# Statistical Analysis

**CS6501: Human-Computer Interaction**

Seongkook Heo

Fall 2020, Department of Computer Science

University of Virginia | School of Engineering & Applied Science

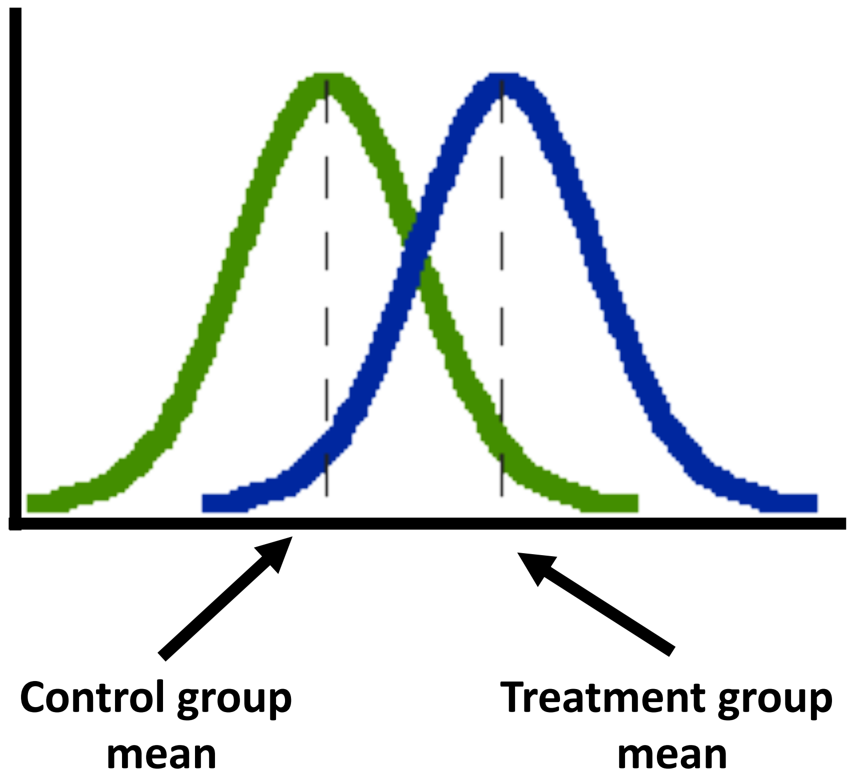
# Conducting an Experiment

# Preparing Data for Analysis

- Record the data
  - Be thorough (if possible: be able to recreate the study)
    - Small file that summarizes each trial + Large log that records everything with time stamp
  - **Check for bugs!**
- Clean the data
  - Detect errors
  - Formatting
- Remove the outliers
  - Follow guidelines
  - Be consistent

