CS 544
Experimental Design

What is experimental design?
What is an experimental hypothesis?
How do I plan an experiment?
Why are statistics used?
What are the important statistical methods?

Acknowledgement: Some of the material in this lecture is based on material prepared for similar courses by Saul Greenberg (University of Calgary)

Quantitative ways to evaluate systems

- Quantitative:
  - precise measurement, numerical values
  - bounds on how correct our statements are

- Methods
  - Controlled Experiments
  - Statistical Analysis

- Measures
  - Objective: user performance (speed & accuracy)
  - Subjective: user satisfaction

Quantitative methods

1. User performance data collection
   - data is collected on system use
     - frequency of request for on-line
       - what did people ask for help with?
     - frequency of use of different parts of the system
       - why are parts of system unused?
     - number of errors and where they occurred
       - why does an error occur repeatedly?
     - time it takes to complete some operation
       - what tasks take longer than expected?
     - collect heaps of data in the hope that something interesting shows up
     - often difficult to sift through data unless specific aspects are targeted (as in list above)

2. 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

Quantitative methods ...

descriptive statistics
The experimental method

a) Begin with a lucid, testable hypothesis

Example 1:

H₀: there is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste

H₁: children and teenagers using crest toothpaste have fewer cavities than those who use no-teeth toothpaste

Example 2:

H₀: 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, regardless of the subject’s previous expertise in using a mouse or using the different menu types

b) Explicitly state the independent variables that are to be altered

Independent variables

- the things you control (independent of how a subject behaves)
- two different kinds:
  1. treatment manipulated (can establish cause/effect, true experiment)
  2. subject individual differences (can never fully establish cause/effect)

in toothpaste experiment

- toothpaste type: uses Crest or No-teeth toothpaste
- age: <= 12 years or > 12 years

in menu experiment

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

c) Carefully choose the dependent variables that will be measured

Dependent variables

- variables dependent on the subject’s behaviour / reaction to the independent variable

in toothpaste experiment

- number of cavities
- frequency of brushing

in menu experiment

- time to select an item
- selection errors made
The experimental method

d) Judiciously select and assign subjects to groups

Ways of controlling subject variability

- recognize classes and make them and independent variable
- minimize unaccounted anomalies in subject group
- use reasonable number of subjects and random assignment

f) Apply statistical methods to data analysis

- 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)
  - means:
    - a 95% chance that your finding is correct
    - a 5% chance you are wrong

The Planning Flowchart

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

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

Stage 3
Conduct research
pilot testing
data collection

Stage 4
Analysis
hypothesis testing
data reductions

Stage 5
Interpretation
interpretation
generalization
reporting

The experimental method

e) Control for biasing factors

- unbiased instructions + experimental protocols
  prepare ahead of time
- double-blind experiments, ...

The Planning Flowchart

Now you get to do the pop-up menus. I think you will really like them... I designed them myself.
### 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
  - eg height difference of males and females,
  - 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
  - we never actually prove the hypothesis true

### Example:

**Is there a significant difference between the means?**

- Condition one: 3, 4, 4, 5, 5, 6
- Condition two: 4, 4, 5, 5, 6, 6, 7

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

- **Mean:**
  - Condition 1: 4.5
  - Condition 2: 5.5

### 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
    - eg two sets of ten tosses with different but fair dice
      - differences between data and means are accountable by expected variation
  - real differences between data
    - eg 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 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            S21–43

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

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

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
  - must be adhered to

Significance level
- decide upon the level before you do the test!
  - typically stated at the .05 or .01 level

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
- X: mean of data points in sample
- Σ(X^2): sum of squares of data points in sample
- s^2: unbiased estimate of population variation
- t: t ratio
- df = degrees of freedom = N1 + N2 – 2

Formulas

\[
S^2 = \frac{\Sigma (X_i^2) - (\Sigma X_i)^2}{n_1} \cdot \frac{\Sigma (X_i^2) - (\Sigma X_i)^2}{n_2}
\]

\[
t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}}
\]
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>

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

Example Calculation

$x_1 = 3, 4, 4, 5, 5, 6$

Condition one: 3, 4, 4, 4, 5, 5, 5, 6

$x_2 = 4, 4, 5, 5, 6, 6, 7, 7$

Condition two: 4, 4, 5, 5, 6, 6, 7, 7

Hypothesis: there is no significant difference between the means at the .05 level.

