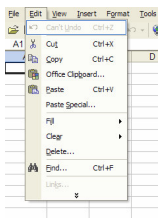
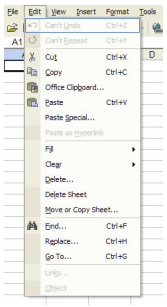
Lecture 20: Experiment Design

UI Hall of Fame or Shame?





UI Hall of Fame or Shame?

Today's Topics

- Experiment design

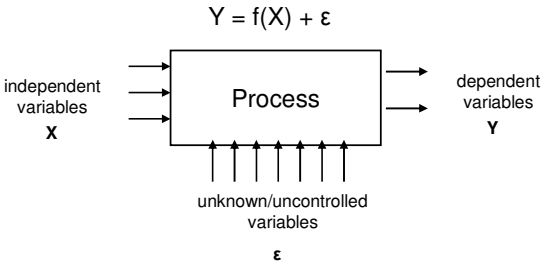
Controlled Experiment

- Start with a testable **hypothesis**
  - e.g. Mac menu bar is faster than Windows menu bar
- Manipulate **independent variables**
  - different interfaces, user classes, tasks
  - in this case, y-position of menubar
- Measure **dependent variables**
  - times, errors, satisfaction
- Use statistical tests to accept or reject the hypothesis

Design of the Menubar Experiment

- Users
  - Windows users or Mac users?
  - Age, handedness?
  - How to sample them?
  - Within- or between-subjects?
- Implementation
  - Real Windows vs. real Mac
  - Artificial window manager that lets us control menu bar position
- Tasks
  - Realistic: word processing, email, web browsing
  - Artificial: repeatedly pointing at fake menu bar
- Measurement
  - When does movement start and end?
- Ordering
  - of tasks and interface conditions
- Hardware
  - mouse, trackball, touchpad, joystick?
  - PC or Mac? which particular machine?

Schematic View of Experiment Design



Concerns Driving Experiment Design

- Internal validity
  - Are observed results actually **caused** by the independent variables?
- External validity
  - Can observed results be **generalized** to the world outside the lab?
- Reliability
  - Will consistent results be obtained by **repeating** the experiment?

Threats to Internal Validity

- Ordering effects
  - People learn, and people get tired
  - Don't present tasks or interfaces in same order for all users
  - Randomize or counterbalance the ordering
- Selection effects
  - Don't use pre-existing groups (unless group is an independent variable)
  - Randomly assign users to independent variables
- Experimenter bias
  - Experimenter may be enthusiastic about interface X but not Y
  - Give training and briefings on paper, not in person
  - Provide equivalent training for every interface
  - Double-blind experiments prevent both subject and experimenter from knowing if it's condition X or Y
    - Essential if measurement of dependent variables requires judgement

Threats to External Validity

- Population
  - Draw a random sample from your real target population
- Ecological
  - Make lab conditions as realistic as possible in important respects
- Training
  - Training should mimic how real interface would be encountered and learned
- Task
  - Base your tasks on task analysis

Threats to Reliability

- Uncontrolled variation
  - Previous experience
    - Novices and experts: separate into different classes, or use only one class
  - User differences
    - Fastest users are **10 times** faster than slowest users
  - Task design
    - Do tasks measure what you're trying to measure?
  - Measurement error
    - Time on task may include coughing, scratching, distractions
- Solutions
  - Eliminate uncontrolled variation
    - Select users for certain experience (or lack thereof)
    - Give all users the same training
    - Measure dependent variables precisely
  - Repetition
    - Many users, many trials
    - Standard deviation of the mean shrinks like the square root of N (i.e., quadrupling users makes the mean twice as accurate)

Blocking

- Divide samples into subsets which are more homogeneous than the whole set
  - Lots of variation between feet of different kids
  - But the feet on the same kid are far more homogeneous
  - Each child is a block
- Apply all conditions within each block
  - Put material A on one foot, material B on the other
- Measure difference within block
  - Wear(A) − Wear(B)
- Randomize within the block to eliminate internal validity threats
  - Randomly put A on left foot or right foot

Between Subjects vs. Within Subjects

- "Between subjects" design
  - Users are divided into two groups:
    - One group sees only interface X
    - Other group sees only interface Y
  - Results are compared **between** different groups
    - Is mean(xi) > mean(yj)?
  - Eliminates variation due to ordering effects
    - User can't learn from one interface to do better on the other
- "Within subjects" design
  - Each user sees both interface X and Y (in random order)
  - Results are compared **within** each user
    - For user i, compute the difference xi-yi
    - Is mean(xi-yi) > 0?
  - Eliminates variation due to user differences
    - User only compared with self