CPSC 444: ADVANCED METHODS IN HUMAN-COMPUTER INTERACTION

Lecture 7 – Experiments III

Joanna McGrenere
ADMINISTRIVIA

• Check-in on Friday’s design reviews

• Reminder: MSIII Blog Update #3 due on this Thursday, 6pm

• Friday’s workshop: get feedback on
  • low-fi prototypes
  • experiment goals
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.

2 days from now: MSIII Blog Update #3 due – covers Steps 1-3

By end of reading week: should have “by the end of 2nd week” deliverables done
TODAY – LEARNING GOALS

significance levels and two types of error
   – what is the difference between a type I and type II error?
   – how does choice of significance levels relate to error types?
   – how do I chose a significance level?

example: C-TOC prep assignment
   – recap motivation/design
   – how to interpret results? main effects and interaction effects
   – how to interpret/explain the lack of predicted results?
   – what is relationship between type I error and multiple comparisons?
SIGNIFICANCE LEVELS & TWO TYPES OF ERRORS
TWO TYPES OF ERRORS

**Type I error**: reject the null hypothesis when it is, in fact, true
  - We conclude that there is a genuine effect, when there isn’t one (false positive)
  - Confidence level for statistical tests, \( \alpha \)-level (e.g., \( \alpha = .05 \)), is probability of a Type I error

**Type II error**: accept the null hypothesis when it is, in fact, false
  - We conclude that there is no effect, when there actually is one (false negative)
  - \( \beta \)-level is probability of a Type II error
    - related to power (which is defined as 1-\( \beta \)), and which depends on \( \alpha \)-level, effect size, and sample size
Trade-off exists between planning for these two types of errors

- If try to protect against Type I errors (e.g., set very high confidence level $\alpha = .001$ to make it harder to mistakenly believe an effect exists when it doesn’t), then a much greater chance of Type II errors

- If we try to protect against Type II errors (e.g., set low confidence level $\alpha = .1$ to make it easier to detect an effect if it exists), then a much greater chance of Type I errors

choice of significance level therefore often depends on effects of result
EXAMINING EFFECT OF EACH TYPE OF ERROR

Consider the comparison of two types of menus for user speed.

H₀ There is no difference between Pie menus and traditional pop-up menus

H₁ Pie menus are faster than traditional pop-up menus

What happens if you make a . . . .

• Type I error: (reject H₀, conclude there is a difference, when there isn’t one)
  – effect of making this error?

• Type II: (fail to reject H₀, believe there is no difference, when there is)
  – effect of making this error?
CHOICE OF SIGNIFICANCE LEVELS AND TWO TYPES OF ERRORS

What happens if you make a . . . .

• Type I: (reject $H_0$, believe there is a difference, when there isn’t)
  – extra work developing software and having people learn a new idiom for no benefit

• Type II: (accept $H_0$, believe there is no difference, when there is)
  – use a less efficient (but already familiar) menu

Consider the follow scenarios, where you want to run an experiment to decide which menu type to implement.

For each, is Type I or Type II error preferable? Why?

• **Scenario 1: Redesigning a traditional GUI interface**
  – your team proposes replacing the existing pop-up menus in your company’s flagship application, which is widely used globally by users with a wide range of expertise, to improve user performance

• **Scenario 2: Designing a new application**
  – Your team is designing a new digital mapping application. It will require expert users to perform extremely frequent menu selections.
Recall: Motivation

Demand for cognitive screening has grown. But the growing aging population in Canada has led to long wait times to receive a cognitive consultation in a clinic.

Researchers at UBC have developed a computerized cognitive testing tool (C-TOC):

- intended for self-administration in clinical or home setting
- consists of a battery of tests to assess cognitive function (e.g., memory, attention, etc.)
for your pre assignment, you studied an implementation of the Picture-Word Pair task

- memory-encoding task where the user is shown four pictures and a word

- must choose the picture that best matches the word.
RECALL: RESEARCH PROBLEM

Generally: need to ensure that testing on personal devices is consistent with paper-based version of test used in clinics.

This experiment investigates effect of two factors:

1. Effect of image size on the screen (100px vs. 250px)
   - Known that generally larger targets are faster to select
   - Want to test if these results generalize to tasks with cognitive component. (What does this really mean?)

