Last Time:

- [ ] reward learning
- [ ] policy learning

lecture 8

HRI, FALL '25

Andrea Bajcsy

This Time:

- [ ] Experimental Design
- [ ] Statistical Analysis

User Studies & Experiments

→ all user experiments are user studies but not all user studies are experiments!

what is a user study or experiment?

↳ **user study** is a broad term for any research conducted to understand users

↳ **user experiment** is a specific type of user study that tests hypotheses via a controlled design, treating users as participants in a structured investigation

ex. interviewing users about their experiences w/ a voice assistant
↳ interview-based user study

vs.

measuring whether AI response time affects perceived trust in the AI system.

! Terms are often used interchangeably in the field, but its important to know the difference.

Why do user studies/experiments?

- validate that a system works as expected
- compare 2+ systems or algorithms
- explore a phenomenon to develop a research Q
- collect training data for a model/algorithm

How to conduct user studies/experiments:

1. Define the research question and hypotheses } experimental design
2. Design a study to address Qs
3. Execute the study
4. Analyze data from the study } statistical analysis
5. Draw conclusions from the analysis

EXPERIMENTAL DESIGN

What is an experiment vs. what is a GOOD experiment?

1. What is an experiment?

"The Arrangement of Field Experiments"

⇒ ORIGINS: - Agriculture: comparing different fertilizers to determine which ⇒ better crops (Fisher, 1926)

- medical Research: testing a drug's effectiveness

⇒ OUR FOCUS: designing experiments to compare robot behaviors / algorithms / policies / models & their effects on human collaboration / perception

⇒ COMPONENTS:

treatments (conditions)	responses (measures)
• drug vs. placebo • random vs. optimal • IRL policy vs. BC policy	• symptom progression • comfort • success rate
experimental units (subjects)	**assignment methods (subject allocation)**
• patients • human users • MDP problems	• random (drug, placebo) • every user sees both • every MDP "sees" both

Let's operationalize these:

- Independent Variables (IV) → Conditions
  - ↳ what you manipulate ex. drug type, motion type, algo.
  - ↳ IV's have levels: - 2 levels: drug vs. no drug
    - 3 levels: 100 mg vs. 200 mg vs. 300 mg

- Dependent Variables (DV) → measures
  - ↳ what you measure ex: symptoms, comfort, task-succ.
  - ↳ DV's can be objective (success rate, time) AND subjective (surveys)

- Population → subjects
  - ↳ how many (i.e. size of population)?
  - ↳ who (e.g. age, gender, education, tech. experience)?

⇒ B/c we work with people, we need to abide by research ethics; i.e. protect participants from physical / mental/emotional harm, violations of privacy & confidentiality, feeling forced to start or continue.

⇒ Every study needs approval from Internal Review Board (IRB)

⇒ Before study starts, we obtain informed consent from all participants.
- tells user about task, risks, benefits, rights
- gets confirmation of voluntary participation

- Assignment
  - ↳ between-subjects: one condition per user
    - → useful if participants seeing multiple levels is a problem (i.e. bias)
    - → good for large participant pools or short study sessions
  - ↳ within-subjects: user experiences all conditions
    - → accounts for interpersonal variability

→ efficiently uses your participant pool

→ susceptible to ordering effects (might want to randomize)

↳ mixed-subjects

⇒ ASK YOURSELF: would seeing multiple conditions be problematic? Is there alot of interpersonal variability? How many participants are available?

- Hypothesis
  - ↳ (weak) IV x affects DV
  - ↳ (stronger) IV x positively affects DV

ex. B/c optimal motion is more predictable, we hypothesize:

H1: optimal robot motions increase user comfort

↳ Good hypotheses:
- make specific predictions (+ will be supported by data or not)
- are measureable ("better" isn't measureable)
- address your research Q + extract a key insight

H: my algo is better than other algo. ← BAD HYPOTHESIS!

↳ think about the IV's! What about your algo is diff. than other algo? How do you quantify "better"? e.g. accuracy?

What is a good experiment?