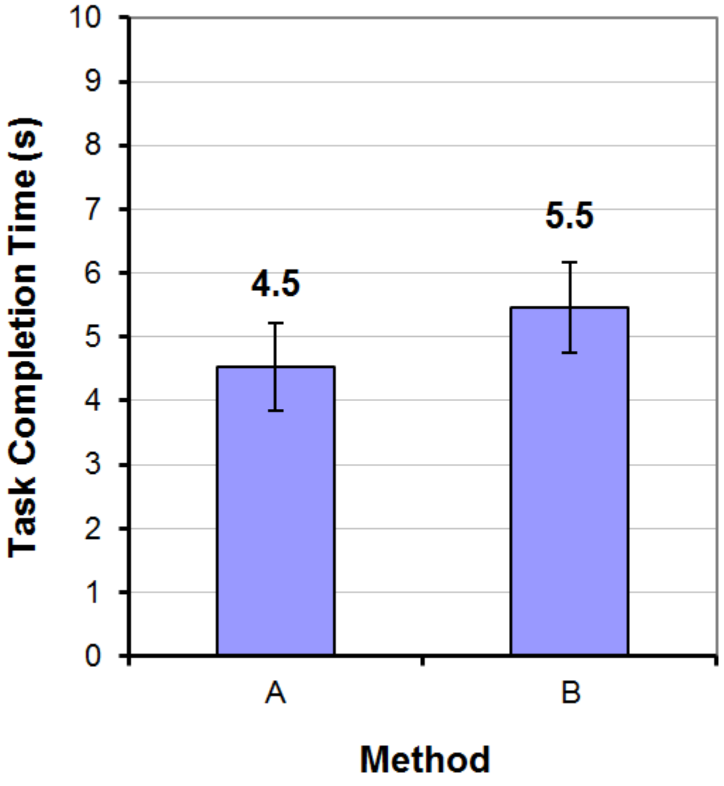
# Descriptive Statistics

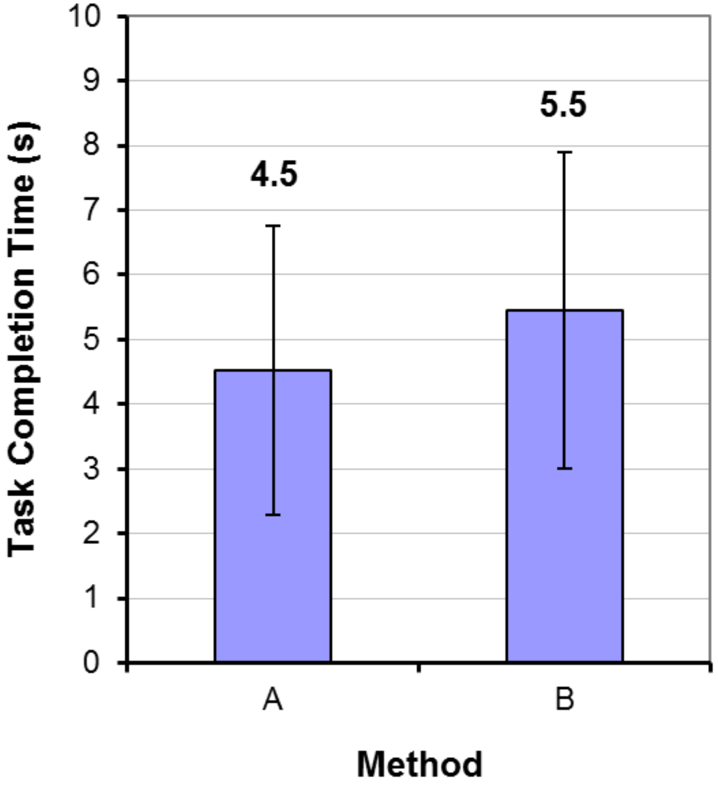
- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of spread
  - Min/Max
  - Range
  - Variance
  - Standard deviations



# Descriptive Statistics

- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of spread
  - Min/Max
  - Range
  - Variance
  - Standard deviations



# Statistical Significance

- Null Hypothesis:
  - IV x has no effect on DV y
- "P-Value":
  - Probability of obtaining your results, assuming the null hypothesis is true
- When $p < .05$
  - Reject the null hypothesis
  - IV x have an effect on DV y

# Statistical Procedures

- Two types:
  - Parametric
    - Data are assumed to come from a distribution, such as the normal distribution, *t*-distribution, etc.
  - Non-parametric
    - Data are not assumed to come from a distribution

  - A reasonable basis for deciding on the most appropriate test is to match the type of test with the measurement scale of the data

# Measurement Scales vs. Statistical Tests

| Measurement Scale | Defining Relations | Examples of Appropriate Statistics | Appropriate Statistical Tests |
|---|---|---|---|
| Nominal | • Equivalence | • Mode<br>• Frequency | • Non-parametric tests |
| Ordinal | • Equivalence<br>• Order | • Median<br>• Percentile | |
| Interval | • Equivalence<br>• Order<br>• Ratio of intervals | • Mean<br>• Standard deviation | • Parametric tests<br>• Non-parametric tests |
| Ratio | • Equivalence<br>• Order<br>• Ratio of intervals<br>• Ratio of values | • Geometric mean<br>• Coefficient of variation | |

