Ethical Considerations

- Sometimes tests can be distressing
  - users have left in tears (embarrassed by mistakes)
- You have a responsibility to alleviate
  - make voluntary with informed consent
  - avoid pressure to participate
  - will not affect their job status either way
  - let them know they can stop at any time
  - stress that you are testing the system, not them
  - make collected data as anonymous as possible
- Get human subjects approval if needed – typically if results are going to be published

Variable types

- Independent Variables: the ones you control
  - Aspects of the interface design
  - Characteristics of the testers
  - Discrete: A, B or C
  - Continuous: Time between clicks for double-click
- Dependent variables: the ones you measure
  - Time to complete tasks
  - Number of errors

Deciding on Data to Collect

- Two types of data
  - process data
    - observations of what users are doing & thinking
  - bottom-line data
    - summary of what happened (time, errors, success, ...)
    - i.e., the dependent variables

Some statistics

- Variables X & Y
- A relation (hypothesis) e.g. $X > Y$
- We would often like to know if a relation is true
  - e.g. $X = \text{time taken by novice users}$
  - $Y = \text{time taken by users with some training}$
- To find out if the relation is true we do experiments to get lots of x's and y's (observations)
  - Suppose $\text{avg}(x) > \text{avg}(y)$, or that most of the x's are larger than all of the y's. What does that prove?

Using Subjects

- Between subjects experiment
  - Two groups of test users
  - Each group uses only 1 of the systems
- Within subjects experiment
  - One group of test users
  - Each person uses both systems
Between subjects

- Two groups of testers, each use 1 system
- **Advantages:**
  - Users only have to use one system (practical).
  - No learning effects.
- **Disadvantages:**
  - Per-user performance differences confounded with system differences:
  - Much harder to get significant results (many more subjects needed).
  - Harder to even predict how many subjects will be needed (depends on subjects).

Within subjects

- One group of testers who use both systems
- **Advantages:**
  - Much more significance for a given number of test subjects.
- **Disadvantages:**
  - Users have to use both systems (two sessions).
  - Order and learning effects (can be minimized by experiment design).

Significance

- The significance or p-value of an outcome is the probability that it happens by chance if the relation does not hold.
- E.g. $p = 0.05$ means that there is a 1/20 chance that the observation happens if the hypothesis is false.
- So the smaller the p-value, the greater the significance.

Normal distributions

- Many variables have a Normal distribution
- At left is the density, right is the cumulative prob.
- Normal distributions are completely characterized by their mean and variance (mean squared deviation from the mean).

Statistics with care:

- What you can do to get better significance:
  - Run each subject several times, compute the average for each subject.
  - Run the analysis as usual on subjects’ average times, with $n$ = number of subjects.
- This decreases the per-subject variance, while keeping data independent.
Some statistics

• Variables X & Y
• A relation (hypothesis) e.g. X > Y
• We would often like to know if a relation is true
  – e.g. X = time taken by novice users
  – Y = time taken by users with some training
• To find out if the relation is true we do experiments to get lots of x’s and y’s (observations)
• Suppose avg(x) > avg(y), or that most of the x’s are larger than all of the y’s. What does that prove?

Empirical Research

• Correlational research
  – Don’t manipulate any variables
  – Look for correlation between variables
  – E.g. “Price of beer is positively correlated with wages of judges.”
• Experimental research
  – Has both dependent and independent variables
  – Can demonstrate causality with control group
  – E.g. “effect of spell-check display”

Pilot Usability Study Tips

• Report results in terms
  – Process data
  – Bottom-line data
• Other questions?

Administrivia

• Changes to grading scheme for hi-fi #1 presentation
  – In response to student feedback,
  – Original 20 points for group grade will now go to the presenters, i.e. all 40 points go to presenters
  – So, rest of group will not be penalized for what presenters did
  – This will be grading scheme for hi-fi #2 presentation

• Online readings for last Monday’s lecture (shopping card, inverted pyramid web design patterns) will be posted on lecture homepage soon