Introduction to HCI
Fall 2021

Evaluation
Designing Controlled Experiments

Mahmood Jasim
UMass Amherst
mjasim@cs.umass.edu
https://people.cs.umass.edu/~mjasim/

Logistics

- Milestone 3 posted
- Feedback survey posted
  - https://forms.gle/H1jhnWWNG4DK2P837

Learning goals

- What is the experimental method?
- What is an experimental hypothesis?
- How do I plan an experiment?
- Why are statistics used?
- Within & between-subject comparisons: how do they differ?

Acknowledgement: Some of the material in this lecture is based on material prepared for similar courses by Saul Greenberg (University of Calgary)

Controlled experiments

The traditional scientific method

- Clear convincing result on specific issues
- In HCI
  - Insights into cognitive process, human performance limitations, ...
  - Allows comparison of systems, fine-tuning of details ...

Strives for

- Lucid and testable hypothesis
- Quantitative measurement
- Measure of confidence in results obtained
- Replicability of experiment
- Control of variables and conditions
- Removal of experimenter bias

Desired outcome of a controlled experiment

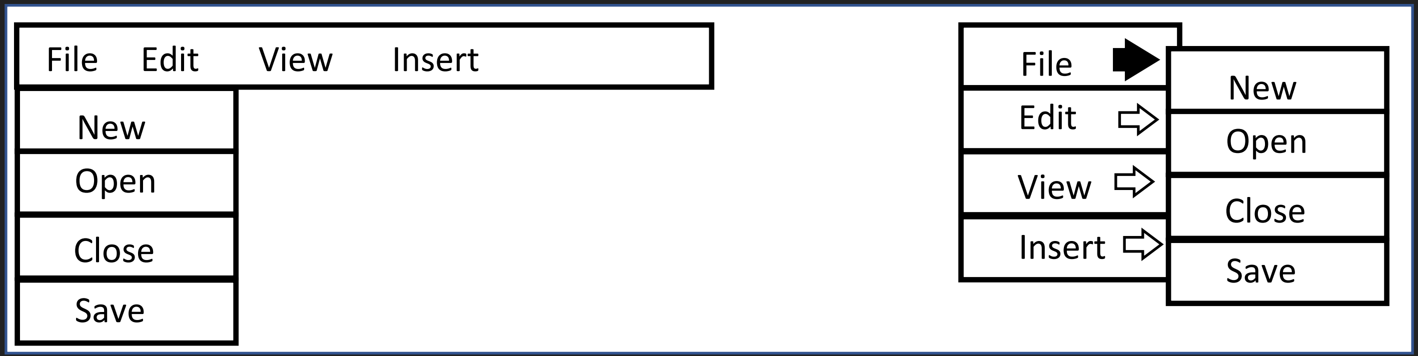
- Statistical inference of an event or situation's probability:
- "Design A is better <in some specific sense> than design B"
- Or, design A meets a target:
- "90% of incoming students who have web experience can complete course registration within 30 minutes"

Summary of steps

- Step 1: begin with a testable hypothesis
- Step 2: explicitly state the independent variables
- Step 3: carefully choose the dependent variables
- step 4: consider possible nuisance variables & determine mitigation approach
- Step 5: design the task to be performed
- Step 6: design experiment protocol
- Step 7: make formal experiment design explicit
- Step 8: carefully select/recruit and assign subjects to groups
- Step 9: apply statistical methods to data analysis
- Step 10: interpret your results

Step 1: Begin with a testable hypothesis

- Example 1:
- Null Hypotheses - H0: there is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu
- Alternate Hypotheses - H1: selecting from a pop-up menu will be faster and less error prone than selecting from a pull down menu



General: Hypothesis testing

- Hypothesis = prediction of the outcome of an experiment.
- Framed in terms of independent and dependent variables:
  - A variation in the independent variable will cause a difference in the dependent variable
- Aim of the experiment: prove this prediction
  - By: disproving the "null hypothesis"
  - Never by: proving the "alternate hypothesis"

- H0: experimental conditions have no effect on performance (to some degree of significance) → null hypothesis
- H1: experimental conditions have an effect on performance (to some degree of significance) → alternate hypothesis

Step 2: Explicitly state the independent variables

Independent variables

- things you control/manipulate (independent of how a subject behaves) to produce different conditions for comparison

