Empirical science

Ethnography

Engineering and design

Research question

Observation

Hypothesis

Descriptive
Relational
Experimental

Manipulate

A

B

Measure

Compare

Validity:
- Internal
- External

Controls

Extraneous variables

Confounding variables

Scales:
- Nominal
- Categorical
- Interval
- Ratio

Hold constant
Matching
Random assignment

media computing group
Basic Statistical Analysis for HCI

- Research Question
  - Do users type on touchscreen mobile phone faster using a stylus than using a finger?

- Between-subjects, 11 participants each

- Result
  - The choice of method had a significant effect on the completion time, $t(20) = 4.03, p < .001$.
  - Finger ($M=39.96 \; 95\% \; CI \; [25.30, \; 54.62]$) is faster than Stylus ($M=80.01 \; [65.35, \; 94.67]$). Effect size Cohens’ $d = 1.74$ (large effect).
Describing Each Condition

- **Measures of central tendency**
  - **Mean:** “average”
  - **Median:** the middle point of the sorted data

- **Measures of spread**
  - **SD:** Standard deviation
  - 95% Confidence Interval (CI)

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \quad SD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}
\]

(Different data from previous slide)

<table>
<thead>
<tr>
<th>Distributions Label=Finger</th>
<th>Data</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Dev</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper 95% Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lower 95% Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Dev</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper 95% Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lower 95% Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
</tr>
</tbody>
</table>

| 9 |
| 3.3166248 |
| 11.228139 |
| 6.7718611 |
| 11 |
| 8 |

<table>
<thead>
<tr>
<th>Distributions Label=Stylus</th>
<th>Data</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Dev</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper 95% Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lower 95% Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Dev</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper 95% Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lower 95% Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
</tr>
</tbody>
</table>

| 9 |
| 3.3166248 |
| 11.228139 |
| 6.7718611 |
| 11 |
| 9 |
Different Plots, Different Purposes

Too complex to be useful
No change as $N$ changes
Abstraction losses details
95% Confidence Interval of Mean

\[ \pm 1.96 \times \frac{SD}{\sqrt{N}} \]

- In an infinite number of experiments, 95% of the CIs will include the population mean
- Changes systematically as \( N \) change
  - Better than SD
- Report both mean and confidence interval
  - E.g., \( M = 39.96 \) 95% CI [25.30, 54.62]
Sample Size Influences Confidence

Mean ± 95% CI

N = 22

N = 44

N = 66

N = 88
Effect Size

• **Effect sizes** indicate the strength of the phenomenon
  • In experimental studies, they indicate how strong does the manipulation of independent variables results in the changes of the dependent variables.

• **Difference between two means**
  • E.g., Stylus is 40s slower than Touch
  • In original unit, intuitive

• **Percentage and ratio**
  • E.g., Stylus is twice slower than Touch
  • Emphasize the magnitude of effect
Effect Size

- **Cohen’s $d$**
  - E.g., effect size Cohen’s $d = 2.0$
  - The mean difference is roughly two SD
  - Allow comparison across different measurement units
  - Reference values:
    - 0.2 (small)
    - 0.5 (medium)
    - 0.8 (large)
  - Reporting: “Cohen’s $d = 0.25$ (small effect)”
Basic Statistical Analysis for HCI

• Research Question
  • Do users type on touchscreen mobile phone faster using a stylus than using a finger?

• Between-subjects, 11 participants each

• Result
  • The choice of method had a significant effect on the completion time, $t(20) = 4.03, p < .001$.
  • Finger ($M=39.96$ 95% CI [25.30, 54.62]) is faster than Stylus ($M=80.01$ [65.35, 94.67]). Effect size Cohens’ $d = 1.74$ (large effect).
NHST: Null Hypothesis Significance Testing

- **Assuming no effect** of IV
  - E.g., keyboard type does *not* influence completion time

- **Then $p$ value** is the probability that our measurements would occur
  - E.g., $p = 0.05$:
    - “Assuming keyboard type does *not* influence completion time, then there would be a 5% probability that our measurement turns out as it did.”

- **De facto** cutoff level of $p = .05$ for statistical significance
**t-test**

- **t ratio**: ratio between
  - Variance explained by the model (Here: mean difference $80.01 - 39.96 = 40.05$)
  - Variance that the model can’t explain (Here: Standard Error of mean difference: $9.93$)
  - $t$ ratio: $40.05 / 9.93 = 4.03$

- Theoretical probability distribution of $t$ varies by degrees of freedom

- **Degrees of freedom**: number of values that are free to vary given the statistics
  - Here: 22 participants – 2 means = 20 DOF

- **Direction of difference**
  - By default, a significant result in a t-test indicates differences without stating the direction. (known as **two-tailed tests**)
Probability Distribution of $t$

- $t(20) = 2.086$, $p = .05$
- $t(20) = 4.03$, $p = .0007$

$df = 20$
In-class Exercise: \( p \) value (Fine Prints)

• Suppose you want to compare the number of hours that people watch TV between school students and college students.
  • You gathered survey data from 100 respondents.
  • Results: On average, school students watch 3.4 hours per day, and college students watch 3.0 hours per day. \( t(98) = 1.04 \), \( p = .03 \).

