LPSC 444:
ADVANCED METHODS IN
HUMAN-COMPUTER INTERACTION

Lecture 6 – Experiments II

Joanna McGrenere

© 2017 Joanna McGrenere
includes slides from Jessica Dawson
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)
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.
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

• 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?
• other tests: what are correlation & regression?
ANALYSIS OF VARIANCE (ANOVA)

• a workhorse
  – allows moderately complex experimental designs (relative to t-test)

• terminology
  – factor
    • independent variable
    • i.e., Keyboard, Toothpaste, Age
  – factor level
    • specific value of independent variable
    • i.e., Qwerty, Crest, 5-10 years old
ANOVA TERMINOLOGY

**between subjects**
- a subject is assigned to only one factor level of treatment
- problem: greater variability, requires more subjects

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

Qwerty S1-20
Dvorak S1-20
Alphabetic S1-20
within group variability (WG)
- individual differences
- measurement error

between group variability (BG)
- treatment effects
- individual differences
- measurement error

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. \(\bar{x}_1 > \bar{x}_2\))

\[
f = \frac{BG}{WG} = \frac{\text{treatment} + \text{id} + \text{m.error}}{\text{id} + \text{m.error}} = ?
\]

= 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
F STATISTIC

• similar to the t-test, we look up the f value in a table, for a given $\alpha$ and degrees of freedom to determine significance

• thus, f statistic is sensitive to sample size
  – Big N $\rightarrow$ Big Power $\rightarrow$ Easier to find significance
  – Small N $\rightarrow$ Small Power $\rightarrow$ Difficult to find significance

• what we (should) want to know is the effect size
  – does the treatment make a big difference (i.e., large effect)?
  – or does it only make a small difference (i.e., small effect)?
  – depending on what we are doing, small effects may be important findings
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;
    mean 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></th>
<th>Qwerty</th>
<th>Alphabetic</th>
<th>Dvorak</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>25 secs</td>
<td>S21: 40 secs</td>
<td>S51: 17 secs</td>
</tr>
<tr>
<td>S2</td>
<td>29</td>
<td>S22: 55</td>
<td>S52: 45</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>S20</td>
<td>33</td>
<td>S40: 33</td>
<td>S60: 23</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
  – unique treatment combination
  – e.g., qwerty x non-typist

2-way factorial ANOVA

Keyboard
Qwerty
Dvorak
Alphabetic

expertise
non-typist
typist
ANOVA TERMINOLOGY

- **mixed factor**
  - contains both between and within subject combinations

<table>
<thead>
<tr>
<th>expertise</th>
<th>Qwerty</th>
<th>Dvorak</th>
<th>Alphabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-typist</td>
<td>S1-20</td>
<td>S1-20</td>
<td>S1-20</td>
</tr>
<tr>
<td>typist</td>
<td>S21-40</td>
<td>S21-40</td>
<td>S21-40</td>
</tr>
</tbody>
</table>
**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

<table>
<thead>
<tr>
<th>Keyboard</th>
<th>Qwerty</th>
<th>Dvorak</th>
<th>Alphabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-typist</td>
<td>S1-20</td>
<td>S1-20</td>
<td>S1-20</td>
</tr>
<tr>
<td>typist</td>
<td>S21-40</td>
<td>S21-40</td>
<td>S21-40</td>
</tr>
</tbody>
</table>
in reality we rarely want to look at one variable at a time
example:

- t-test:
  subjects faster on dvorak than 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 link 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: https://www.cs.ubc.ca/~jchendy/geka.htm
• how should it be evaluated?
ANOVA CASE STUDY

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

• independent variables:
  – interface: WIMP, hybrid
  – command line expertise: novice, expert
  – command parameters: zero, one, two
    • E.g., bold (zero), font ariel (one), print –copies 2 –color greyscale (two)
    • Note: zero parameter commands can be done using shortcuts keys in WIMP

• dependent variables:
  – performance: speed, error
  – satisfaction
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
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

<table>
<thead>
<tr>
<th></th>
<th>F-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface (I)</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Expertise (E)</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>Parameters (P)</td>
<td>31.0</td>
<td></td>
</tr>
<tr>
<td>IxE</td>
<td>15.2</td>
<td></td>
</tr>
<tr>
<td>IxP</td>
<td>8.0</td>
<td></td>
</tr>
<tr>
<td>ExP</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>IxExP</td>
<td>14.1</td>
<td></td>
</tr>
</tbody>
</table>

**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)
STATISTICAL RESULTS: SPEED (TIME)

Interface x Expertise (IxE)

<table>
<thead>
<tr>
<th>Time</th>
<th>WIMP</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>novice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>expert</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interface x Parameters (IxP)

<table>
<thead>
<tr>
<th>Time</th>
<th>WIMP</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero</td>
<td></td>
<td></td>
</tr>
<tr>
<td>one</td>
<td></td>
<td></td>
</tr>
<tr>
<td>two</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interface x Expertise x Parameters (IxExP)

<table>
<thead>
<tr>
<th>Time</th>
<th>WIMP</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero</td>
<td></td>
<td></td>
</tr>
<tr>
<td>one</td>
<td></td>
<td></td>
</tr>
<tr>
<td>two</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

shortcuts!
SUMMARY OF RESULTS

Assuming same results for errors as speed...

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

- Examples of ANOVA and t-test from lit

Will be a prep assignment for Tuesday, Feb 14

- little bit longer than usual (2\textsuperscript{nd} part to C-TOC experiment)
- will need to collect some data! (work with your team)
- will try to post on Connect by Thursday