that you think is even better. How many participants do you need to find an effect that is at least the same size? (assume 85% power)

G*Power demo

You conducted a linear regression testing the effect of number of previous privacy violations on 35 Facebook users' privacy concerns (controlling for age and gender).

The number of previous violations was not significant.

The model without this variable had an R^2 of 0.15.

The model with this variable had an R^2 of 0.30.

What was your power? What sample size should you use to find an effect of this size with 85% power?

G*Power demo

You want to test the combined effect of 6 text sizes and 6 background colors on text readability. You only have money for 150 study participants.

What is the maximum effect size you can find (with 85% power) for a main effects of text size and background color?

What about the interaction effect?

Would it help if you only test 2 sizes and colors?

Final thoughts...

Your Mileage May Vary!

Because power cannot be 100%, there is no guarantee you will find an effect!

The effect in your study might be smaller than in previous work!

Your may need to exclude faulty/outlying participants!

Better to estimate conservatively!

Or check out the graphs to see what would happen...

Tiny samples

Be aware of tiny samples (even when they report significant results)

- Randomization doesn't work well in tiny samples
- Tiny samples fall prey to the "publication bias"
- Due to the "winner's curse", tiny samples overestimate the real effect size

These problems are worst for counter-intuitive results

- Ask your friendly neighborhood Bayesian statistician

A final cool trick...

Let's say you need to collect 150 participants...

Ugh... 3 weeks of my time!

...why not run a quick analysis after the first 50 to see if the results are significant?

That's called "p-hacking", and is not allowed

Why? Because you inflate alpha by "peeking"

But what if you compensate by reducing your alpha?

That's allowed! It's called sequential analysis

A final cool trick...

After 50 participants, you do an analysis

3 options:

- No significance, low effect size (reaction: abandon study)
- Significant result (reaction: stop study, take 2 weeks off)
- No significance, but decent effect size (reaction: continue collecting data)

See http://dx.doi.org/10.1002/ejsp.2023 for more details...