Experimental design

- WHAT is an experiment
- what is a GOOD experiment
  - identify insight, test without confounds

1. What is an experiment?

1.1. Origins: medical (drug effectiveness)
agriculture (pesticide effectiveness)
- Fisher (1926)

then: psychology (understand people)
focus: HRI (test design decisions)
ROBOTICS (compare algorithms, test insights)

1.2. Components:

treatments (conditions)	responses (measures)
• drug vs placebo • random vs optimal • CHOMP vs GS CHOMP	• symptom progression • comfort • cost
experimental units (subjects)	**assignment method (subj. allocation)**
• patients • users • MP Problems	• random (drug, placebo) • every user sees both • every problem "sees" both

1.3. Assignment

- Between-subjects: one condition per user
  no biases

- within-subjects: all conditions
  eliminate inter-user variation
  cases where within not possible
  - give patients both drugs ✗
  - not realistic: robot collaboration / coordination
    ↳ once you finish the task, you know it
- Mixed

1.4. Operationalization of Variables
- Independent variables → conditions
  what you manipulate e.g. drug type, motion type
- Dependent Variables = measures
  what you measure e.g. symptoms, comfort

1.5. Hypothesis
- IV x affects DV y
  ↳ motivate by a mechanism or analogous studies
- Better: IV x <u>positively</u> affects DV y.

Because optimal motion is more predictable, we hypothesize that:

H1. Optimizing motion increases user comfort.

Because goal sets add flexibility, we hypothesize that:

H1. Including goal sets in optimization reduces the final trajectory cost.

Hypotheses extract key insights.

H. My alg > other alg.
bad hypothesis

↳ think about the IV!
Science not just engineering

2. What is a good experiment?

2.1. Good experiments are controlled.

"controlled" = experimenter assigns experimental units to treatments

as opposed to observational

e.g. How does estrogen treatment affect health outcomes?

Obs study:

- 93676 women
- 8 years
- estrogen yes/no
- Heart Health (HH)

⇒ estrogen positively related to HH.

What's wrong?

! CONFOUND ! health consciousness

controlled study: estrogen negatively affects HH

eg 2. left handedness

Obs study:

- 2000 people who died ? – [illegible]

LH died 9 years younger } ⇒ LH die younger
what's wrong?

! CONFOUND ! LH condemned: younger people equal amounts, but older people are right-handed

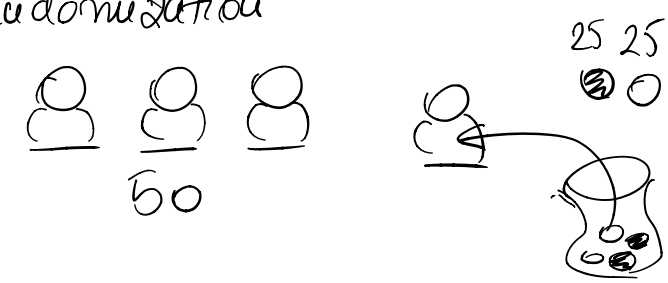
2.2. Good experiments avoid confounds

"confound": a variable whose effect can not be distinguished from the effect of the IV.

e.g. health conciousness
artificial decrease of LH
gender
experience with robots
trajectory execution time
algorithm metaparameters? eg optimizer step size

Tools for avoiding confounds

- Randomization



Note: haphazard ≠ randomized
eg alternate ABABAB
↳ all A in morning, all B in evening } → diff population

→ similar population for each condition

- within subjects: counter-balance the order
  D P N conditions → N! orderings

↳ "latin square"

- stack the cards against yourself
  - raud – path length
  - CHOMP – squared velocity

  DM: path length

- optimize the baseline alg
  - STOMP – step size
  - RRT* – improve RRT as well in all other ways
    but < parent selection, rewiring
    ↳ test the key insight

2.2. Good experiments are reliable.

"reliability": low experimental error (low variance)

Tools for improving reliability

- within-subjects – exact same user, no eg Likert bias
- "Blocking"; block – group of homogenous experimental units; blocking – arrangement of units into blocks; reduces known but irrelevant sources of variation between conditions

$$\underset{\text{minimize}}{\mathrm{Var}(x-y)} = \mathrm{Var}(x) + \mathrm{Var}(y) - 2\,\underset{\text{maximize}}{\mathrm{cov}(x,y)}$$

blocking factor
covariates - secondary variable that can affect the DM

- Multi-item scales

predictable expected surprising

2.3. Good experiments have construct validity.

"construct validity": the measures actually measure what you want

IQ test → intelligence?

