ADMINISTRIVIA

• prep assignment for W06
  – there isn’t one! focus on your project report
  – W07 assignment will be released next week
ADMINISTRIVIA

- project: MSII
  - % of teams who have started observations?
  - % of teams who’ve had TA sign off on study instruments?
  - TAs need turnaround time for approval (~24h)
  - any other issues?

MSII Report & Blog #2 Due at 9AM next Tues
— will not be accepted late!
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?
• what’s involved in computing a t-test?
• what are the different types of t-tests?

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

• much of this material is well covered in today’s reading:
  – Newman & Lamming, Ch 10
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

There are some 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 HCl
    • 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
DESIGNED 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 1: 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.
GENERAL: HYPOTHESIS TESTING

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)  

in menu experiment

1. menu type: pop-up or pull-down  
2. menu length: 3, 6, 9, 12, 15  
3. expertise: expert or novice  

(worksheet)
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) \(\text{[not covered in the reading]}\)

\textit{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?
**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, ...

Now you get to do the pop-up menus. I think you will really like them... I designed them myself!
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) must be between

-> 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

Novice  

Expert
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 arrows connect the stages, indicating the iterative nature of the planning process.
GOAL OF EXPERIMENT DESIGN

- guard against ambiguous or misleading results

\[ \leftarrow \text{a good (definitive) result} \]
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?
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
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”
EXAMPLE: DIFFERENCES BETWEEN MEANS

given: two data sets measuring a condition
  – e.g., time to select an item from different menu styles ...

question:
  – is the difference between the means of the data statistically significant?

null hypothesis:
  – there is no difference between the two means
  – statistical analysis can only reject the hypothesis at a certain level of confidence
  – remember: we never actually prove the null hypothesis true
This time let's just hypothesize about error rate:

- **H₀**: there is no difference in error rate when selecting a single item from a pop-up or a pull down menu
- **H₁**: selecting from a pop-up menu will be less error prone than selecting from a pull down menu
EXAMPLE:

pull down vs. pop-up menus

assume we ran a between-subjects experiment, where we counted the # of errors under each condition

condition 1 (pop-up) : 0, 1, 1, 1, 2, 2, 2, 3

condition 2 (pull down) : 1, 1, 2, 2, 3, 3, 4, 4

Is there a significant difference between the means?
THE PROBLEM WITH VISUAL INSPECTION OF DATA

there is almost always variation in the collected data

differences between data sets may be due to:

• *normal variation*
  
  e.g., two sets of ten tosses with different but fair dice
  
  • differences between data and means are accountable by expected variation

• *real differences between data*
  
  e.g., two sets of ten tosses with loaded dice and fair dice
  
  – differences between data and means are not accountable by expected variation
T-TEST

a statistical test

allows one to say something about differences between two means at a certain confidence level

null hypothesis of the t-test:
  – no difference exists between the means

possible results:
• I am 95% sure that null hypothesis is rejected
  – there is probably a true difference between the means

• I cannot reject the null hypothesis
  – the means are likely the same
DIFFERENT TYPES OF T-TESTS

comparing two sets of independent observations
usually different subjects in each group (number may differ as well)
  – Condition 1     Condition 2
  –  S1–S20            S21–S43

paired observations
usually single group studied under separate experimental conditions

data points of one subject are treated as a pair
  – Condition 1     Condition 2
  –  S1–S20            S1–S20
DIFFERENT TYPES OF T-TESTS

comparing two sets of independent observations *(between subjects)*
usually different subjects in each group (number may differ as well)
  – Condition 1     Condition 2
  – S1–S20            S21–S43

paired observations *(within subjects)*
usually single group studied under separate experimental conditions

data points of one subject are treated as a pair
  – Condition 1     Condition 2
  – S1–S20            S1–S20
DIFFERENT TYPES OF T-TESTS

non-directional vs directional alternatives

non-directional (two-tailed)

– no expectation that the direction of difference matters

directional (one-tailed)

– only interested if the mean of a given condition is greater than the other
T-TESTS

• Assumptions of t-tests
  – data points of each sample are normally distributed
    • but t-test very robust in practice
  – sample variances are equal
    • t-test reasonably robust for differing variances
    • deserves consideration
  – individual observations of data points in sample are independent
    • can’t be seen in data, must be accounted for in research design
      (can you think of examples where they are not independent?)

• Significance level
  – decide upon the level before you do the test!
  – typically stated at the .05 or .01 level
  – .10 can be considered a trend, but is controversial
WHAT THE T-TEST IS TESTING?

a) the two samples come from two different populations;  
b) the two samples are part of the same population.

Which represents $H_0$ and which represents $H_1$?
TWO-TAILED UNPAIRED T-TEST

n: number of data points in the one sample (N = n₁ + n₂)

ΣX: sum of all data points in one sample

X̄: mean of data points in sample

Σ(X²): sum of squares of data points in sample

s²: combined sample variance

t: t ratio

df = degrees of freedom = n₁ + n₂ − 2

How to maximize t?