2. Effect of input device (mouse vs. touchscreen vs. trackpad)
   - Previous work has shown older adults may be faster with touchscreen, but this may depend on type of task.
RECALL: EXPERIMENT DESIGN

x 3 input device conditions
  • mouse, trackpad, touchscreen

x 2 image size conditions
  • 100px, 250px

x 5 sets of 4 trials (1 image selection per trial)
  • a control variable – we ignore in our analysis for simplification

Independent variables as used in experiment:
  • counterbalanced? randomized? something else?
  • within-subjects or between-subjects?

vs. ideally what should have been done?
Recall: Experiment Design

x 3 input device conditions *between-subject, hopefully randomly assigned (ideally within-subject)*
- mouse, trackpad, touchscreen

x 2 image size conditions *within-subject, randomized (ideally counterbalanced)*
- 100px, 250px

x 5 sets of 4 trials (1 image selection per trial) *within-subject, randomized (fine as is)*
- a control variable – we ignore in our analysis for simplification

Independent variables as *used in experiment*:
- counterbalanced? randomized? something else?
- within-subjects or between-subjects?

vs. *ideally what should have been done?*
**RECALL: HYPOTHESES**

**H1:** Users will be faster in performing the task with 250px images than with 100px images (i.e. a main effect of image size).

**H2:** There will be a difference in performance based on whether the participant used a touchscreen, trackpad or mouse (i.e. a main effect of input device).

**H3:** The difference in performance for 250px and 100px images will be similar for each input device (i.e. there will be no interaction effect between devices).

Note: up until now we have generally discussed/presented both $H_0$ and the alternative hypothesis for each independent variable — but researchers often only list the expected results as hypotheses (i.e., they do not list $H_0$).
EXAMPLE ANALYSIS

• Now we will go through sample results (based on submitted results)

• Note that your own results are likely at least somewhat different (Why?)

• We will start by looking at the plots, and then consider the results of the statistical tests
**H1:** Users will be faster in performing the task with 250px images than with 100px images (i.e. a main effect of image size).
H2: There will be a difference in performance based on whether the participant used a touchscreen, trackpad or mouse (i.e. a main effect of input device).
**H3:** The difference in performance for 250px and 100px images will be similar for each input device (i.e. there will be no interaction effect between devices).
**H1:** Users will be faster in performing the task with 250px images than with 100px images (i.e. a main effect of image size).

2-factor mixed ANOVA showed *sig. main effect for image size* ($F_{1,7} = 44.55, p < 0.05$)
**H2:** There will be a difference in performance based on whether the participant used a touchscreen, trackpad or mouse (i.e. a main effect of input device).

2-factor mixed ANOVA showed **NO** sig. main effect for input device ($F_{2,7} = 1.37$, $p = 0.31$)
**H3:** The difference in performance for 250px and 100px images will be similar for each input device (i.e. there will be no interaction effect between devices).

*2-factor mixed ANOVA showed sig. interaction effect (F_{2,7} = 5.31, p < 0.05)*
RECALL: HYPOTHESES

**H1:** Users will be faster in performing the task with 250px images than with 100px images (i.e. a main effect of image size).

*Supported*

**H2:** There will be a difference in performance based on whether the participant used a touchscreen, trackpad or mouse (i.e. a main effect of input device).

*Not supported*

**H3:** The difference in performance for 250px and 100px images will be similar for each input device (i.e. there will be no interaction effect between devices).

*Not supported*

Despite the trends observed in the plots, we did not see all the predicted main effects

– what might explain this result?
ANALYSIS: MAKING MULTIPLE COMPARISONS

Recall:
- t-tests are for comparing two samples
- ANOVA are for comparing more than one sample

• So if our only hypothesis was H1 (i.e. only want to compare performance with two image sizes), we could have use a t-test.

• But we because we compared more than two samples, we ultimately used a factorial ANOVA for everything (including testing H1)

why can’t we just carry out several t-tests to compare all combinations of groups? e.g.,
- touchscreen (100px) vs. mouse (100px)
- mouse(100px) vs. trackpad (100px)
- trackpad (100px) vs. mouse (100px)
- etc. . . .
INFLATED ERROR RATES

• imagine comparing each device type with a t-test
  1. touchscreen vs. trackpad
  2. trackpad vs. mouse
  3. mouse vs. touchscreen

• if each t-test uses $\alpha=0.05$ significance level, probability of not having a Type I error for each test is .95 (95%)

• because we have 3 tests, overall probability of NO Type I error is actually:

  \[ 0.95^3 = 0.95 \times 0.95 \times 0.95 = 0.857 \]

• probability of Type I error is now \(1 - 0.857 = 0.143\) (14.3%)!
DIGGING INTO AN INTERACTION

What might we conclude from the interaction found?

• to test differences between combinations statistically, would need to do additional t-tests between each group

• ways to do this that account for error inflation!
  – but you don’t need to know specifics for this class
Next time

• in experiments IV:
  – ANOVA Case Study from literature
  – Reporting experimental results
  – Types of validity

• Prep assignment (TO BE posted during reading week)
  – research paper posted
  – outline on how to write a report posted

Have a Happy Reading week!