CPSC 544: EXPERIMENTS II

Prof. Narges Mahyar

© 2017 Prof. Joanna McGrenere & Dr. Leila Aflatoony
includes slides from Prof. Karon MacLean and Jessica Dawson
Learning Goals

• why are statistics used?
• What is a T-test?
• what is an analysis of variance (ANOVA)?
• what is the important terminology in ANOVA?
• what are the different types of ANOVA?
• when would one choose to use an ANOVA?
• what other statistics are relevant to HCI?

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

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

cnull 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

Which one is within-subject?  
Between-subject?
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
TWO-TAILED UNPAIRED T-TEST

n: number of data points in the one sample \((N = n_1 + n_2)\)

\[ \sum X: \text{sum of all data points in one sample} \]

\[ \bar{X}: \text{mean of data points in sample} \]

\[ \sum (X^2): \text{sum of squares of data points in sample} \]

\[ s^2: \text{combined sample variance} \]

\[ t: t \text{ ratio} \]

\[ \text{df} = \text{degrees of freedom} = n_1 + n_2 - 2 \]

Formulas

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

\[ t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

How to maximize t?
**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.

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, 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

How does the outcome change for a confidence level of 0.10?

probability that means are from the same underlying population
Recall Menu Hypotheses

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. *- 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 \( p \)
ANALYSIS OF VARIANCE (ANOVA)

- a workhorse
  - allows moderately complex experimental designs (relative to t-test)

- terminology
  - factor
    - independent variable
    - e.g., Keyboard, Expertise, Age
  - factor level
    - specific value of independent variable
    - e.g., Qwerty, novice, 10-12 year olds
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
• error (random + systematic)

between group variability (BG)
• treatment effects
• individual differences
• error (random + systematic)

these two variability's combine to give total variability
• we are mostly interested in ____________ variability because we are trying to understand the effect of the treatment
ANOVA is what we call an omnibus test

- tells us if (\(\bar{x}_1 = \bar{x}_2 = \bar{x}_3\)) IS NOT true
- doesn’t tell us HOW the means differ (i.e. \(\bar{x}_1 > \bar{x}_2\))

Intuition...

\[
f = \frac{BG}{WG} = \frac{\text{treatment} + \text{id} + \text{error}}{\text{id} + \text{error}} = ?
\]

= 1, if there are no treatment effects
> 1, if there are treatment effects

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

• similar to the t-test, we look up the f value in a table, for a given \( \alpha \) and degrees of freedom to determine significance

• thus, f statistic is sensitive to sample size
  – Big N  \( \rightarrow \)  Big Power  \( \rightarrow \)  Easier to find significance
  – Small N  \( \rightarrow \)  Small Power  \( \rightarrow \)  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
  – e.g., 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

• e.g.:
  – dependent variable: typing speed (time)
  – independent variable (factor): keyboard
  – between subject design

\[
\begin{array}{c|c|c}
\text{Qwerty} & \text{Alphabetic} & \text{Dvorak} \\
\hline
S1: & 25 \text{ secs} & S21: \quad 40 \text{ secs} & S51: \quad 17 \text{ secs} \\
S2: & 29 & S22: \quad 55 & S52: \quad 45 \\
\ldots & \ldots & \ldots & \ldots \\
S20: & 33 & S40: \quad 33 & \ldots \\
S51: & \ldots & \ldots & \ldots \\
S52: & \ldots & \ldots & \ldots \\
S60: & \ldots & \ldots & \ldots \\
\end{array}
\]

also called a one-way ANOVA
ANOVA TERMINOLOGY

• factorial design
  – cross combination of levels of one factor with levels of another
  – e.g., keyboard type (3) x expertise (2)

• Cell [or condition]
  – unique treatment combination
  – e.g., qwerty x non-typist

2-way factorial ANOVA

<table>
<thead>
<tr>
<th>Keyboard</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Qwerty</td>
<td></td>
</tr>
<tr>
<td>Dvorak</td>
<td></td>
</tr>
<tr>
<td>Alphabetic</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-typist</td>
</tr>
<tr>
<td>typist</td>
</tr>
</tbody>
</table>
ANOVA TERMINOLOGY

• mixed factor [*split-plot*]
  – contains both between and within subject combinations

```
<table>
<thead>
<tr>
<th></th>
<th>Qwerty</th>
<th>Dvorak</th>
<th>Alphabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-typist</td>
<td>S1-20</td>
<td>S1-20</td>
<td>S1-20</td>
</tr>
<tr>
<td>typist</td>
<td>S21-40</td>
<td>S21-40</td>
<td>S21-40</td>
</tr>
</tbody>
</table>
```
ANOVA

- compares the relationships between many factors
- provides more informed results
  - considers the interactions between factors
  - e.g.,
    - typists type faster on Dvorak, than on alphabetic and Qwerty
    - non-typists are fastest on alphabetic

<table>
<thead>
<tr>
<th>expertise</th>
<th>Qwerty</th>
<th>Dvorak</th>
<th>Alphabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-typist</td>
<td>S1-20</td>
<td>S1-20</td>
<td>S1-20</td>
</tr>
<tr>
<td>typist</td>
<td>S21-40</td>
<td>S21-40</td>
<td>S21-40</td>
</tr>
</tbody>
</table>
Other statistical tests commonly used in H C I

• Your reading does a very good job of covering these, and we won’t cover them further
  – Correlation
  – Regression
  – Non-parametric tests
    • Chi-squared
    • Mann-Whitney
    • Wilcoxon signed-rank
    • Kruskal-Wallis
    • Friedman’s