# Which Statistical Test to Use?

- Parametric
  - Analysis of variance (ANOVA)
    - Used for ratio data and interval data
    - Most common statistical procedure in HCI research
- Non-parametric
  - Chi-square test
    - Used for nominal data
  - Mann-Whitney U, Wilcoxon Signed-Rank, Kruskal-Wallis, and Friedman tests
    - Used for ordinal data

# Which Statistical Test to Use?

| | Interval/Ratio (Normality assumed) | Interval/Ratio (Normality not assumed), Ordinal | Dichotomy (Binomial) |
|---|---|---|---|
| **Compare two unpaired groups** | Unpaired t test | Mann-Whitney test | Fisher's test |
| **Compare two paired groups** | Paired t test | Wilcoxon test | McNemar's test |
| **Compare more than two unmatched groups** | ANOVA | Kruskal-Wallis test | Chi-square test |
| **Compare more than two matched groups** | Repeated-measures ANOVA | Friedman test | Cochran's Q test |
| **Find relationship between two variables** | Pearson correlation | Spearman correlation | Cramer's V |
| **Predict a value with one independent variable** | Linear/Non-linear regression | Non-parametric regression | Logistic regression |
| **Predict a value with multiple independent variables or binomial variables** | Multiple linear/non-linear regression | | Multiple logistic regression |

http://yatani.jp/teaching/doku.php?id=hcistats:start

# Analysis of Variance

- The *analysis of variance* (ANOVA) is the most widely used statistical test for hypothesis testing in factorial experiments
- Determine if an IV has a significant effect on a DV
  - e.g., one of the test conditions is faster/slower than the other
- Remember, an IV has at least two levels

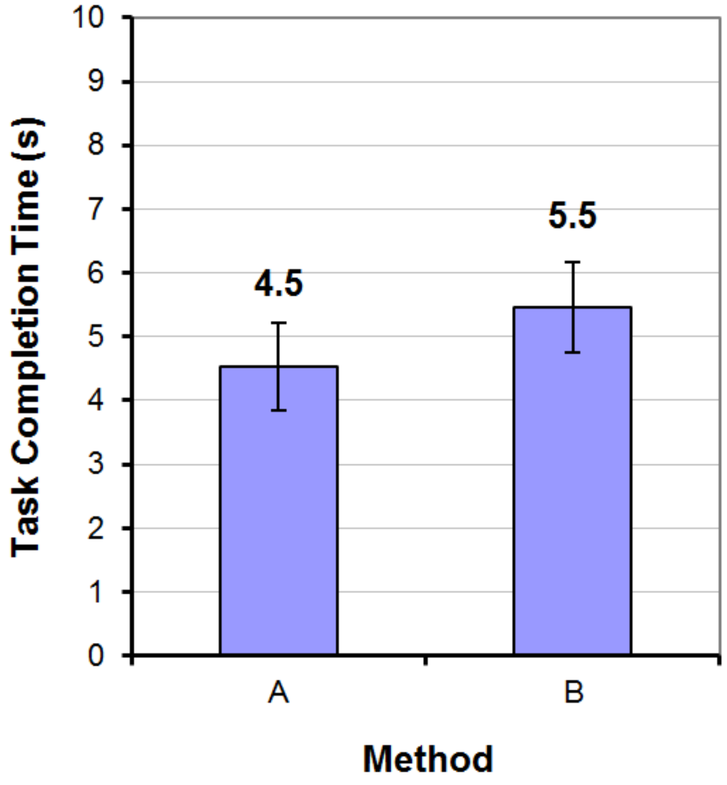
# Why Analyze the Variance?

