Evaluation 3

John Stasko

Spring 2007

Agenda (for 3 evaluation lectures)

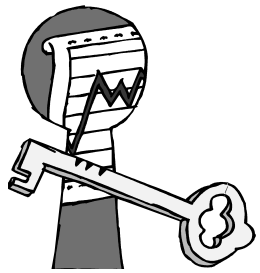
- Evaluation overview
- Designing an experiment
  - Hypotheses
  - Variables
  - Designs & paradigms
- Participants, IRB, & ethics
- Gathering data
  - Objective; Subjective data
- Analyzing & interpreting results
- Using the results in your design

Evaluation, Day 3

- Inspecting your data
- Analyzing & interpreting results
- Using the results in your design
- Usability specifications

Data Inspection

- Look at the results
- First look at each participant's data
  - Were there outliers, people who fell asleep, anyone who tried to mess up the study, etc.?
- Then look at aggregate results and descriptive statistics



Inspecting Your Data

- "What happened in this study?"
- Keep in mind the goals and hypotheses you had at the beginning
- Questions:
  - Overall, how did people do?
  - "5 W's" (Where, what, why, when, and for whom were the problems?)

Descriptive Statistics

- For all variables, get a feel for results:
- Total scores, times, ratings, etc.
- Minimum, maximum
- Mean, median, ranges, etc.

- e.g. "Twenty participants completed both sessions (10 males, 10 females; mean age 22.4, range 18-37 years)."
- e.g. "The median time to complete the task in the mouse-input group was 34.5 s (min=19.2, max=305 s)."

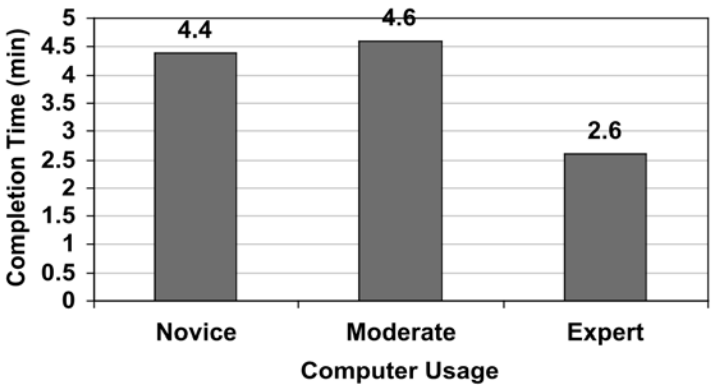
What is the difference between mean & median? Why use one or the other?

Subgroup Stats

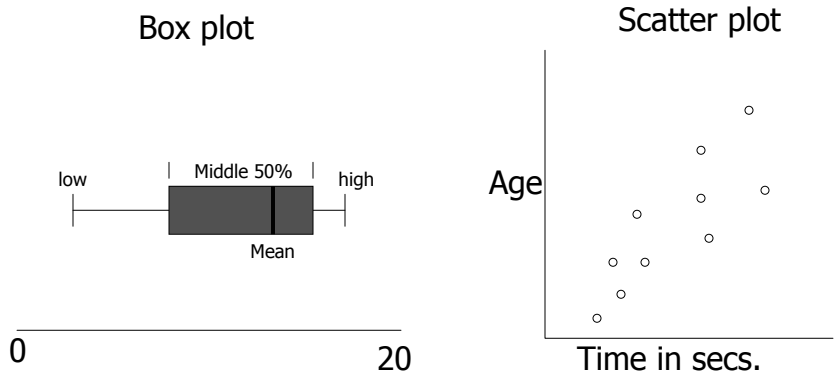
- Look at descriptive stats (means, medians, ranges, etc.) for any subgroups
  - e.g. "The mean error rate for the mouse-input group was 3.4%. The mean error rate for the keyboard group was 5.6%."
  - e.g. "The median completion time (in seconds) for the three groups were: novices: 4.4, moderate users: 4.6, and experts: 2.6."

