EVALUATION/TESTING
EXPERIMENTS 1

CPSC 544 FUNDAMENTALS IN DESIGNING INTERACTIVE COMPUTATION TECHNOLOGY FOR PEOPLE (HUMAN COMPUTER INTERACTION)

WEEK 10 – CLASS 18

© Joanna McGrenere and Leila Aflatoony
Includes slides from Karon MacLean and Jessica Dawson
LEARNING GOALS

• what is the experimental method?
• what is an experimental hypothesis?
• how do I plan an experiment?
• why are statistics used?
• within- & between-subject comparisons: how do they differ?
• 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?

Acknowledgement: Some of the material in this lecture is based on material prepared for similar courses by Saul Greenberg (University of Calgary)
• some portion of the material in these lectures on experimental design should be familiar from undergrad stats class, although perhaps presented here from a slightly different perspective

• much of this material is well covered in today’s readings:

- Experimental research. Chapter 2.
- Experimental design. Chapter 3.
Who has designed/run an experiment?
MATERIAL | I ASSUME YOU ALREADY KNOW AND WILL NOT BE COVERED IN LECTURE

- types of variables
- samples & populations
- normal distribution
- variance and standard deviation

*a couple slides on these topics at the end of this lecture if you need review on your own time; largely repeat what was in the readings.*
CONTROLLED EXPERIMENTS

the traditional scientific method
  – reductionist
    • clear convincing result on specific issues
  – in HCI
    • insights into cognitive process, human performance limitations, ...
    • allows comparison of systems, fine-tuning of details ...

strives for
  – lucid and testable hypothesis (usually a causal inference)
  – quantitative measurement
  – measure of confidence in results obtained (inferential statistics)
  – replicability of experiment
  – control of variables and conditions
  – removal of experimenter bias
DESIRED OUTCOME OF A CONTROLLED EXPERIMENT

**statistical inference** of an event or situation’s probability:

“Design A is better <in some specific sense> than Design B”

or, **Design A meets a target:**

“90% of incoming students who have web experience can complete course registration within 30 minutes”
STEPS IN THE EXPERIMENTAL METHOD
STEP 1: BEGIN WITH A LUCID, TESTABLE HYPOTHESIS

Example 1:

• $H_0$: there is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu

• $H_1$: selecting from a pop-up menu will be faster and less error prone than selecting from a pull down menu
STEP I: BEGIN WITH A LUCID, TESTABLE HYPOTHESIS

Example 2:

• $H_0$: there is no difference in the security of passwords/pins generated for people who have attended a security training program compared to those who have not.

• $H_1$: people who have attended a security training program generate more secure passwords/pins compared to those who have not been to the training.
hypothesis = prediction of the outcome of an experiment.

• framed in terms of independent and dependent variables:
  – a variation in the independent variable will cause a difference in the dependent variable.

• aim of the experiment: prove this prediction
  – by: disproving the “null hypothesis”
  – never by: proving the “alternate hypothesis”

H₀: experimental conditions have no effect on performance (to some degree of significance) → null hypothesis

H₁: experimental conditions have an effect on performance (to some degree of significance) → alternate hypothesis
STEP 2: EXPLICITLY STATE THE INDEPENDENT VARIABLES

Independent variables

• things you control/manipulate (independent of how a subject behaves) to produce different conditions for comparison

• two different kinds:
  – treatment manipulated (can establish cause/effect, true experiment)
  – subject individual differences (can never fully establish cause/effect) [not covered in the reading]

in menu experiment

1.
2.
3.
Menu:

• $H_1$: selecting from a pop-up menu will be faster and less error prone than selecting from a pull down menu

Password:

• $H_1$: people who have attended a security training program generate more secure passwords/pins compared to those who have not been to the training.
**STEP 2: EXPLICITLY STATE THE INDEPENDENT VARIABLES**

Independent variables

- things you control/manipulate (independent of how a subject behaves) to produce different conditions for comparison
- two different kinds:
  - treatment manipulated (can establish cause/effect, true experiment)
  - subject individual differences (can never fully establish cause/effect) \[not covered in the reading\]