Step 1: Calculating $s^2$

Step 2: Calculating $t$

Step 3: Looking up critical value of $t$

Unpaired t-test

<table>
<thead>
<tr>
<th>Group</th>
<th>Count</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>8</td>
<td>4.5</td>
<td>0.96</td>
<td>0.327</td>
</tr>
<tr>
<td>two</td>
<td>8</td>
<td>5.5</td>
<td>1.195</td>
<td>0.423</td>
</tr>
</tbody>
</table>

What the results would look like in stats software.
Choice of significance levels and two types of errors

- Type I error: reject the null hypothesis when it is, in fact, true ($\alpha = .05$)
- Type II error: accept the null hypothesis when it is, in fact, false ($\beta$)

<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>$\alpha$ (Type I error)</td>
<td>$1 - \beta$ (Power)</td>
</tr>
<tr>
<td>Not Reject $H_0$</td>
<td>$1 - \alpha$</td>
<td>$\beta$ (Type II error)</td>
</tr>
</tbody>
</table>

- Effects of levels of significance
  - Very high confidence level (e.g., $\alpha = .0001$) gives greater chance of Type II errors
  - Very low confidence level (e.g., $\alpha = .1$) gives greater chance of Type I errors
  - Tradeoff: choice often depends on effects of result

Hypothesis $H_0$: There is no difference between Pie menus and traditional pop-up menus

- 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

Case 1: Redesigning a traditional GUI interface
- A Type II error is preferable to a Type I error. Why?

Case 2: Designing a digital mapping application where experts perform extremely frequent menu selections
- A Type I error is preferable to a Type II error. Why?

Other Tests: Correlation

- Measures the extent to which two concepts are related
  - E.g., years of university training vs computer ownership per capita

- How?
  - Obtain the two sets of measurements
  - Calculate correlation coefficient
    - $+1$: positively correlated
    - $0$: no correlation (no relation)
    - $-1$: negatively correlated

- Dangers
  - Attributing causality
    - A correlation does not imply cause and effect
    - Cause may be due to a third “hidden” variable related to both other variables
    - E.g., (above example) age, affluence
  - Drawing strong conclusions from small samples
    - Unreliable with small groups
    - Be wary of accepting anything more than the direction of correlation unless you have at least 40 subjects
Sample Study: Cigarette Consumption
Crude Male death rate for lung cancer in 1950 per capita consumption of cigarettes in 1930 in various countries.

Correlation

Regression
- Calculate a line of "best fit"
- Use the value of one variable to predict the value of the other
  - e.g., 60% of people with 3 years of university own a computer

Analysis of Variance (Anova)
- A Workhorse
  - Allows moderately complex experimental designs and statistics
- 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

F statistic

- Within group variability (WG)
  - individual differences
  - measurement error

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

- These two variabilities are independent of one another
- They combine to give total variability
- We are mostly interested in between group variability because we are trying to understand the effect of the treatment

\[ F = \frac{BG}{WG} = \frac{\text{treatment} + \text{id} + \text{m.error}}{\text{id} + \text{m.error}} = 1.0 \]

If there are treatment effects then the numerator becomes inflated

Within-subjects design: the id component in numerator and denominator factored out, therefore a more powerful design

\[ F = \frac{BG}{WG} \]

Thus, F statistic sensitive to sample size.
- Big N → Big Power → Easier to find significance
- Small N → Small Power → 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
  - eg menu choice: mean selection time of menu A is 3 seconds; menu 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
- example:
  - dependent variable: typing speed
  - independent variable (factor): keyboard
  - between subject design

Anova terminology

- Factorial design
  - cross combination of levels of one factor with levels of another
  - eg keyboard type (3) x expertise (2)
- Cell
  - unique treatment combination
  - eg qwerty x non-typist

Anova terminology

- Mixed factor
  - contains both between and within subject combinations
Anova

- Compares the relationships between many factors
- Provides more informed results
  - considers the interactions between factors
  - eg
e.g.
  - typists type faster on Qwerty, than on alphabetic and Dvorak
  there is no difference in typing speeds for non-typists across all keyboards

