Questions

- Final presentation (see wiki)
- Comments expectations

- Course evaluations
  [www.courseevalum.umd.edu](http://www.courseevalum.umd.edu)
- No office hours today

---

**Qualitative vs. Quantitative**

- Qualitative: Develop understanding of human experience
- Quantitative: Objectively measure human performance

- Less about more vs. more about less

When are each appropriate?
Quantitative Evaluation

- Gather (performance) measurements

- Methods
  - User interaction collection (i.e., logging)
    - Mouse clicks, keys pressed, ...
    - Data collected during system use
      - Google, Amazon
  - Controlled experiments
    - Set forth a testable hypothesis
    - Manipulate one or more independent variable
    - Observe effect on one or more dependent variable
    - Can be reproduced by others

Controlled experiment

- State a lucid, testable hypothesis
- Identify independent and dependent variables
- Design the experimental protocol
- Choose the user population
- Apply for human subjects protocol review (IRB)
- Run some pilot participants
- Fix the experimental protocol
- Run the experiment
- Perform statistical analysis
- Draw conclusions
- Communicate results
**Question Experiment**

- Is it reliable?
  - Will you get the same result if someone else repeats the experiment?
  - Does the experiment take into account variations between subjects?
    - *Need for testing a sample of subjects*

- Is it valid?
  - Does the experiment reflects target use?
    - *Were users typical?*
    - *Were tasks typical?*
    - *Was the setting realistic?*
    - *Was the experience biased?*

---

**Are results significant?**

- Statistical significance
  - Chance that observed result is due to chance
  - Type I errors are the most disruptive

<table>
<thead>
<tr>
<th>Researcher’s Decision</th>
<th>Actual Situation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NO effect</td>
<td>Effect</td>
</tr>
<tr>
<td>NO effect</td>
<td>Correct decision</td>
<td>Type II error</td>
</tr>
<tr>
<td>Effect</td>
<td>Type I error</td>
<td>Correct decision</td>
</tr>
</tbody>
</table>

- Design significance?
  - 3.00s versus 3.05s?
Are results significant?

- **Statistical significance**
  - Comparing to the null hypothesis: "There is no effect"
  - Type I errors are the most disruptive

<table>
<thead>
<tr>
<th>Researcher’s Decision</th>
<th>Actual Situation: Null Hypothesis is</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail to reject the null hypothesis</td>
<td>Correct decision</td>
<td><strong>Type II error</strong></td>
</tr>
<tr>
<td>Reject the null hypothesis</td>
<td><strong>Type I error</strong></td>
<td>Correct decision</td>
</tr>
</tbody>
</table>

- **Design significance?**
  - 3.00s versus 3.05s?

Example

- Examine value of animation during scrolling
  [Klein & Bederson 2005]
  - Various reading tasks
  - For various document types
  - While scrolling with varying kinds of animated transitions
State a lucid, testable hypothesis

“Animated scrolling will speed up reading and decrease errors, especially for plain documents.”

Choose the variables

- Manipulate one or more independent variable (the thing you change)
  - Document type
  - Animation speed
- Observe effect on one or more dependent variable (the thing you measure)
  - Time to completion
  - Accuracy (i.e., error rate)
Design the experimental protocol

• Between or within subjects?
  – Between subjects: each subject runs one condition
    - Need more subjects
    - Difference between subjects might introduce a bias
      + No learning effects across conditions
  – Within subjects: each subject runs several conditions
    + Need fewer subjects
    + No bias across participants
    - No learning effects across conditions
  – Very important for the statistical analysis phase

• Which tasks?
  – Must reflect the hypothesis
  – Must avoid bias
    • Instructions, ordering...
    • In doubt, always favor the null hypothesis

Design the experimental protocol

• Running Example:
  – Reading while scrolling

Confound!
Chose the user population

• Pick a well balanced sample
  – Novices, experts, average
  – Age group
  – Sex…

• Population group may be one of the independent variable

• Running example
  – Varied population, did not control

Run the experiment

• Always run pilots first!
  – There are always unexpected problem!
  – When the experiment has started you cannot pick and choose

• Use a check-list so that all subjects follow the same steps

• IRB - Don’t forget the consent form!

• Don’t forget to debrief each subject
Run statistical analysis

• Properties of our result data
  – Mean, variance…

• How different data sets relate to each other
  – Are we sampling from similar or different distributions?

