Experimental Design

- **What is an experiment?**
- **What is a good experiment?**
  - Identify insight and test without confounds.

1. **What is an experiment?**

1.1. **Origins:**
   - Medical (drug effectiveness)
   - Agriculture (pesticide effectiveness)
   - Psychology (understand people)
   - Games: TR1 (test design decisions)
   - Robotics (compare algorithms)
   - Test insights

1.2. **Components:**

<table>
<thead>
<tr>
<th>Treatments (conditions)</th>
<th>Responses (measures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug vs placebo</td>
<td>symptom progression</td>
</tr>
<tr>
<td>random vs optimal</td>
<td>cost</td>
</tr>
<tr>
<td>CHOMP vs GS CHOMP</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experimental units (subjects)</th>
<th>Assignment method (Subj. allocation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>patients</td>
<td>random (drug, placebo)</td>
</tr>
<tr>
<td>users</td>
<td>every user sees both</td>
</tr>
<tr>
<td>MP Problems</td>
<td>every problem &quot;sees&quot; both</td>
</tr>
</tbody>
</table>

1.3. **Assignment:**

- Between-subjects: one condition per user, no biases.
Within-subjects: all conditions

- cases where within not possible

- give patient 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. symptom, comfort

1.5. Hypothesis

- IV \times DV y affects DV y
  - motivate by a mechanism n analogous studies

- Better: IV \times positively 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.

- **My alg**
- **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:
  - 93,676 women
  - 8 years
  - estrogen yes/no
  - Heart Health (HT)

  \[ \rightarrow \text{estrogen positively related to HT} \]

  ✗ **CONFOUND?** Health consciousness
  - controlled study: estrogen negatively affects HT

  ✗ 2. left-handedness

  Obs study:
  - 2000 people who died
  - 1/11 die younger
LH died 9 years younger does younger mean 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 cannot be distinguished from the effect of the IV.

E.g., health consciousness, artificial decrease of LH level, experience with robots, trajectory execution time, algorithm metaparameters? e.g., optimize some step size.

Tools for avoiding confounds:

- Randomization

![Randomization Diagram]

Note: Randomized ≠ Randomized

E.g., alternate ABABAB

\( \text{all A in morning} \rightarrow \text{all B in evening} \) → all population

→ Similar population in each condition.

- Within subjects: counter-balance the order

E.g., D P N conditions \( \rightarrow N! \) orderings.
o. stack the cards against yourself
  rand - path length
  CHOMP - squared velocity

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

2.2. Good experiments are reliable.
  "reliability" : low experimental error (low variance)
  tools for improving reliability
  o. within-subjects - exact same user, no eg diet Bias
  o. "Blocking" ; block-group of homogenous experimental units; blocking = arrangement of units
     into blocks; reduces known but irrelevant sources of variation between conditions

\[ \text{Var}(x-y) = \text{Var}(x) + \text{Var}(y) - 2 \text{cov}(x,y) \]

minimize
blocking factor
covariates - secondary variable that can affect HDM
- 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
  - 
  - 
  - 
  - 

<table>
<thead>
<tr>
<th>Construct Validity</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unreliable invalid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unreliable valid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliable invalid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliable valid</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>NO</th>
<th>CHOMP</th>
<th>TrajOpt</th>
<th>STOMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>CHOMP</td>
<td>TrajOpt</td>
<td>STOMP</td>
</tr>
</tbody>
</table>

Representation

<table>
<thead>
<tr>
<th>WAYPOINTS</th>
<th>CHOMP-NO</th>
<th>CHOMP-YE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RKHS</td>
<td>CHOMP-RKHS</td>
<td>CHOMP-RKHS</td>
</tr>
</tbody>
</table>

2.5. Good experiments are factorial?

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

Common pitfall: 2 changes at once

CHOMP vs TrajOpt: 2\\textsuperscript{nd} order, treatment of obstacles

1\\textsuperscript{st} vs 2\\textsuperscript{nd} order

Soft, hard

TrajOpt, CHOMP

2\\textsuperscript{nd} order CHOMP, point-SDF

Soft, hard

TrajOpt, CHOMP
CHOMP - mp vs CHOMP - RKTS vs Trajopt - RKTS

What if RKTS helps CHOMP but not Trajopt

Factorials break down the changes and extract what actually matters

Recap

<table>
<thead>
<tr>
<th>treatment</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp units</td>
<td>assignment</td>
</tr>
</tbody>
</table>

H21: manipulated variables
dependent measures
subj. allocation
controlled, not confounded, reliable, valid

3. Analysis (starting point)

3.1. 1 IV, 2 levels, within subjects

<table>
<thead>
<tr>
<th>MP</th>
<th>Problem</th>
<th>Level</th>
<th>Cost</th>
<th>Δ Cost</th>
<th>MP</th>
<th>Problem</th>
<th>Δ Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[ t\text{-test} \quad t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}} \]

\((t, N) = p\text{-value}\) probability of obtaining a result equal to or more extreme than what was actually observed when the null hypothesis is true.

The artificial threshold for significance: \(\leq 0.05\)

- \(\bar{x}\) larger: more confident
- \(N\) larger: more confident
- \(s\) larger: less confident
- If \(\text{DV}\) is binary/categorical: chi-squared \(\chi^2\)

3.2. \(1\text{ IV}, 2\text{ levels, between-subjects}\)
\[ t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2_1}{N_1} + \frac{s^2_2}{N_2}}} \]

3.3. \(2\text{ IV}, 2\text{ levels each, between-subjects}\)

![6+tests?](image)

No "multiple comparison problem"

Each comparison: \(P\text{-error}) = 0.05 \)

\(P(C > 1\text{ errors in } 100\text{ comparisons}) = 1 - P(C\text{ always correct}) \)
\[ (1.96, 0.95^{100} = 0.9841 \] for conservative correction: Bonferroni: \( \alpha < \alpha / m \)

better: Tuckey HSD

Factorial designs or IV with 2 levels

ANOVA - Analysis of Variance

- main effects \( IV_1, IV_2 \)
- interaction effects \( IV_1 \times IV_2 \)

\[ a = \begin{cases} 1 & \text{if } IV_1 = 1 \\ 0 & \text{otherwise} \end{cases} \quad b = \begin{cases} 1 & \text{if } IV_2 = 1 \\ 0 & \text{otherwise} \end{cases} \]

response = \( \beta_0 + \beta_1 a + \beta_2 b + \beta_3 ab + \beta_4 \)

- add user problem # for within subjects

useful in robotics when - multiple jack

- (fast multiple levels)

JMP Minitab