- e.g., age and time

Step 3: Carefully choose the dependent variables

Dependent variables

- Things that are measured
- Expectation that they depend on the subject's behavior / reaction to the independent variable (but unaffected by other factors)

- e.g. height

Step 4: Consider possible nuisance variables & determine mitigation approach

- Undesired variations in experiment conditions which cannot be eliminated, but which may affect dependent variable
  - Critical to know about them

- Experiment design & analysis must generally accommodate them:
  - Treat as an additional experiment independent variable (if they can be controlled)
  - Randomization (if they cannot be controlled)

- Common nuisance variable: subject (individual differences)

Step 5: Design the task to be performed

- Be externally valid
  - External validity = do the results generalize?
  - Will they be an accurate predictor of how well users can perform tasks as they would in real life?
- Exercise the designs, bringing out any differences in their support for the task
  - E.g., If a design supports website navigation, test task should not require subject to work within a single page
- Be feasible - supported by the design/prototype, and executable within experiment time scale

Step 6: Design experiment protocol

- Steps for executing experiment are prepared well ahead of time
- Includes unbiased instructions + instruments (questionnaire, interview script, observation sheet)
- Double-blind experiments, …

Step 7: Make formal experiment design explicit

- Simplest: 2-sample (2-condition) experiment

- Based on comparison of two sample means:
  - Performance data from using design A & design B
    - e.g., New design & status quo design
    - e.g., 2 new designs

- Or, comparison of one sample mean with a constant:
  - Performance data from using design A, compared to performance requirement
    - Determine whether single new design meets key design requirement

Step 7: Make formal experiment design explicit

- More complex: factorial design

- In menu experiment:
  - 2 menu types (pop-up, pull down)
  - X 5 menu lengths (3, 6, 9, 12, 15)
  - X 2 levels of expertise (novice, expert)

Within/between subject comparisons

Within-subject design:

- subjects exposed to multiple treatment conditions
- primary comparison internal to each subject
- allows control over subject variable
- greater statistical power, fewer subjects required
- not always possible (exposure to one condition might "contaminate" subject for another condition; or session too long)

Within/between subject comparisons

- Between-subject design:
  - Subjects only exposed to one condition
  - Primary comparison is from subject to subject
  - Less statistical power, more subjects required
  - Why? Because greater variability due to more individual differences

Step 8: Carefully select/recruit and assign subjects to groups

Subject pool: similar issues as for informal and field studies

- Match expected user population as closely as possible
- Age, physical attributes, level of education
- General experience with systems similar to those being tested
- Experience and knowledge of task domain

Sample size: more critical in experiments than other studies

- Going for “statistical significance”
- Should be large enough to be “representative” of population
- Guidelines exist based on statistical methods used & required significance of results
- Pragmatic concerns may dictate actual numbers
- “10” is often a good place to start

Step 8: Carefully select/recruit and assign subjects to groups

- If there is too much variability in the data collected, you will not be able to achieve statistical significance
- You can reduce variability by controlling subject variability
  - Recognize classes and make them an independent variable
    - e.g., Older users vs. Younger users
    - e.g., Superstars versus poor performers
  - Use reasonable number of subjects and random assignment