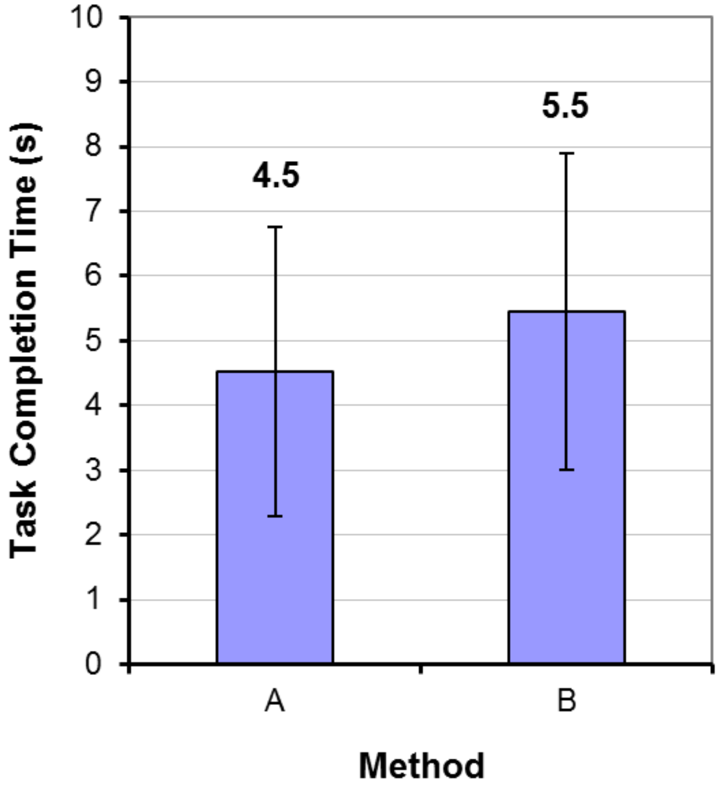
- Seems odd that we analyze the variance, when the research question is concerned with the overall means:

> Is the time to complete a task less using Method A than using Method B?

# Why Analyze the Variance?

- Two examples:

