ADMINISTRIVIA

- next Mon is a holiday—BC’s family day
  - no office hours
  - post questions to Piazza
- MS II - due this morning
- Design reviews in workshop on Friday (attendance is mandatory!)
- MS III - briefly introduce (posted)

*** Family day is not until Feb 18th! During reading week. Not next week.

MSIII: WORKPLAN RECOMMENDATION

To allow yourself adequate time for all Parts A, B & C, we suggest the following workplan. You have approximately 3.5 weeks (including reading break):

- **By the end of the 1st week**: Complete Part A (Steps 1 & 2, start Step 3).
- **By the end of the 2nd week**: Complete Step 3, and have Step 4 underway. Complete Step 5, and have Step 6 underway.
  - Decide which team members will be working on refining the experiment and which will be focused on prototype implementation.
- **By halfway through the 3rd week**: Close to completing Step 4, and have significant progress on Step 6.
  Use workshop to get feedback on the experiment design and the plans for the prototype. There should be a clear plan about which team members will be completing which steps in the deliverable. There will be less than a week left in this stage.

  **MSIII Blog Update #3 due 1.5 weeks from now**: covers Steps 1-3

LAST TIME . . . STATISTICAL ANALYSIS

- what is a statistic?
  - a number that describes a sample
  - sample is a subset (hopefully representative) of the population we are interested in understanding
- statistics are calculations that tell us
  - mathematical attributes about our data sets (sample)
    - mean, amount of variance, ...
  - how data sets relate to each other
    - whether we are “sampling” from the same or different populations
  - the probability that our claims are correct
    - “statistical significance”
**Learning Goals**

- Why are statistics used?
- What is an analysis of variance (ANOVA)?
- What is the important terminology in ANOVA?
- What are the different types of ANOVA?
- When would one choose to use an ANOVA?
- What is the difference between statistical and practical significance?
- What other statistics are relevant to HCI?

Acknowledgement: Some of the material in this lecture is based on material prepared for similar courses by Saul Greenberg (University of Calgary)

**Analysis of Variance (ANOVA)**

- A workhorse
  - Allows moderately complex experimental designs (relative to t-test)

- Terminology
  - Factor
    - Independent variable
    - I.e., Keyboard, Menu, Age
  - Factor level
    - Specific value of independent variable
    - I.e., Qwerty, Pop-up menu, 5-10 years old

**ANOVA Terminology**

**Between subjects**
- A subject is assigned to only one factor level of treatment
- Requires more subjects
- Problem: Greater variability, requires more subjects

<table>
<thead>
<tr>
<th>Qwerty</th>
<th>Keyboard</th>
<th>Alphabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-20</td>
<td>S21-40</td>
<td>S41-60</td>
</tr>
</tbody>
</table>

**Within subjects**
- Subjects assigned to all factor levels of a treatment
- Requires fewer subjects
- Less variability as subject measures are paired
- Problem: Order effects (e.g., learning)
- Partially solved by counter-balanced ordering

<table>
<thead>
<tr>
<th>Qwerty</th>
<th>Keyboard</th>
<th>Alphabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-20</td>
<td>S1-20</td>
<td>S1-20</td>
</tr>
</tbody>
</table>

**F Statistic**

Within group variability (WG)
- Individual differences
- Error

<table>
<thead>
<tr>
<th>Qwerty</th>
<th>Keyboard</th>
<th>Alphabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.9, 7.6,...</td>
<td>3.9, 11.2,...</td>
<td>3.5, 3.4,...</td>
</tr>
<tr>
<td>3.7, 3.7</td>
<td>3.10, 2.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Between group variability (BG)
- Treatment effects
- Individual differences
- Error

<table>
<thead>
<tr>
<th>Qwerty</th>
<th>Keyboard</th>
<th>Alphabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.9, 7.6,...</td>
<td>5.9, 11.2,...</td>
<td>3.5, 3.4,...</td>
</tr>
<tr>
<td>3.7, 3.7</td>
<td>3.10, 2.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

These two variability's combine to give total variability
- We are mostly interested in ______ variability because we are trying to understand the effect of the treatment
**F STATISTIC**

ANOVA is what we call an *omnibus* test

- tells us if ($\bar{x}_1 = \bar{x}_2 = \bar{x}_3$) IS NOT true
- doesn’t tell us HOW the means differ (i.e. $\bar{x}_1 > \bar{x}_2$)

$$f = \frac{BG}{WG} = \frac{\text{treatment} + \text{id} + \text{error}}{\text{id} + \text{error}} = ?$$

= 1, if there are no treatment effects

> 1, if there are treatment effects

within-subjects design: the id component in numerator and denominator factored out, therefore a more powerful design

**STATISTICAL SIGNIFICANCE VS. PRACTICAL SIGNIFICANCE**

- when $N$ is large, even a trivial difference (small effect) may be large enough to produce a statistically significant result
  - e.g., menu choice:
    mean selection time of menu A is 3 seconds;
    menu B is 3.05 seconds

- statistical significance does not imply that the difference is important!
  - a matter of interpretation, i.e., subjective opinion
  - should always report means to help others make their opinion

- there are measures for effect size
  - regrettably they are not widely used in HCI research

**SINGLE FACTOR ANALYSIS OF VARIANCE**

- compare means between two or more factor levels within a single factor

- e.g.:
  - dependent variable: typing speed (time)
  - independent variable (factor): keyboard
  - between subject design