**in menu experiment**

1. menu type: pop-up or pull-down
2. menu length: 3, 6, 9, 12, 15
3. expertise: expert or novice
STEP 2: EXPLICITLY STATE THE INDEPENDENT VARIABLES

Independent variables

- things you control/manipulate (independent of how a subject behaves) to produce different conditions for comparison

- two different kinds:
  - treatment manipulated (can establish cause/effect, true experiment)
  - subject individual differences (can never fully establish cause/effect) *[not covered in the reading]*

*in menu experiment*

1. menu type: pop-up or pull-down *(treatment)*
2. menu length: 3, 6, 9, 12, 15 *(treatment)*
3. expertise: expert or novice *(often subject, but can train an expert)*
STEP 2: EXPLICITLY STATE THE INDEPENDENT VARIABLES

Independent variables

• things you control/manipulate (independent of how a subject behaves) to produce different conditions for comparison

• two different kinds:
  – treatment manipulated (can establish cause/effect, true experiment)
  – subject individual differences (can never fully establish cause/effect) [not covered in the reading]

in password experiment

1.
2.
3.
STEP 2: EXPLICITLY STATE THE INDEPENDENT VARIABLES

Independent variables

• things you control/manipulate (independent of how a subject behaves) to produce different conditions for comparison

• two different kinds:
  – treatment manipulated (can establish cause/effect, true experiment)
  – subject individual differences (can never fully establish cause/effect) [not covered in the reading]

in password experiment

1. training: yes, no
2. type of online service: financial, e-commerce, other
3. general computer expertise: expert or novice
STEP 2: EXPLICITLY STATE THE INDEPENDENT VARIABLES

Independent variables

- things you control/manipulate (independent of how a subject behaves) to produce different conditions for comparison
- two different kinds:
  - treatment manipulated (can establish cause/effect, true experiment)
  - subject individual differences (can never fully establish cause/effect) [not covered in the reading]

in password experiment

1. training: yes, no  (could be either)
2. type of online service: financial, e-commerce, other  (treatment)
3. general computer expertise: expert or novice  (subject)
STEP 3: CAREFULLY CHOOSE THE DEPENDENT VARIABLES

Dependent variables

– things that are measured

– expectation that they depend on the subject’s behaviour / reaction to the independent variable (but unaffected by other factors)

What else could we measure?

• in menu experiment:

• in password experiment:
STEP 4: CONSIDER POSSIBLE NUISANCE VARIABLES & DETERMINE MITIGATION APPROACH

– undesired variations in experiment conditions which cannot be eliminated, but which may affect dependent variable
  • critical to know about them
– experiment design & analysis must generally accommodate them:
  • treat as an additional experiment independent variable (if they can be controlled)
  • randomization (if they cannot be controlled)
– common nuisance variable: subject (individual differences)

• in menu experiment:
• in password experiment: how to manage?

“Systematic errors” in reading worksheet
STEP 5: DESIGN THE TASK TO BE PERFORMED

tasks must:

be externally valid

– external validity = do the results generalize?
– ... will they be an accurate predictor of how well users can perform tasks as they would in real life?
– for a large interactive system, can probably only test a small subset of all possible tasks.

exercise the designs, bringing out any differences in their support for the task

– e.g., if a design supports website navigation, test task should not require subject to work within a single page

be feasible - supported by the design/prototype, and executable within experiment time scale
STEP 5: DESIGN THE TASK TO BE PERFORMED

• in menu experiment:

• in password experiment:
STEP 6: DESIGN EXPERIMENT PROTOCOL

• steps for executing experiment are prepared well ahead of time
• includes unbiased instructions + instruments (questionnaire, interview script, observation sheet)
• double-blind experiments, ...
STEP 7: MAKE FORMAL EXPERIMENT DESIGN EXPLICIT

simplest: 2-sample (2-condition) experiment

- based on comparison of **two sample means**:
  - performance data from using Design A & Design B
    - e.g., new design & status quo design
    - e.g., 2 new designs