• Which of the following statements are correct?
  • There are 3% probability that school students watch TV more than college students
  • There are 3% probability that school students watch TV in different amount that college students
  • Assuming that school students watch TV in different amount than college students, there is a 3% probability that this result occur.
  • Assuming that school students and college students watch TV at the same amount, there is a 3% probability that this result occur.
In-class Exercise: $p$ value (Fine Prints)

• Which of the following statements are correct?
  • There are 3% probability that school students watch TV more than college students
    \textbf{Incorrect:} not the definition of $p$-value, specifying direction of the comparison
  • There are 3% probability that school students watch TV in different amount that college students
    \textbf{Incorrect:} not the definition of $p$-value, specifying direction of the comparison
  • Assuming that school students watch TV in different amount than college students, there is a 3% probability that this result occur.
    \textbf{Incorrect:} assuming the difference in population
  • Assuming that school students and college students watch TV at the same amount, there is a 3% probability that this result occur.
    \textbf{Correct:} assuming no difference in the population and does not specify the direction
Basic Statistical Analysis for HCI

• **Research Question**
  • Do users type on touchscreen mobile phone faster using a stylus than using a finger?

• **Between-subjects, 11 participants each**

• **Result**
  • The choice of method had a significant effect on the completion time, $t(20) = 4.03, p < .001$.
  • Finger ($M=39.96$ 95% CI [25.30, 54.62]) is faster than Stylus ($M=80.01$ [65.35, 94.67]). Effect size Cohens’ $d = 1.74$ (large effect).
Statistical Assumptions

• **Normality:** distribution of sampled means are normally distributed
  • Check from the normality of the data in each group
  • Plotting data and use Shapiro-Wilk test

• **Homogeneity of variance:** sampled data from the populations of the same variance
  • Check that variance across groups are roughly equal
  • Plotting data and Leven’s test

• **Independence:** Sampled from different participants

• **Interval data**
Non-parametric Tests

- Used when normality, homogeneity of variance, or interval data assumptions are violated

- Lower statistical power
  - Need larger sample size for the same $p$-value

- E.g., Wilcoxon rank-sum test

### t Test

<table>
<thead>
<tr>
<th>Stylus-Finger</th>
<th>Assuming equal variances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference</td>
<td>40.0500</td>
</tr>
<tr>
<td>Std Err Dif</td>
<td>9.9371</td>
</tr>
<tr>
<td>Upper CL Dif</td>
<td>60.7784</td>
</tr>
<tr>
<td>Lower CL Dif</td>
<td>19.3216</td>
</tr>
<tr>
<td>Confidence</td>
<td>0.95</td>
</tr>
<tr>
<td>t Ratio</td>
<td>4.030356</td>
</tr>
<tr>
<td>DF</td>
<td>20</td>
</tr>
<tr>
<td>Prob &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Prob &gt; t</td>
<td>0.0003*</td>
</tr>
<tr>
<td>Prob &lt; t</td>
<td>0.9997</td>
</tr>
</tbody>
</table>

### Wilcoxon (Rank Sums)

| S  | Z    | Prob>|Z| |
|----|------|------|
| 175| 3.15192 | 0.0016* |
Paired Tests

• For within-subject designs (violate independence assumption)
  • E.g., paired t-tests, Wilcoxon signed rank test

• More statistical power

### t Test

<table>
<thead>
<tr>
<th></th>
<th>Stylus-Finger</th>
<th>Assuming equal variances</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Difference</strong></td>
<td>40.0500</td>
<td>4.030356</td>
</tr>
<tr>
<td><strong>Std Err Dif</strong></td>
<td>9.9371</td>
<td>20</td>
</tr>
<tr>
<td><strong>Upper CL Dif</strong></td>
<td>60.7784</td>
<td>Prob &gt; ltl 0.0007*</td>
</tr>
<tr>
<td><strong>Lower CL Dif</strong></td>
<td>19.3216</td>
<td>Prob &gt; t 0.0003*</td>
</tr>
<tr>
<td><strong>Confidence</strong></td>
<td>0.95</td>
<td>Prob &lt; t 0.9997</td>
</tr>
</tbody>
</table>

### Wilcoxon Signed Rank

<table>
<thead>
<tr>
<th></th>
<th>Finger-Stylus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S</strong></td>
<td>175</td>
</tr>
<tr>
<td><strong>Z</strong></td>
<td>3.15192</td>
</tr>
<tr>
<td>**Prob&gt;</td>
<td>Z</td>
</tr>
</tbody>
</table>

### Difference: Finger-Stylus

|                | Finger | Stylus | t-Ratio | DF | Prob > |t| Prob > t | Prob < t |
|----------------|--------|--------|---------|----|---------|---------|---------|
| **Mean Difference** | -40.05 | Prob > ltl <.0001* |
| **Std Error**       | 4.45   | Prob > t 1.0000 |
| **Upper 95%**       | -30.135| Prob < t <.0001* |
| **Lower 95%**       | -49.965| 11      |
| **N**               |        | 1       |
| **Correlation**     |        |         |