<table>
<thead>
<tr>
<th>Subject</th>
<th>Qwerty</th>
<th>Alphabetic</th>
<th>Dvorak</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1:</td>
<td>25 secs</td>
<td>40 secs</td>
<td>17 secs</td>
</tr>
<tr>
<td>S2:</td>
<td>29</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>540: 33</td>
<td>S50: 45</td>
</tr>
<tr>
<td>S10:</td>
<td></td>
<td>S60: 23</td>
<td></td>
</tr>
</tbody>
</table>

also called a one-way ANOVA
ANOVA TERMINOLOGY

• factorial design
  – cross combination of levels of one factor with
  levels of another
  – e.g., keyboard type (3) x expertise (2)

• Cell [or condition]
  – unique treatment combination
  – e.g., qwerty x non-typist

2-way factorial ANOVA

ANOVA

• compares the relationships between many factors
• provides more informed results
  – considers the interactions between factors
  – e.g.,
    • typists type faster on Dvorak, than on alphabetic and Qwerty
    • non-typists are fastest on alphabetic

ANOVA

in reality we rarely want to look at one variable at a time
example:
  – t-test:
    subjects faster on dvorak qwerty
  – anova: keyboard x expertise
    alphabetic fastest for non-typists
dvorak fastest for typists
ANOVA CASE STUDY

WIMP (GUI) vs. HYBRID (graphical command line)

motivation:
• WIMP interfaces are slow because of the mouse
• Command line interfaces are fast (but harder to learn)
• can we create a hybrid interface that is graphical but can be fully operated through the keyboard?
  – assume that one has been designed: inspired by Jeff Hendy’s work on GEKA, Keyboard based dialog boxes
  • how should it be evaluated?

possible hypotheses:
• H1: experts will perform better than novices (not that interesting)
• H2: novices will perform better with WIMP than hybrid
• H3: experts will perform better with hybrid than WIMP, but only for commands with one or more parameters

2 level (interface) x 2 level (expertise) x 3 level (parameters)

mixed-factor design (split plot)
TASK

• assume that the task is to enter a whole series of commands, one after the other

• there is an equal number of 0, 1, and 2 parameter commands used

• identical commands are used in both interface conditions

EFFECTS IN ANOVA

- main effect: the effect of the variable collapsing across all levels of other variables in the experiment
  - i.e. holding all other variable

- interaction effect: the effect of one variable differs depending on the level of another (other) variable(s)

<table>
<thead>
<tr>
<th>Effect</th>
<th>F-ratio</th>
<th>p</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface (I)</td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expertise (E)</td>
<td>5.5</td>
<td>&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>Parameters (P)</td>
<td>31.0</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>IxE</td>
<td>15.2</td>
<td>&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>IxP</td>
<td>8.0</td>
<td>&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>ExP</td>
<td>5.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IxEExP</td>
<td>14.1</td>
<td>&lt;0.05</td>
<td></td>
</tr>
</tbody>
</table>

STATISTICAL RESULTS: SPEED

- Interface x Expertise x Parameters (IxExP)
  - main effect: the effect of the variable collapsing across all levels of other variables in the experiment
  - interaction effect: the effect of one variable differs depending on the level of another (other) variable(s)

<table>
<thead>
<tr>
<th>Effect</th>
<th>F-ratio</th>
<th>p</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface (I)</td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expertise (E)</td>
<td>5.5*</td>
<td>&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>Parameters (P)</td>
<td>31.0**</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>IxE</td>
<td>15.2*</td>
<td>&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>IxP</td>
<td>8.0*</td>
<td>&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>ExP</td>
<td>5.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IxEExP</td>
<td>14.1*</td>
<td>&lt;0.05</td>
<td></td>
</tr>
</tbody>
</table>
SUMMARY OF RESULTS

Assuming same results for errors as speed, and that post-hoc pair-wise comparisons are statistically significant...

H1: experts will perform better than novices (not that interesting)
   Supported: main effect of expertise, showing experts better

H2: novices will perform better with WIMP than hybrid
   Supported: 2-way interaction effect of interface and expertise, showing novices overall better with WIMP

H3: experts will perform better with hybrid than WIMP, but only for commands with one or more parameters
   Supported: 3-way interaction effect of interface, expertise, and number of parameters, showing experts better with hybrid, but only with one and two parameters

CASE STUDY CONCLUSIONS

• expertise makes a big difference
• WIMP interaction should be kept for novices
• hybrid interaction should be available for experts

OTHER STATISTICAL TESTS COMMONLY USED IN HCI

• If we have time, we will cover them in a future lecture
• Your TA may recommend one or more of these for your analysis in your project
  – Correlation
  – Regression
  – Non-parametric tests
    • Chi-squared
    • Mann-Whitney
    • Wilcoxon signed-rank
    • Kruskal-Wallis
    • Friedman's

RECALL

that ANOVA is an OMNIBUS test
• tells us that there IS a difference
• but does NOT tell us the direction of the difference, or which differences are significant

In an exam, I will generally only ask you to identify overall effects
• If I want you to determine the differences within two specific levels specifically, you’ll be allowed to eyeball it.
**Next Time**

in experiments III:
- Significance levels and two types of error
- Interpreting results from C-TOC prep assignment (see next)

Will be a prep assignment for next Tuesday
- little bit longer than usual (2nd part to C-TOC experiment)
  - will need to collect some data (work with your team)
- posted on Canvas already

---

**Excellent Supplementary Readings**


- **Chapter 2. Experimental research** (pp. 25-44).
- **Chapter 3. Experimental design** (pp. 45-69).
- **Chapter 4. Statistical analysis** (pp. 71-104).

Available to UBC students: