Introduction to HCI

Statistical Analysis

Prof. Ali Sarvghad
UMass Amherst

asarvr@cs.umass.edu
Courses, projects, papers, and more:
http://groups.cs.umass.edu/nmahyar/
Introduction

• **Inferential statistics** methods for hypothesis testing:
  • t-test
  • ANOVA

• You need to know:
  • Basics of descriptive statistics
    • Mean
    • Variance
    • Standard deviation
  • Normal distribution
  • Basics of probability
Introduction

• We will use an example of a designed experiment to talk about t-test & ANOVA
• In a designed experiment, you can establish causation
• Control some factors (independent variables) to find their effects on some other factors (dependent variables)
Example Designed Experiment

We designed a new technique for visually compare and understating changes in a hierarchical data
Example Designed Experiment

Change type: a node can be: deleted, moved, added
Example Designed Experiment

Comparison technique 1: **side by side**

Before

After

Added
Example Designed Experiment

Comparison technique 2: **reduced side by side (RSS)**
Controlled Lab Experiment

• **Goal:** Compare Side by Side, and Reduced Side by Side (RSS) techniques

• **H0:** There will be no difference between the two conditions

• **H1:** Users will be **faster** to identify the change using RSS

• **Measured**
  • Performance **time**
Target population.

It is often not possible to access and involve the entire target population in your study.
Sampling is the technique that we use when we can’t access the entire target population.

Example, a sample of size = 20
Randomly assign participants to your two experimental condition

Condition 1, Side by side, 10
Condition 2, RSS, 10
Condition 1, Side by Side

Condition 2, Reduce Side by Side

Study:

All the participants worked on the same set of tasks identifying changes in hierarchical data.

Participants in condition 1 used Side by Side to identify the changes and participants in condition 2 used Reduced Side by Side.

We collected time and accuracy per task for each participant.
Average time and accuracy for performing for the two conditions

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg.</td>
<td>36.0, 19.25</td>
</tr>
<tr>
<td>StDev.</td>
<td>21.12, 18.12</td>
</tr>
</tbody>
</table>

Is the difference between the time averages significant?
If I replicate the study, will I get the same results? t-test will tell you!
T-Test

T-test tells you the **probability (p-value)** of getting the same outcomes if you replicate your experiments with a different sample from the target population.
T-Test

Formula for independent samples, \( df = n1 + n2 - 2 \)
Side by side

Reduced side by side

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg.</td>
<td>36.0, 19.25</td>
</tr>
<tr>
<td>StDev.</td>
<td>21.12, 18.12</td>
</tr>
<tr>
<td>n</td>
<td>10, 10</td>
</tr>
</tbody>
</table>

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

\[ t \text{ value} = 1.93 \]

What does it mean?
How do we calculate the p value?
We use a number called **critical value** to decide whether we reject the null hypothesis based on our t value.

- **Our t value < critical value**, we don’t reject the null hypothesis
- **Our t value > critical value**, we reject the null hypothesis
Degrees of freedom (df)

\[ df = (n_1 + n_2) - 2 \]

\[ df = 10 + 10 - 1 = 18 \]

Significance threshold

One-tailed t-table

Degrees of freedom: [https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/](https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/)
Side by side

Reduced side by side

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg.</td>
<td>36.0, 19.25</td>
</tr>
<tr>
<td>StDev.</td>
<td>21.12, 18.12</td>
</tr>
<tr>
<td>n</td>
<td>10, 10</td>
</tr>
</tbody>
</table>

t value 1.903 > 1.734 critical value
p-value = .03476 < .05

We CAN reject the null hypothesis!
In other words, the difference between 36.0 and 19.25 is significant and not random
t-test: Important points to note

There are fundamental questions you ask before doing a t-test:

1. Is your data is normally distributed?

2. Do you have enough samples? (Ideally between 20-30)

3. Are you doing a two-tailed or one-tailed t-test?

4. Is data paired or unpaired (independent)?
Unpaired & Paired Samples

Comparing two sets of unpaired observations
Usually different subjects in each group (number may differ as well)

- Condition 1: $S_1$–$s_{20}$
- Condition 2: $s_{21}$–$s_{43}$

Paired observations
usually single group studied under separate experimental conditions
Data points of one subject are treated as a pair

- Condition 1: $S_1$–$s_{20}$
- Condition 2: $s_1$–$s_{20}$

Which one is within-subject?
Between-subject?
T-Test

The mean difference

Formula for paired samples, \( df = n - 1 \)
t-test: One-tailed

If you have **two sample means, A & B:**

You do a one-tailed test when you restrict your null hypothesis is $A < B$, $A > B$, ...

Example, the average height 8 year of boys is less than the average height of 8 year old girls.
t-test: Two-tailed

If you have **two sample means, A & B:**

You do a two-tailed test when your null hypothesize $A=B$, so you combine the possibilities of $A>B$ and $A<B$

Example, the average height 8 year of boys and girls are different.
<table>
<thead>
<tr>
<th>Degrees of freedom</th>
<th>20% (0.20)</th>
<th>10% (0.10)</th>
<th>5% (0.05)</th>
<th>2% (0.02)</th>
<th>1% (0.01)</th>
<th>0.1% (0.001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>636.619</td>
</tr>
<tr>
<td>2</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
<td>31.598</td>
</tr>
<tr>
<td>3</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
<td>12.941</td>
</tr>
<tr>
<td>4</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
<td>8.610</td>
</tr>
<tr>
<td>5</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
<td>6.859</td>
</tr>
<tr>
<td>6</td>
<td>1.440</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
<td>5.959</td>
</tr>
<tr>
<td>7</td>
<td>1.415</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
<td>5.405</td>
</tr>
<tr>
<td>8</td>
<td>1.397</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
<td>5.041</td>
</tr>
<tr>
<td>9</td>
<td>1.383</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
<td>4.781</td>
</tr>
<tr>
<td>10</td>
<td>1.372</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
<td>4.587</td>
</tr>
<tr>
<td>11</td>
<td>1.363</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
<td>4.437</td>
</tr>
<tr>
<td>12</td>
<td>1.356</td>
<td>1.782</td>
<td>2.179</td>
<td>2.681</td>
<td>3.055</td>
<td>4.318</td>
</tr>
<tr>
<td>13</td>
<td>1.350</td>
<td>1.771</td>
<td>2.160</td>
<td>2.650</td>
<td>3.012</td>
<td>4.221</td>
</tr>
<tr>
<td>14</td>
<td>1.345</td>
<td>1.761</td>
<td>2.145</td>
<td>2.624</td>
<td>2.977</td>
<td>4.140</td>
</tr>
<tr>
<td>15</td>
<td>1.341</td>
<td>1.753</td>
<td>2.131</td>
<td>2.602</td>
<td>2.947</td>
<td>4.073</td>
</tr>
<tr>
<td>16</td>
<td>1.337</td>
<td>1.746</td>
<td>2.120</td>
<td>2.583</td>
<td>2.921</td>
<td>4.015</td>
</tr>
<tr>
<td>17</td>
<td>1.333</td>
<td>1.740</td>
<td>2.110</td>
<td>2.567</td>
<td>2.898</td>
<td>3.965</td>
</tr>
<tr>
<td>18</td>
<td>1.330</td>
<td>1.734</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
<td>3.922</td>
</tr>
<tr>
<td>19</td>
<td>1.328</td>
<td>1.729</td>
<td>2.093</td>
<td>2.539</td>
<td>2.861</td>
<td>3.883</td>
</tr>
<tr>
<td>20</td>
<td>1.325</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.850</td>
</tr>
</tbody>
</table>

Two-tailed t table
Example Designed Experiment

What happen if we add a third condition to our experiment?

Comparison technique 3: **animation**

$H_0 =$ the mean of performance time is the same between three conditions
How do we compare **three means** of the three experimental conditions?

ANOVA (Analysis of Variances) is a technique that we can use to do this.

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. \(\bar{x}_1 > \bar{x}_2\))
ANOVA

Within group variability (WG)
• Participants’ differences
• Error (random + systematic)

Between group variability (BG)
• Conditions effects
• Individual differences
• Error (random + systematic)

These two variability's combine to give total variability
You want to make sure that the difference between conditions are because of the differences between the groups (BG), not the differences within the groups (WG)!
To do ANOVA, we calculate the *f statistic*

\[ f = \frac{\text{Between group variability (BG)}}{\text{Within group variability (WG)}} \]

- \( f \leq 1 \), if there are no treatment effects
- \( f > 1 \), if there are treatment effects
Cheers!
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
f statistic

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        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
  • 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

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: 25 secs</td>
<td>S21: 40 secs</td>
<td>S51: 17 secs</td>
</tr>
<tr>
<td>S2: 29</td>
<td>S22: 55</td>
<td>S52: 45</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>S20: 33</td>
<td>S40: 33</td>
<td>S60: 23</td>
</tr>
</tbody>
</table>
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
ANOVA terminology

• Mixed factor [split-plot]
  • Contains both between and within subject combinations
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
Other statistical tests commonly used in HCI

• 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