Step 9: Apply statistical methods to data analysis

- Examples: t-tests, ANOVA, correlation, regression

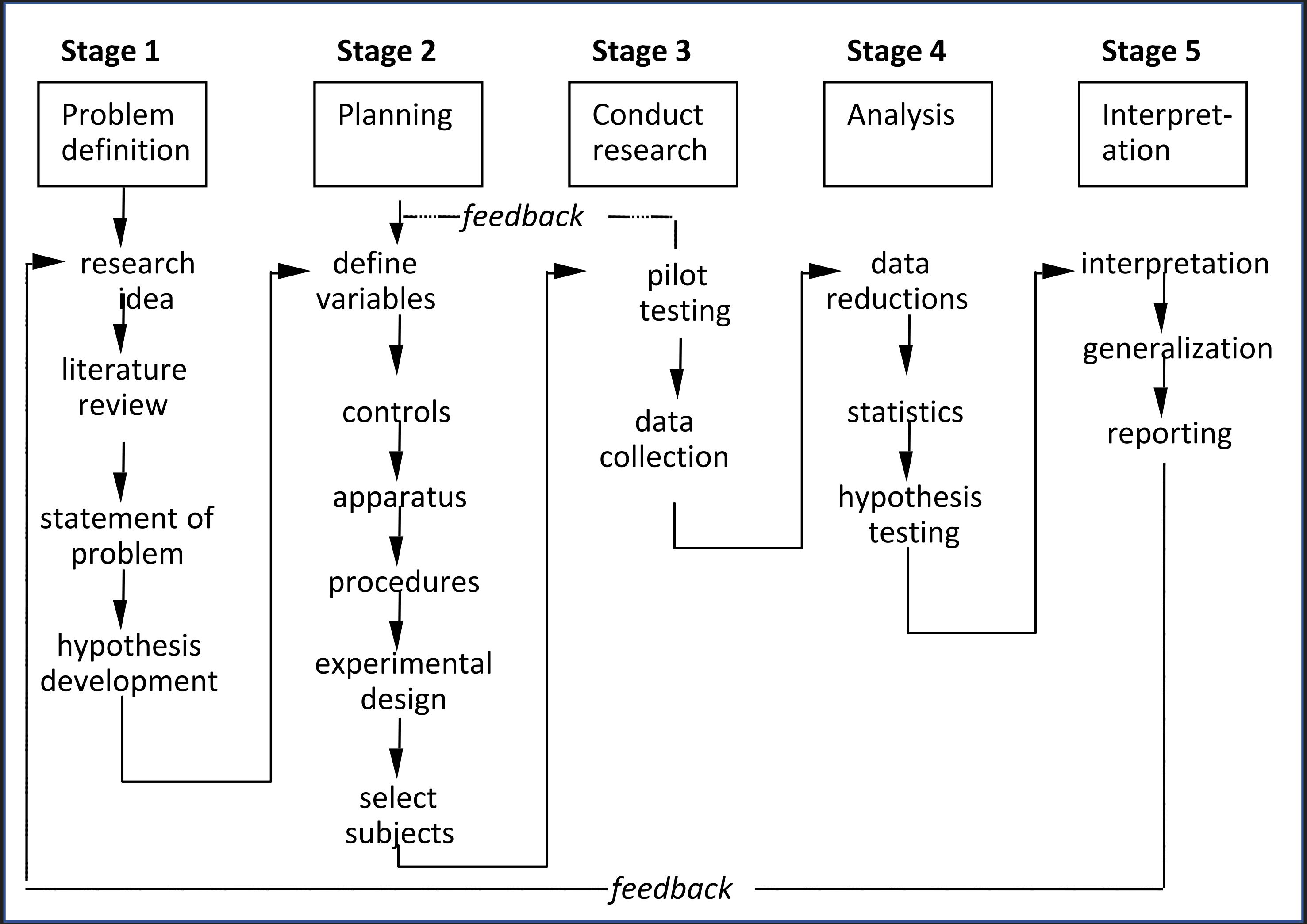
- Confidence limits: the confidence that your conclusion is correct
  - "The hypothesis that mouse experience makes no difference is rejected at the .05 level" (i.e., Null hypothesis rejected)
  - This means:
    - A 95% chance that your finding is correct
    - A 5% chance you are wrong

Step 10: Interpret your results

- What you believe the results mean, and their implications

- Yes, there can be a subjective component to quantitative analysis

The experiment planning flowchart



To summarize: how a controlled experiment works

- Formulate an alternate and a null hypothesis:
    - H1: experimental conditions have an effect on performance
    - H0: experimental conditions have no effect on performance
- Through experimental task, try to demonstrate that the null hypothesis is false (reject it),
    - For a particular level of significance
- If successful, we can accept the alternate hypothesis,
    - And state the probability p that we are wrong (the null hypothesis is true after all) → this is result's confidence level
    - e.g., Selection speed is significantly faster in menus of length 5 than of length 10 ($p<.05$)

5% chance we've made a mistake, 95% confident

In-class activity

- Work in groups
- Write down names of participating group members
- Design experiments for your own projects
- Focus on
  - Hypothesis
  - Independent and Dependent Variables
  - Tasks
- Link to worksheet - https://tinyurl.com/94kest79

Optional reading

- Research Methods in Human-Computer Interaction, 2nd edition. Jonathan Lazar, Jinjuan Heidi Feng, Harry Hochheiser.
  - Chapter 3 – Experimental Design
  - https://learning.oreilly.com/library/view/research-methods-in/9780128093436/?sso_link=yes&sso_link_from=UMassAmherst