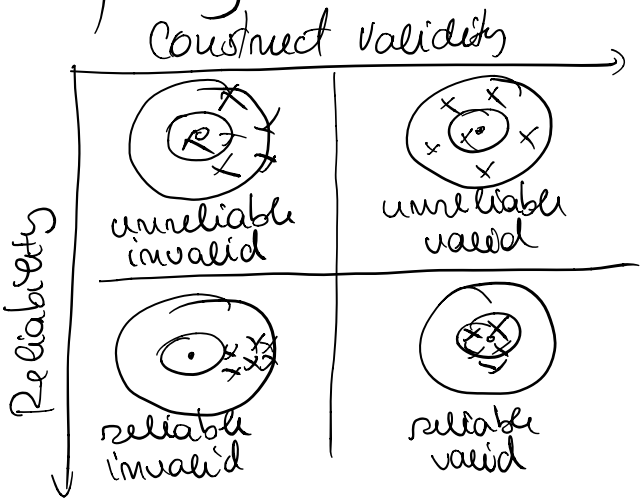
rating of predictability → predictability?

Tools:

- objective + subjective measures

pred
expected
surprising

choose correct motion



2.4. Good experiments have external validity.

"external validity": conclusions

Tools:

- Sample from the target population.

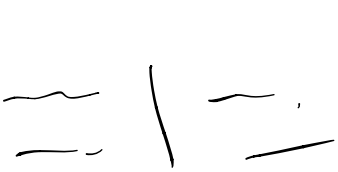
Apply your insight across problems, problem types, even algorithms.

Goal Sets

NO	CHOMP	TrajOpt	STOMP
YES	GSCHOMP	GSTrajopt	GS.STOMP

Representation

WAYPOINTS	CHOMP-WP	
RKHS	CHOMP-RKHS	



25. Good experiments are factorial !

"factorial" : conditions = all combinations of all levels of the IVs

GS

RKHS	0	1
0	✓	✓
1	✓	✓

common pitfall : 2 changes at once

CHOMP vs TrajOPT : 2nd order, treatment of obstacles

1st	vs	2nd	soft	hard
soft, SDF	CHOMP	2nd order CHOMP	point-SDF	
hard, geom	hard constr. geom CHOMP	TrajOPT	geom-geom	

CHOMP-WP vs CHOMP-RKHS vs Trajopt-RKHS
↳ what if RKHS helps CHOMP but not TrajOpt

Factorials break down the changes and extract what actually matters

<u>Recap</u>

treatment	response
exp units	assignment

HR1: manipulated variables
dependent measures
subj. allocation
controlled, not confounded, reliable, valid,
factorial
construct external

3. Analysis (starting point)

3.1. 1 IV, 2 levels, within subjects

MP Problem	Level	Cost	MP Problem	Δ Cost
1	0	10	1	2
1	1	8	2	4
2	0	11		
2	1	7		

t-test

$$t = \frac{\bar{x}}{\bar{s}/\sqrt{N}}$$

$\bar{x}$ — sample mean

$\bar{s}$ — sample SD

N — sample size

$(t, N) \Rightarrow$ p value: probability of obtaining a result equal to or more extreme than what was actually observed when the null hypothesis is true

artificial threshold for significance: $\leq .05$

$\bar{x}$ larger – more confident

N larger – more confident

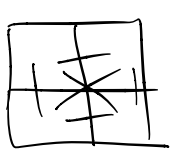
$\bar{s}$ larger – less confident

if DV is binary/categorical: chi-squared χ^2

3.2. 1 IV, 2 levels, between-subjects

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\bar{s}_1^2}{N_1}} + \sqrt{\frac{\bar{s}_2^2}{N_2}}}$$

3.3. 2 IV, 2 levels each, between-subjects



6 t-tests? <u>NO</u> "multiple comparisons problem"

each comparison: P(error) = .05

P(≥1 errors in 100 comparisons) =

1 − P(always correct)

= (as indep.) $1 - .95^{100} = .99411$!

conservative correction: Bonferroni: $\alpha \leftarrow \alpha / n$

comparisons

better: Tuckey HSD

Factorial designs OR IV with >2 levels

ANOVA - Analysis of Variance

precursor to multiple comparisons

- main effects IV1 IV2
- interaction effects IV1 × IV2

$a = \begin{cases} 1 & \text{if } IV1 = 1 \\ 0 \end{cases}$ $b = \begin{cases} 1 & \text{if } IV2 = 1 \\ 0 \end{cases}$

response $= \beta_1 a + \beta_2 b + \beta_3 ab + \beta_4$

- add user / problem # for within subjects

useful in robotics when - multiple factors
- 1 factor multiple levels

JMP MINITAB