Plot the Data

- Look for the trends graphically



Other Presentation Methods



Experimental Results

- How does one know if an experiment's results mean anything or confirm any beliefs?
- Example: 40 people participated,
  28 preferred interface 1,
  12 preferred interface 2
- What do you conclude?

Inferential (Diagnostic) Stats

- Tests to determine if what you see in the data (e.g., differences in the means) are reliable (replicable), and if they are likely caused by the independent variables, and not due to random effects
  - e.g., t-test to compare two means
  - e.g., ANOVA (Analysis of Variance) to compare several means
  - e.g., test "significance level" of a correlation between two variables

Means Not Always Perfect

Experiment 1

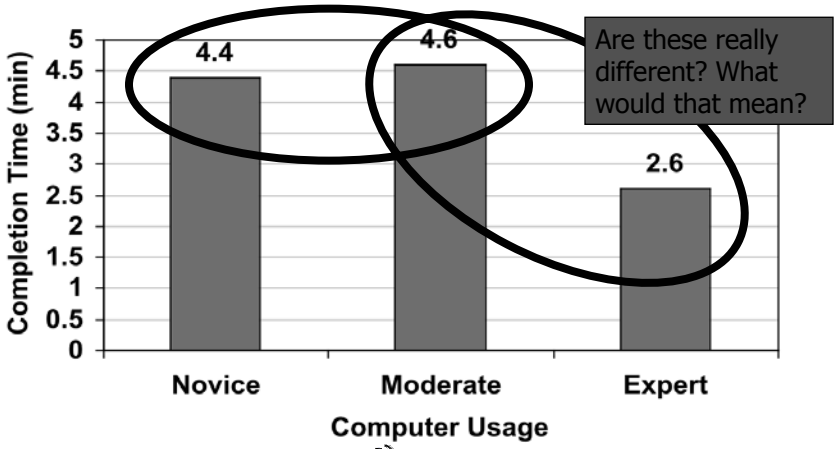
Group 1	Group 2
Mean: 7	Mean: 10
1,10,10	3,6,21

Experiment 2

Group 1	Group 2
Mean: 7	Mean: 10
6,7,8	8,11,11

Inferential Stats and the Data

- Ask diagnostic questions about the data



Hypothesis Testing

- Recall: We set up a "null hypothesis"
  - e.g., there should be no difference between the completion times of the three groups
  - Or, H_0: $Time_{Novice} = Time_{Moderate} = Time_{Expert}$
- Our real hypothesis was, say, that experts should perform more quickly than novices

Hypothesis Testing

- "Significance level" (p):
  - The probability that your null hypothesis was wrong, simply by chance
  - Can also think of this as the probability that your "real" hypothesis (not the null), is wrong
  - The cutoff or threshold level of p ("alpha" level) is often set at 0.05, or 5% of the time you'll get the result you saw, just by chance
  - e.g. If your statistical t-test (testing the difference between two means) returns a t-value of t=4.5, and a p-value of p=.01, the difference between the means is statistically significant

Errors

- Errors in analysis do occur
- Main Types:
  - Type I/False positive - You conclude there is a difference, when in fact there isn't
  - Type II/False negative - You conclude there is no different when there is
  - Dreaded Type III

Drawing Conclusions

- Make your conclusions based on the descriptive stats, but back them up with inferential stats
  - e.g., "The expert group performed faster than the novice group $t(1,34) = 4.6$, $p > .01$."
- Translate the stats into words that regular people can understand
  - e.g., "Thus, those who have computer experience will be able to perform better, right from the beginning…"

Beyond the Scope…

- Note: We cannot teach you statistics in this class, but make sure you get a good grasp of the basics during your student career, perhaps taking a stats class.