### Wilcoxon Signed Rank

<table>
<thead>
<tr>
<th></th>
<th>Finger-Stylus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Statistic S</strong></td>
<td>-33.000</td>
</tr>
<tr>
<td>**Prob&gt;</td>
<td>S</td>
</tr>
<tr>
<td><strong>Prob&gt;S</strong></td>
<td>0.9995</td>
</tr>
<tr>
<td><strong>Prob&lt;S</strong></td>
<td>0.0005*</td>
</tr>
</tbody>
</table>
Statistical Analysis So Far

Experimental design

Between

- Normality
- Homogeneity of variance
- Interval data

- Yes → t-test
- No → Wilcoxon Rank sum test

Within

- Normality
- Homogeneity of variance
- Interval data

- Yes → paired t-test
- No → Wilcoxon signed rank test
Statistical Analysis So Far

Number of Levels in the IV

- **Two**
  - **Between**
    - Normality
      - Yes: t-test
      - No: Wilcoxon
    - Homogeneity of variance
      - Yes: paired t-test
      - No: Wilcoxon signed rank test
  - **Within**
    - Interval data

- **More than two**
  - **Between**
    - Normality
      - Yes: One-way ANOVA
      - No: Kruskal-Wallis
    - Homogeneity of variance
      - Yes: Repeated-Measure ANOVA
      - No: Friedman’s ANOVA
  - **Within**
    - Interval data
Type I and Type II Error

• Each time we do a t-test ($p < .05$), we have 5% probability to be false positive
  • Probability of no false positive = 95%

• Three t-tests: $0.95^3 = 0.857$
  • Actual probability to be false positive: $1 - 0.857 = 0.143$
  • Overtesting increase probability to be false positive
**ANOVA: Analysis of Variance**

- Fit different models and determine how good the models explain the data
  - **Maximal model:** one parameter per data point
  - **Null model:** one parameter (e.g., mean) represents all data points
  - Determine just adequate candidate model that fits the data
ANOVA

- Candidate model fits better than null model ⇒ The effect is statistically significant

- Candidate model fits as well as null model ⇒ The effect is not statistically significant

- Conclusion: The differences among the levels are statistically significant

Statistically significant

E.g., $F_{2,28} = 73.07, p < .001$
Post-hoc Test

• Compare each pair of conditions as a follow-up of ANOVA
  • E.g., t-tests

• Need to prevent the false-positive

• E.g., **Bonferroni correction**: set lower cut-off for $p$-value to be significant
  • Three conditions: cut-off $0.05 / 3 = .0167$
  • Apply this cut-off to all tests
Statistical Analysis So Far

Number of Levels

- Two
  - Between
    - Normality
      - Yes
        - t-test
      - No
        - Wilcoxon Rank sum test
    - Homogeneity of variance
    - Interval data
  - Within
    - Normality
      - Yes
        - paired t-test
      - No
        - Wilcoxon signed rank test

- More than two
  - Between
    - Normality
      - Yes
        - One-way ANOVA
      - No
        - Kruskal-Wallis
  - Within
    - Normality
      - Yes
        - Repeated-Measure ANOVA
      - No
        - Friedman’s ANOVA

ANOVA is significant

Post hoc Tests
Statistical Analysis So Far

**Number of IVs**

- One
- More than one

**Number of Levels**

- Two
- More than two

**Experimental design**

- Between
  - Normally homogeneity of variance
    - Yes
      - t-test
    - No
      - Wilcoxon rank sum test
  - Not normally homogeneity of variance
    - pooled t-test
    - Wilcoxon signed rank test

- Within
  - Normally homogeneity of variance
    - Yes
      - One-way ANOVA
    - No
      - Kruskal-Wallis
  - Not normally homogeneity of variance
    - Repeated-Measures ANOVA
    - Friedman's ANOVA

**Post hoc Tests**

- ANOVA is significant

**Factorial ANOVA**

Mixed-design ANOVA

...
Reporting

• Result
  • The choice of method had a significant effect on the completion time, \( t(20) = 4.03, p < .001 \).
  • Finger (\( M = 39.96, 95\% \ CI [25.30, 54.62] \)) is faster than Stylus (\( M = 80.01, [65.35, 94.67] \)). Effect size Cohens’ \( d = 1.74 \) (large effect).

• Two-digit after the decimal point
  • Except \( p \)-value: report exact iff more than 0.001

• Use 95% confidence interval as error bar and indicate so
Reading Assignment

• Required
  • (Dragicevic et al., alt.chi 2014) Running an HCI experiment in multiple parallel universes

• Recommended
  • Practical Statistics for HCI by Jacob O. Wobbrock, U. of Washington
    Independent study material with examples from HCI
    Uses SPSS and JMP (trial version: free download)
    http://depts.washington.edu/aimgroup/proj/ps4hci/
Summary

- Effect size (mean) and their confidence interval describes the data

- Cohen’s $d$ (standardized effect size) allows comparison across experiments

- $p$-value is the probability of that the result occurs assuming no effect of IV.

- Statistical assumptions and experimental design indicate appropriate type of the test

- Overtesting increase probability to be false positive