<table>
<thead>
<tr>
<th></th>
<th>Qwerty</th>
<th>Alphabetic</th>
<th>Dvorak</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-typ</td>
<td>S1-S10</td>
<td>S11-S20</td>
<td>S21-S30</td>
</tr>
<tr>
<td>typist</td>
<td>S31-S40</td>
<td>S41-S50</td>
<td>S51-S60</td>
</tr>
</tbody>
</table>

Anova

- In reality, we can rarely look at one variable at a time
- Example:
  - t-test:
    Subjects who use crest have fewer cavities
  - anova: toothpaste x age
    Subjects who are 12 or less have fewer cavities with crest.
    Subjects who are older than 12 have fewer cavities with no-teeth.

Anova case study

- The situation
  - text-based menu display for very large telephone directory
  - names are presented as a range within a selectable menu item
  - users navigate until unique names are reached
  - but several ways are possible to display these ranges

- Question
  - what display method is best?

Range Delimiters

- Truncation

1) Arbor - Barney
2) Barrymore - Dacker
3) Danby - Estovitch
4) Farquar - Kalmer
5) Kalmerson - Moreen
6) Moriarty - Proctor
7) Practer - Sagin
8) Sagin - Unger
9) Unger - Zlotsky

- (Arbor)
1) A - Barn
2) Barr - Dac
3) Dan - E
4) F - Kalmer
5) Kalman - More
6) Mori - Pra
7) Pra - Sag
8) Sag - Ul
9) Ul - Z

- (A)
1) A
2) Barr
3) Dan
4) F
5) Kalman
6) Mori
7) Pra
8) Sag
9) Ul
10) Z
Span

as one descends the menu hierarchy, name suffixes become similar

Statistical results

<table>
<thead>
<tr>
<th>Scanning speed</th>
<th>F-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range delimiter (R)</td>
<td>2.2*</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Truncation (T)</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Experience (E)</td>
<td>6.5*</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Menu Span (S)</td>
<td>216.0**</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

- main effects
- interactions

Anova case study

Null hypothesis
- six menu display systems based on combinations of truncation and delimiter methods do not differ significantly from each other as measured by people's scanning speed and error rate
- menu span and user experience has no significant effect on these results
- mixed design

Statistical results

Null hypothesis
- 2 level (truncation) x 2 level (menu span) x 3 level (delimiter)
- Full: Truncated / Not Truncated
- Narrow: upper / lower
- Expert Novice

Main effects on selection time
- Full range delimiters slowest
- Truncation has no effect on time
- Narrow span menus are slowest
- Novices are slower
**Statistical results**

<table>
<thead>
<tr>
<th>Error rate</th>
<th>F-ratio</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range delimiter (R)</td>
<td>3.7*</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Truncation (T)</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>Experience (E)</td>
<td>5.6**</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Menu Span (S)</td>
<td>77.9**</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>RxT</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>RxE</td>
<td>4.7*</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>RxS</td>
<td>5.4*</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>TxE</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>TxS</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>ExS</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>RxTxE</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>RxTxS</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>RxExS</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>TxExS</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>RxTxExS</td>
<td>0.1</td>
<td></td>
</tr>
</tbody>
</table>

- **main effects**
- **interactions**

---

**Conclusions**

- upper range delimiter is best
- truncation up to the implementers
- keep users from descending the menu hierarchy
- experience is critical in menu displays

---

**You know now**

- Controlled experiments can provide clear convincing result on specific issues
- Creating testable hypotheses are critical to good experimental design
- Experimental design requires a great deal of planning
- Statistics inform us about
  - mathematical attributes about our data sets
  - how data sets relate to each other
  - the probability that our claims are correct

---

**Graphs:** whenever there are non-parallel lines, we have a potential interaction effect.
You now know

- There are many statistical methods that can be applied to different experimental designs
  - T-tests
  - Correlation and regression
  - Single factor Anova
  - Factorial Anova
- Anova terminology
  - factors, levels, cells
  - factorial design
    - between, within, mixed designs

For more information…

…I strongly recommend that you take EPSE 592: Design and Analysis in Educational Research (Educational Psychology and Special Education)