Feeding Back Into Design

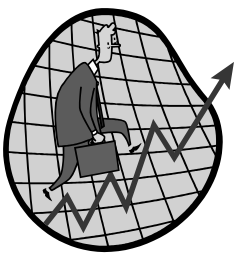
- Your study, was designed to yield information you can use to redesign your interface
- What were the conclusions you reached?
- How can you improve on the design?
- What are quantitative benefits of the redesign?
  - e.g., 2 minutes saved per transaction, which means 24% increase in production, or $45,000,000 per year in increased profit
- What are qualitative, less tangible benefit(s)?
  - e.g., workers will be less bored, less tired, and therefore more interested --> better cust. service

Usability Specifications

"Is it good enough…

…to stop working on it?

…to get paid?"



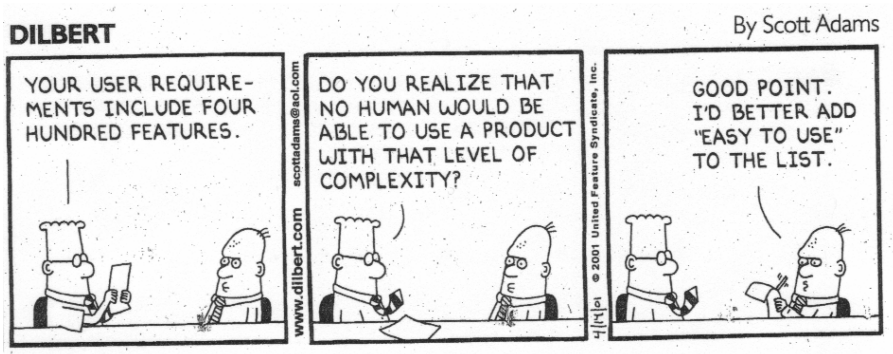
How do we judge these things?

Usability Specifications

- Quantitative usability goals, used a guide for knowing when interface is "good enough"
- Should be established as early as possible
  - Generally a large part of the Requirements Specifications at the center of a design contract
  - Evaluation is often used to demonstrate the design meets certain requirements (and so the designer/developer should get paid)
  - Often driven by competition's usability, features, or performance

Formulating Specifications

- They're often more useful than this…



Measurement Process

- "If you can't measure it, you can't manage it"

- Need to keep gathering data on each iterative evaluation and refinement
- Compare benchmark task performance to specified levels
- Know when to get it out the door!

What is Included?

- Common usability attributes that are often captured in usability specs:
  - Initial performance
  - Long-term performance
  - Learnability
  - Retainability
  - Advanced feature usage
  - First impression
  - Long-term user satisfaction

Quantitative

Assessment Technique

How will you judge whether your design meets the criteria?

Usability attribute	Measure instrum.	Value to be meas.	Current level	Worst perf. level	Planned target level	Best poss level	Observ results
Initial perf	Benchmk task	Length of time to successfully add appointment on the first trial	15 secs (manual)	30 secs	20 secs	10 secs	
First impression	Quest	-2..2	??	0	0.75	1.5	

Explain

Fields

- Measuring Instrument
  - Questionnaires, Benchmark tasks
- Value to be measured
  - Time to complete task
  - Number of percentage of errors
  - Percent of task completed in given time
  - Ratio of successes to failures
  - Number of commands used
  - Frequency of help usage
- Target level
  - Often established by comparison with competing system or non-computer based task

Summary

- Usability specs can be useful in tracking the effectiveness of redesign efforts
- They are often part of a contract
- Designers can set their own usability specs, even if the project does not specify them in advance
- Know when it is good enough, and be confident to move on to the next project

P2 Recap

- Things we noticed

P3

- Due Tuesday
- Key parts
  - Quick recap of problem
  - Design choice --- why?
  - Description of prototype
  - Some "gut" assessment
  - Beginnings of eval plan for P4

P4 Tips

- Consider running different kinds of evaluation
  - Empirical (necessary)
  - Heuristic eval
  - …
- Consider running different kinds of sessions
  - Trained, performance evaluation
  - Untrained, learning evaluation
- Make it appropriate for your system

Upcoming

Transition to focus topics

- Universal Design
- WWW design and evaluation