- or, comparison of **one sample mean with a constant**:
  - performance data from using Design A, compared to performance requirement
    - determine whether single new design meets key design requirement
STEP 7: MAKE FORMAL EXPERIMENT DESIGN EXPLICIT

more complex: factorial design

in menu experiment:
  – 2 menu types (pop-up, pull down)
  – x 5 menu lengths (3, 6, 9, 12, 15)
  – x 2 levels of expertise (novice, expert)

in password experiment:
  – 2 training (yes, no)
  – x 3 types of online service (financial, e-commerce, other)
  – x 2 general computer expertise (novice, expert)
WITHIN/BETWEEN SUBJECT COMPARISONS

within-subject design:
subjects exposed to multiple treatment conditions
→ primary comparison internal to each subject
– allows control over subject variable
– greater statistical power, fewer subjects required
– not always possible (exposure to one condition might “contaminate” subject for another condition; or session too long)

between-subject design:
subjects only exposed to one condition
→ primary comparison is from subject to subject
– less statistical power, more subjects required
– why? because greater variability due to more individual differences

split-plot design (also called mixed factorial design)
combination of within-subject and between-subject in a factorial design
WITHIN/BETWEEN SUBJECT COMPARISONS

• in menu experiment:
  • 2 menu types (pop-up, pull down)
  • x 5 menu lengths (3, 6, 9, 12, 15)
  • x 2 levels of expertise (novice, expert)

• in password experiment:
  • 2 training (yes, no)
  • x 3 types of online service (financial, e-commerce, other)
  • x 2 general computer expertise (novice, expert)
WITHIN/BETWEEN SUBJECT COMPARISONS

• in menu experiment:
  • 2 menu types (pop-up, pull down) likely within
  • x 5 menu lengths (3, 6, 9, 12, 15) between
  • x 2 levels of expertise (novice, expert) likely between

-→ split plot design (mixed factorial design)

• in password experiment:
  • 2 training (yes, no) must be between
  • x 3 types of online service (financial, e-commerce, other)
  • x 2 general computer expertise (novice, expert)

-→ split plot design (mixed factorial design)
STEP 8: JUDICIOUSLY SELECT/RECRUIT AND ASSIGN SUBJECTS TO GROUPS

**subject pool:** similar issues as for informal and field studies
- match expected user population as closely as possible
- age, physical attributes, level of education
- general experience with systems similar to those being tested
- experience and knowledge of task domain

**sample size:** more critical in experiments than other studies
- going for “statistical significance”
- should be large enough to be “representative” of population
- guidelines exist based on statistical methods used & required significance of results
- pragmatic concerns may dictate actual numbers
- “10” is often a good place to start
STEP 8: JUDICIOUSLY SELECT/RECRUIT AND ASSIGN SUBJECTS TO GROUPS

• if there is too much variability in the data collected, you will not be able to achieve statistical significance
• you can reduce variability by controlling subject variability
• how?
  – recognize classes and make them an independent variable
    • e.g., older users vs. younger users
    • e.g., superstars versus poor performers
  – use reasonable number of subjects and random assignment
STEP 9: APPLY STATISTICAL METHODS TO DATA ANALYSIS

examples: t-tests, ANOVA, correlation, regression (more on these in upcoming lectures)

confidence limits: the confidence that your conclusion is correct

– “The hypothesis that mouse experience makes no difference is rejected at the .05 level” (i.e., null hypothesis rejected)

– this means:
  • a 95% chance that your finding is correct
  • a 5% chance you are wrong
STEP 10: INTERPRET YOUR RESULTS

• what you believe the results mean, and their implications

• yes, there can be a subjective component to quantitative analysis
THE PLANNING FLOWCHART

Stage 1
- Problem definition
  - research idea
  - literature review
  - statement of problem
  - hypothesis development

Stage 2
- Planning
  - define variables
  - controls
  - apparatus
  - procedures
  - experimental design
  - select subjects

Stage 3
- Conduct research
  - pilot testing
  - data collection

Stage 4
- Analysis
  - data reductions
  - statistics
  - hypothesis testing

Stage 5
- Interpretation
  - interpretation
  - generalization
  - reporting

feedback
GOAL OF EXPERIMENT DESIGN

Guard against ambiguous or misleading results

↩️ A good (definitive) result
POOR EXPERIMENT DESIGN OR RESULTS

less distinguishable results:

perhaps task was poorly chosen – OR there’s really no difference
POOR EXPERIMENT DESIGN

misleading results

e.g. subject assignment not controlled: one design tested on novices, other on experts, disguising actual trend
POOR EXPERIMENT DESIGN OR RESULTS

large spread in values

perhaps conditions were not well controlled?

