CS 352
Introduction to Usability Engineering

Empirical Evaluation with Users

Part II

Think-Aloud Research Studies

- Analyze data for patterns, surprises, etc.
- No stats: not enough subjects for this
- Sample think-aloud results: from VL/HCC'03 (Prabhakararao et al.)

Example

People's Strategies with a spreadsheet debug feature

- Research questions:
  - RQ1: Perceived value?
  - RQ3; What debugging strategies used?
  - RQ4- Influence of feature on their strategies?
- Sample results:
  - RQ1: 4/5 users used it at least once. (#3 forgot it, but wished he had remembered)
  - RQ3&4: Dataflow debugging was a success and the feature encouraged it

Statistical Studies

- We will not use them, but …
  - You need to know the basics
- Goal: answer a binary question
  - eg. Does system X help users create animations?
  - eg. Are people better debuggers using X than Y?
- Advantage: your audience believes it
- Disadvantage: you may not find out about "why or why not?"

Hypothesis Testing

- Need to be specific and provable/refutable
  - e.g. “users will debug better using X than Y”
  - Strictly speaking we use the “null” hypothesis, which says there won’t be a difference
  - Pick a significance value (rule of thumb is 0.05)
    - If you get a p-value <=0.05 this says you’ve shown a significant difference, but there’s a 5% chance the difference is a fluke

Design the Experiment

- Identify outputs (dependent variables) for the hypotheses:
  - eg, more bugs fixed?
  - eg. Fewer minutes to fix the same number of bugs
- Identify independent variables we'll manipulate (treatments):
  - Which system used, X or Y?

Design the Experiment (cont)

- Decide on within vs. between subject
  - “Within”: 1 group experiments all treatments
    - In random order
    - “within is best, if possible (Why?)
- How many subjects?
- Rule of thumbe: 30/treatment
- More subjects -> more statistical power -> more likely to get p<=0.5 if there really is a difference

Design the Experiment (cont)

- Design the task they will do
  - Since you usually run a lot of these at one time and you're comparing them, you need to be careful with length
    - Long enough to get over the learning curve
    - Big enough to be convincing
    - Small enough to be do-able in the amount of time subjects have available
  - Vary the order if multiple tasks

Design the Experiment (cont)

- Develop the tutorial
  - Practice like crasy! (must work the same for everyone!)
- Plan the data to gather
  - Log files?
  - Questionnaires before/after?
  - Saved results files at end?

Design the Experiment (cont)

- Water in the beer:
  - Sources of uncontrolled variation spoil results
- Sources
  - Too much variation in subject background
  - Not good enough tutorial
  - Task not a good match for goal
  - and so on …
- Result: no significant difference

Finally, Analyze the Data

- Choose an appropriate statistical test. (there are entire courses on this!)
- Run it
- Hope for p<=0.5
- Summary
  - Statistical studies are a lot of work (too much for this class)
  - Right choice for answering X>Y questions

Recommendation

- How to Lie with Statistics, 1st ed., Darrel Huff, 1954.
- First edition is still in print!
- Excellent description of the (unintended?) implicit bias that can arise in presenting statistical data