ADMINISTRIVIA

• Design reviews went well
• Friday’s workshop:
  – get feedback on completed
    • 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.

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
  – interpreting results
  – relationship between type I error and multiple comparisons
  – discussion
SIGNIFICANCE LEVELS & TWO TYPES OF ERRORS

Type I error: reject the null hypothesis when it is, in fact, true
  – In other words, 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
  – In other words, 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
    • won’t worry about calculating it in 444

TRADEOFFS AND SIGNIFICANCE LEVELS

<table>
<thead>
<tr>
<th></th>
<th>H(_0) True</th>
<th>H(_0) False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject H(_0)</td>
<td>Type I error (false positive)</td>
<td>Correct inference (true positive)</td>
</tr>
<tr>
<td>Fail to Reject H(_0)</td>
<td>Correct inference (true positive)</td>
<td>Type II error (false negative)</td>
</tr>
</tbody>
</table>

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... \( H_0 \) There is no difference between Pie menus and traditional pop-up menus

\( H_1 \) Pie menus will be faster than traditional pop-up menus

What happens if you make a...

• Type I error: (reject \( H_0 \), conclude there is a difference, when there isn’t one)
  – effect of making this error?

• Type II: (fail to reject \( H_0 \), 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 H₀, 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 H₀, believe there is no difference, when there is)
  – use a less efficient (but already familiar) menu

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

In 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.

C-TOC CASE STUDY

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.)

RECALL: MOTIVATION

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.
  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

- 3 input device conditions
  - mouse, trackpad, touchscreen
- 2 image size conditions
  - 100px, 250px
- 5 sets of 4 trials (1 image selection per trial) for each image size

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: so far we have generally discussed both $H_0$ and the alternative hypothesis for each specific statistical test
  - but researchers often only list the expected results as hypotheses

EXAMPLE ANALYSIS

- Now we will go through sample results (based on submitted results)
- Note that your own results are likely 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).

![Graph](image1.png)

**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).

![Graph](image2.png)

**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).

![Graph](image3.png)

**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 used a t-test.
- But because we compared more than two samples, we ultimately used an 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
  - touchscreen vs. trackpad
  - trackpad vs. mouse
  - 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 probably 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 TBD
  - outline on how to write a report