![Graph showing task performance time for designs A and B.](image)
TO SUMMARIZE SO FAR:

HOW A CONTROLLED EXPERIMENT WORKS

1. formulate an alternate and a null hypothesis:
   – H₁: experimental conditions have an effect on performance
   – H₀: experimental conditions have no effect on performance

2. through experimental task, try to demonstrate that the null hypothesis is false (reject it),
   – for a particular level of significance

3. if successful, we can accept the alternate hypothesis,
   and state the probability p that we are wrong (the null hypothesis is true after all) ⇒ this is result’s confidence level
e.g., selection speed is significantly faster in menus of length 5 than of length 10 (p<.05)
⇒ 5% chance we’ve made a mistake, 95% confident
SIGNIFICANCE LEVELS & TWO TYPES OF ERRORS
**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
Trade-offs and Significance Levels

<table>
<thead>
<tr>
<th>Outcome of Exp’t</th>
<th>Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H₀ True</td>
</tr>
<tr>
<td>Reject H₀</td>
<td><strong>Type I error</strong> (false positive)</td>
</tr>
<tr>
<td>Fail to Reject H₀</td>
<td>Correct inference (true positive)</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.

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.
INTERIM STUDENT FEEDBACK

• Lots of positive feedback – thanks!

• Things feel rushed
  – Classes: in-class activity, discussion, design reviews
    • considering 2 x 2 hour sessions – thoughts?
  – Project:
    • Classic challenge

• Readings are long (but mostly well liked!)
  – Appropriate for grad level
  – Interacts with lecture – we are working to further reduce lecture length
Next time

Thurs: design reviews for prototype phase

Next week:

in experiments II:
– Statistic, including t-test and ANOVA, hopefully case study
– Readings posted by Thurs

in experiments III:
– case study
ADDITIONAL SLIDES: MATERIAL I ASSUME YOU KNOW

• types of variables
• samples & populations
• normal distribution
• variance and standard deviation
TYPES OF VARIABLES
(INDEPENDENT OR DEPENDENT)

• discrete: can take on finite number of levels
  – e.g. a 3-color display can only render in red, green or blue;
  – a design may be version A, or version B

• continuous: can take any value (usually within bounds)
  – e.g. a response time that may be any positive number (to resolution of measuring technology)

• normal: one particular distribution of a continuous variable
POPULATIONS AND SAMPLES

• statistical sample = approximation of total possible set of, e.g.
  – people who will ever use the system
  – tasks these users will ever perform
  – state users might be in when performing tasks

• “sample” a representative fraction
  – draw randomly from population
  – if large enough and representative enough, the sample mean should lie somewhere near the population mean
CONFIDENCE LEVELS

• “the sample mean should lie somewhere near the population mean”
• how close?
• how sure are we?
• a confidence interval provides an estimate of the probability that the statistical measure is valid:
  • “We are 95% certain that selection from menus of five items is faster than that from menus of seven items”
• how does this work?
  important aspect of experiment design
ESTABLISHING CONFIDENCE LEVELS: NORMAL DISTRIBUTIONS

• fundamental premise of statistics:
  – predict behavior of a population based on a small sample

• validity of this practice depends on the distribution
  – of the population and of the sample

• many populations are normally distributed:
  – many statistical methods for continuous dependent variables are based on the assumption of normality

• if your sample is normally distributed, your population is likely to be,
  – and these statistical methods are valid,
  – and everything is a lot easier.
WHAT’S A NORMAL DISTRIBUTION?

population ➔

sample ➔
VARIANCE AND STANDARD DEVIATION

• all normal distributions are not the same:

• population variance is a measure of the distribution’s “spread”
  all normal population distributions still have the same shape
HOW DO YOU GET THE POPULATION’S VARIANCE?

- estimate the population’s (true) variance from the (measured) sample’s standard deviation
WHAT'S THE BIG DEAL?

• if you know you’re dealing with samples from a normal distribution,
• and you have a good estimate of its variance
  – (i.e. your sample’s std dev)