↳ good experiments are controlled

= experimenter assigns experimental units to treatments.... as opposed to observational

ex. Left handedness (LH)

obs. study: (1980s, 1990s) – 2,000 people who recently died – left handed died 9 yrs younger ⇒ LH die younger

what's wrong? → changing prevalence of reporting left handedness over time

! CONFOUND ! left handedness was condemned in older generations, so majority of people identified as right handed But younger generations naturally show higher proportion of people who are left handed

↳ good experiments avoids confounds

a variable whose effect cannot be distinguished from the effect of the actual IV.

ex. artificial decrease of LH, gender, experience w/robots, algorithm hyper-parameters

⇓ How to avoid confounds?

→ Between subjects: randomized group assignment

ex: if 50 participants and 2 conditions: rand. assign. 25 to condition 1 and 25 to condition 2.

NOTE: randomized ≠ haphazard → want similar populations per condition

→ Within subjects: counter-balance the condition

Order 1: Alg. A, Alg. B

Order 2: Alg. B, Alg. A

N conditions ⇒ N! orderings

→ pre-study practice: unrecorded familiarization stage

↳ good experiments are reliable

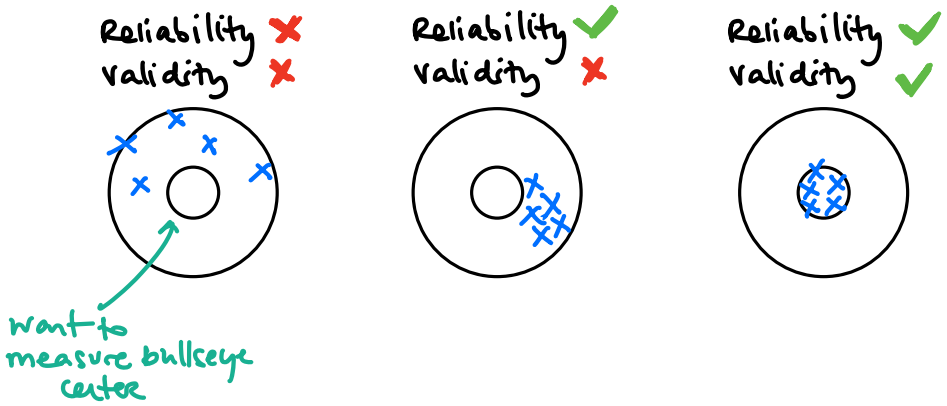
= low experimental error (low variance) → repeating produces similar outcomes

↳ good experiments have **construct validity**

= the measures actually measure what you want

ex. IQ test → intelligence?
rating of predictability → predictability?

Q How to improve? Have objective + subjective measures



↳ good experiments have **external validity**

= conclusions drawn from exp. apply to real world setting

Q How to improve? Sample from target pop., apply insights across problems / algos / etc.

↳ good experiments are **factorial**

= conditions are all combinations of all levels of IVs

ex. IV 1: Value Iter vs. Q-learning } 2x2 Design
IV 2: Sparse Rew. vs. Dense

	VI	Q
Sparse	C1	C2
Dense	C3	C4

common pitfall: 2 changes @ once

↳ w/o factorial design, we may only test Q + Dense vs. VI + sparse and then have a confound!

STATISTICAL ANALYSIS

Q "Has the DV changed as a result of manipulating IV?"

ex. 1 IV (alg), 2 levels (alg 0, alg 1), within subjects (2 MDP Problems)

	MDP Problem	level (IV)	Total Reward (DV)
"participant 1"	1	0	10
	1	1	8
"participant 2"	2	0	11
	2	1	7

how participant behaved under 0 / how they behaved under 1

MDP	Diff Reward
1	2
2	4

→ for each participant, calculate diff. btwn their pre & post intervention score

t-test: what's the probability 2 populations are different from each other?