• Probability that our claims are correct
  – Statistical significance:
    “The hypothesis that technique X is faster is accepted (p < .05)”
    means that there is a higher than 95% chance the hypothesis is true
  – Typical levels are .05 and .01 level
  – These levels are socially determined

Statistical tools I - Descriptive

• Mean
  – Average

• Median
  – Central value

• Standard Deviation
  – Measure of variation
Statistical tools I - Descriptive

• Correlation
  – Measure the extent to which 2 concepts are related
  – Caveats
    • \textit{Correlation does not imply cause and effect (hidden variable)}
      – Ice cream consumption and drowning
      – Third variable problem
      – Directionality problem
    • \textit{Need a large enough group}
  – \( r = 0 \) (no correlation)
  – \( r = 1 \) (positively correlated)
  – \( r = -1 \) (negatively correlated)

Example (random data)

\begin{tabular}{|c|c|c|}
  \hline
  Person & Height & Self Esteem \\
  \hline
  1 & 68 & 4.1 \\
  2 & 71 & 4.6 \\
  3 & 62 & 3.8 \\
  4 & 75 & 4.4 \\
  5 & 58 & 3.2 \\
  6 & 60 & 3.1 \\
  7 & 67 & 3.8 \\
  8 & 68 & 4.1 \\
  9 & 71 & 4.3 \\
  10 & 69 & 3.7 \\
  11 & 68 & 3.5 \\
  12 & 67 & 3.2 \\
  13 & 63 & 3.7 \\
  14 & 62 & 3.3 \\
  15 & 60 & 3.4 \\
  16 & 63 & 4.0 \\
  17 & 65 & 4.1 \\
  18 & 67 & 3.8 \\
  19 & 63 & 3.4 \\
  20 & 61 & 3.6 \\
  \hline
\end{tabular}

\begin{tabular}{|c|c|}
  \hline
  MEAN & 65.4 \\
  SD & 4.4 \\
  \hline
\end{tabular}
Example (continued)

Correlation: $r = 0.73$

Statistical tools I - Descriptive

- Regression
  - Calculate the “best fit”

<table>
<thead>
<tr>
<th></th>
<th>$r$</th>
<th>$R^2$</th>
<th>Slope</th>
<th>Intercept (s)</th>
<th>IP (bps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScrollPoint</td>
<td>0.97</td>
<td>0.94</td>
<td>0.84</td>
<td>0.42</td>
<td>1.19</td>
</tr>
<tr>
<td>Accel W1</td>
<td>0.90</td>
<td>0.81</td>
<td>1.16</td>
<td>-0.51</td>
<td>0.86</td>
</tr>
<tr>
<td>Accel W3</td>
<td>0.97</td>
<td>0.95</td>
<td>0.80</td>
<td>0.18</td>
<td>1.25</td>
</tr>
<tr>
<td>Wheel Std</td>
<td>0.94</td>
<td>0.88</td>
<td>1.25</td>
<td>-0.42</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Statistical tools II - Analytical

• **T-test**  
  - Compare the mean of 2 populations  
    • *Null hypothesis: no difference between means*  
    • *Can only examine a single independent variable*  
  - Assumptions  
    • *Samples are normally distributed*  
      - Very robust in practice  
    • *Population variances are equal*  
      - Reasonably robust for differing variances  
    • *Individual observations in samples are independent*  
      - Very important

• **ANOVA**  
  - Single factor analysis of variance  
    • *Compare three or more means*  
  - Analysis of variance  
    • *Compare relationship between many factors*  
      - Beginners type at the same speed on all keyboards,  
      - Touch-typist type fastest on the qwerty

• Your protocol influences the kind of test you can use  
  - If in doubt, consult with a statistician before starting the experiment!
Reporting Results

• “This analysis revealed a significant main effect for Device, F(2,15)=15.2, p<0.001.”

• “As one would expect, movement times increased as either W decreased or D increased (i.e., as the task got more difficult: for W, F(2,25)=801, p<0.001; and for D, F(3,54)=1429, p<0.001).”

Running example data

Error bars show +/- one standard deviation
Running Example Analysis

- Measured speed, error and RSD (relative subjective duration)
- Significant results for
  - reading time
  - reading error
  - counting time
- Animated scrolling
  - reduces reading errors by up to 54%
  - task time by up to 3%
  - task time by up to 24% for counting tasks
- 300ms seems to be best overall animation time for reading.
- Formatted documents had a higher base rate perf, and lower improvement – thus visual landmarks appear to be a good idea

In Class Experiment

- Compare mouse vs. keyboard
- Independent variable:
  - Interaction type (conditions: mouse, keyboard)
- Dependent variable:
  - Speed
- Experiment:
  - Bring up two text editor windows with 10 lines of text in one
  - Copy each line of text (one at a time) from one window to the other
  - Conditions:
    - Select text with mouse and press ctrl-c, then alt-tab, ctrl-v
    - Select text with mouse and selected edit->copy menu, then edit->paste
  - Manually measure entire task and divide by 10
  - Train on 3 lines
- Enter data at [http://bederson.etherpad.com/4](http://bederson.etherpad.com/4)