Formulas

\[ s^2 = \frac{\Sigma(X_1^2) - \left(\frac{\Sigma X_1}{n_1}\right)^2 + \Sigma(X_2^2) - \left(\frac{\Sigma X_2}{n_2}\right)^2}{n_1 + n_2 - 2} \]

\[ t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \]
**Level of Significance for Two-Tailed Test**

<table>
<thead>
<tr>
<th>df</th>
<th>.05</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.706</td>
<td>63.657</td>
</tr>
<tr>
<td>2</td>
<td>4.303</td>
<td>9.925</td>
</tr>
<tr>
<td>3</td>
<td>3.182</td>
<td>5.841</td>
</tr>
<tr>
<td>4</td>
<td>2.776</td>
<td>4.604</td>
</tr>
<tr>
<td>5</td>
<td>2.571</td>
<td>4.032</td>
</tr>
<tr>
<td>6</td>
<td>2.447</td>
<td>3.707</td>
</tr>
<tr>
<td>7</td>
<td>2.365</td>
<td>3.499</td>
</tr>
<tr>
<td>8</td>
<td>2.306</td>
<td>3.355</td>
</tr>
<tr>
<td>9</td>
<td>2.262</td>
<td>3.250</td>
</tr>
<tr>
<td>10</td>
<td>2.228</td>
<td>3.169</td>
</tr>
<tr>
<td>11</td>
<td>2.201</td>
<td>3.106</td>
</tr>
<tr>
<td>12</td>
<td>2.179</td>
<td>3.055</td>
</tr>
<tr>
<td>13</td>
<td>2.160</td>
<td>3.012</td>
</tr>
<tr>
<td>14</td>
<td>2.145</td>
<td>2.977</td>
</tr>
<tr>
<td>15</td>
<td>2.131</td>
<td>2.947</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>df</th>
<th>.05</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>2.120</td>
<td>2.921</td>
</tr>
<tr>
<td>18</td>
<td>2.101</td>
<td>2.878</td>
</tr>
<tr>
<td>20</td>
<td>2.086</td>
<td>2.845</td>
</tr>
<tr>
<td>22</td>
<td>2.074</td>
<td>2.819</td>
</tr>
<tr>
<td>24</td>
<td>2.064</td>
<td>2.797</td>
</tr>
</tbody>
</table>

Critical value (threshold) that t-statistic must reach to achieve significance.

How does critical value change based on df and confidence level?
scenario 2: assume we ran a between-subjects experiment, where we counted the # of errors under each condition

condition 1 (pop-up) : 0, 1, 1, 1, 2, 2, 2, 3

condition 2 (pull down) : 1, 1, 2, 2, 3, 3, 4, 4

Is there a significant difference between the means?
TWO-TAILED UNPAIRED T-TEST

Condition one (pop up): 0, 1, 1, 1, 2, 2, 2, 3
Condition two (pull down): 1, 1, 2, 2, 3, 3, 4, 4

What the results would look like in R.

data: my_data$Condition.1 and my_data$Condition.2
t = -1.8708, df = 13.176, p-value = 0.08374

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-2.1531955 0.1531955

sample estimates:
mean of x mean of y
1.5 2.5

is the difference significant?
TWO-TAILED UNPAIRED T-TEST

Condition one (pop up): 0, 1, 1, 1, 2, 2, 2, 3
Condition two (pull down): 1, 1, 2, 2, 3, 3, 4, 4

data: my_data$Condition.1 and my_data$Condition.2

t = -1.8708, df = 13.176, p-value = 0.08374

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-2.1531955  0.1531955

sample estimates:
mean of x mean of y
1.5  2.5

probability that means are from the same underlying population

How does the outcome change for a confidence level of 0.10?
This time lets just hypothesize about error rate:

- **H0**: there is no difference in error rate when selecting a single item from a pop-up or a pull down menu - cannot reject at 0.5 level
- **H1**: selecting from a pop-up menu will be less error prone than selecting from a pull down menu
SUMMARY OF THE T-TEST

- the point: establish a confidence level in the difference we’ve found between 2 sample means.

- the process (what your stats software does under the hood):
  - compute df
  - choose desired significance, p (aka $\alpha$)
  - calculate value of the t statistic
  - compare it to the critical value of t given $p$, $df$: $t(p,df)$
  - if $t > t(p,df)$, can reject null hypothesis at $p$
Next time

in experiments II:
• ANOVA (including short case study)

no prep assignment for next week
• you will also get experience doing t-tests and ANOVA in W07 prep assignment for CTOC Experiment

in experiments III:
• Types of error
• C-TOC 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)