"null hypothesis" a statement of no effect

$H_0: \mu_1 = \mu_2$ (the population means are equal)

$H_1: \mu_1 \neq \mu_2$ (the population means are not equal)

Paired t-test:

$$t = \frac{\bar{X}_{diff}}{\bar{S}_{diff}/\sqrt{n}}$$

$\bar{X}_{diff}$: mean of sample differences

$\bar{S}_{diff}$: stdev. of the sample differences

n: sample size

ex. $n=2$, $\bar{X}_{diff} = \frac{2+4}{2} = 3$, $\bar{S}_{diff} = \sqrt{\frac{(2-3)^2+(4-3)^2}{2}} = 1$

$$t = \frac{3}{1/\sqrt{2}} = 4.2426$$

$df = n - 1 = 1$

"degrees of freedom"

You use the combo of (t, df) to obtain a p-value (via a lookup table, or statistical software)

p-value: probability of obtaining a result at least as extreme as the results actually observed, assuming the null hypothesis is true.

→ Intuitively: determines if observed result is likely due to chance or a real effect

→ small p-value (≤ 0.05) ⇒ observed results are unlikely to have occurred by chance; provides enough evidence to reject H_0

→ large p-value (> 0.05) ⇒ observed results are likely due to chance; can't reject H_0

↳ if $p = 0.02$ then there is 2% chance of observing results as extreme as the ones obtained, if H_0 is true.

! a small p-value does not prove alternative hypothesis is true, just suggests H_0 is unlikely

↳ in statistics, we aim to falsify the null hypothesis, not to prove the alternative

$\bar{X}$ larger ⇒ more confident
N larger ⇒ more confident
$\bar{s}$ lager ⇒ less confident

→ if DV is binary / categorical rather than continuous:
(T/F) (yes/no/maybe) (reward & IR)

use a chi-squared (χ^2) test instead of t-test

OTHER USEFUL SCENARIOS:

• 1 IV, 2 levels, between subjects ← here, comparing 2 groups where the data is not paired (but in paired t-test it is)

independent t-test → $t = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\bar{s}_1^2}{N_1} + \dfrac{\bar{s}_2^2}{N_2}}}$

- 2 IVs, 2 levels each, between subjects

(Alg, Robot)
↙ ↘ ↙ ↘
BC IRL Arm Vehicle

G1: BC + Arm
G2: BC + Vehicle
G3: IRL + Arm
G4: IRL + Vehicle

If we decided to run t-tests, for all pairwise comparisons of $k=4$ groups, we need:

$$\#\text{ comparisons} = \frac{k(k-1)}{2} = \frac{4(3)}{2} = 6$$

! should we do this? 6 t-tests? No!

Increases Type I error rate (i.e. false positives)

ex: one t-test has a 5% chance ($p = 0.05$) of falsely rejecting H_0 (i.e. detecting significance).

If you do 100 t-tests:

$P(\text{error}_1 \text{ OR } \text{error}_2 \dots \text{OR } \text{error}_{100}) = 1 - P(\text{correct}_1 \text{ AND } \dots \text{ correct}_{100})$

"P(make ≥ 1 error)"

each test indep. of next $= \prod_i P(\text{correct}_i)$

$= 1 - 0.95^{100} = 0.9941$!

! 99.41% chance you falsely reject H_0

Some Solutions:

(A) Bonferoni correction: adjust significance level by # of comparisons

if $\alpha = 0.05$ then now $\bar{\alpha} = \alpha/m$, $m = \#$ comparisons

ex: $m = 6$ from example above, $\bar{\alpha} = 0.008\bar{3}$

! very conservative (↑ increased risk of Type II error i.e. false negatives)

(B) Tukey's HSD (Honestly Significant Differences): apply a single test to compare all groups @ once.

For Factorial Designs or IV with >2 levels

ANOVA (Analysis of variance)

⊛ always run an ANOVA first as a precursor to making multiple comparisons

↳ it tells you if there is a difference, but not where the difference is

- main effect: the effect of each IV individually
- interaction effect: how IVs combine to influence DVs

If there is an interaction effect or the main effect is not clear, then run a "post-hoc" test (e.g. Tukey's HSD) to see which groups differ.

- Use a one-tailed ANOVA to test diff. in a specific direction (ex. method A leads to higher user efficiency than method B)

- Use a two-tailed ANOVA to test for any significant difference btwn. group means, whether its increase or decrease (ex. there is a diff. btwn method A and B)