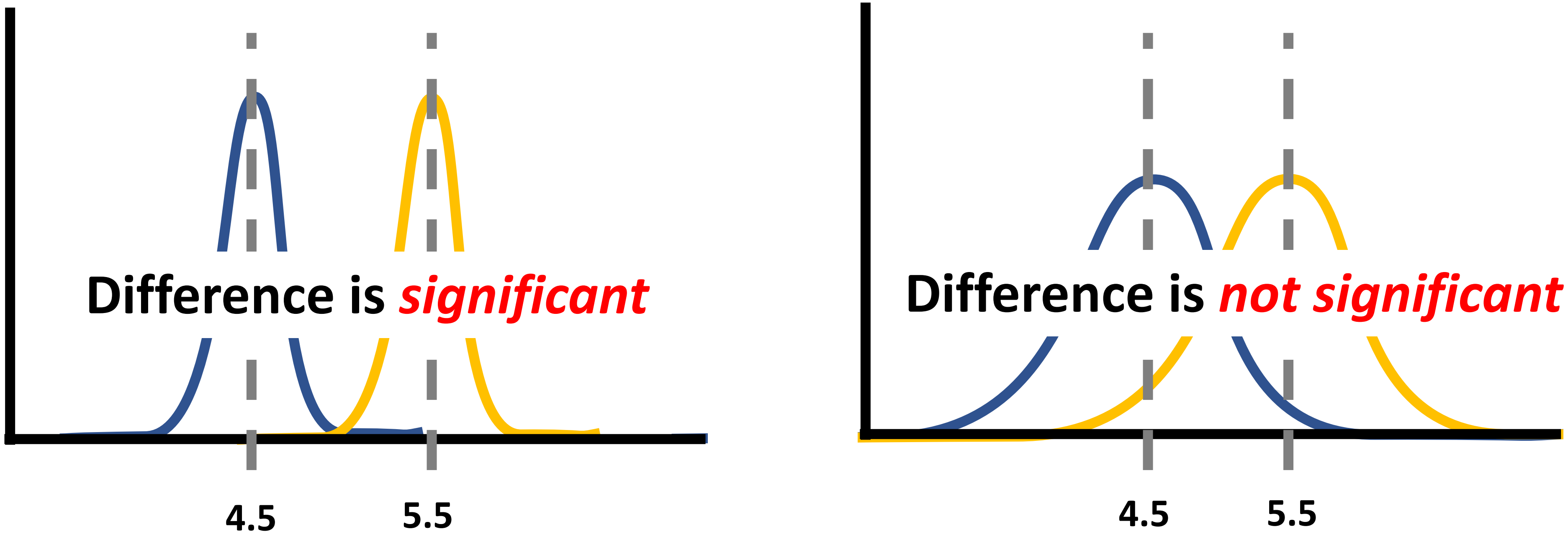



**Error bars show ±1 standard deviation**

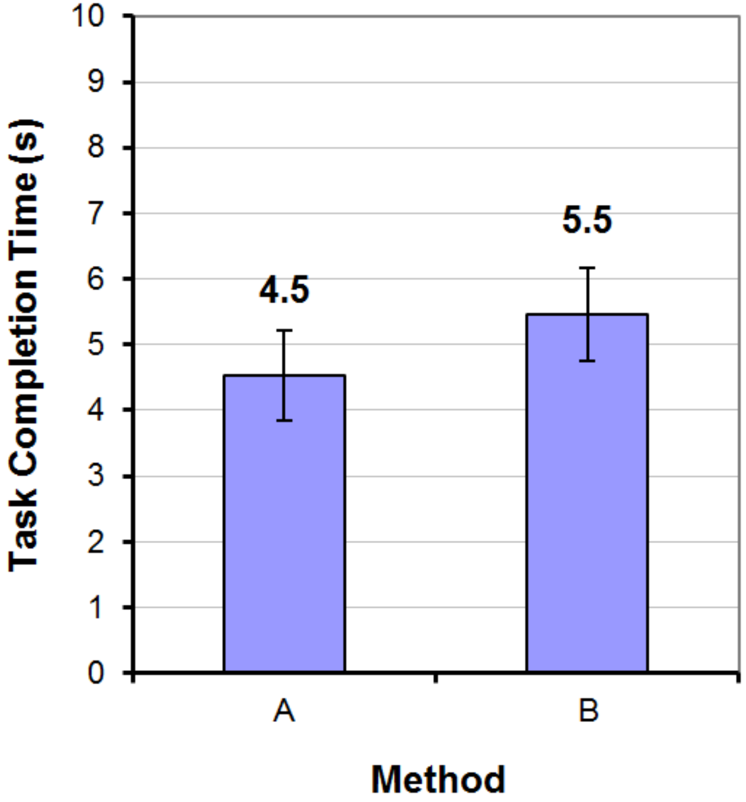
# Why Analyze the Variance?

- Two examples:




$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

# Example #1: ANOVA Analysis




| Participant | Method | |
|---|---|---|
| | A | B |
| 1 | 5.3 | 5.7 |
| 2 | 3.6 | 4.8 |
| 3 | 5.2 | 5.1 |
| 4 | 3.6 | 4.5 |
| 5 | 4.6 | 6.0 |
| 6 | 4.1 | 6.8 |
| 7 | 4.0 | 6.0 |
| 8 | 4.8 | 4.6 |
| 9 | 5.2 | 5.5 |
| 10 | 5.1 | 5.6 |
| *Mean* | 4.5 | 5.5 |
| *SD* | 0.68 | 0.72 |

# Example #1: ANOVA Analysis

**ANOVA Table for Task Completion Time (s)**

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Subject | 9 | 5.080 | .564 | | | | |
| Method | 1 | 4.232 | 4.232 | 9.796 | .0121 | 9.796 | .804 |
| Method * Subject | 9 | 3.888 | .432 | | | | |

**Probability of obtaining the observed data if the null hypothesis is true**

Reported as…

$F_{1,9} = 9.80$, $p < .05$

Thresholds for "p"

- .05
- .01
- .005
- .001
- .0005
- .0001

# Example #1: ANOVA Analysis

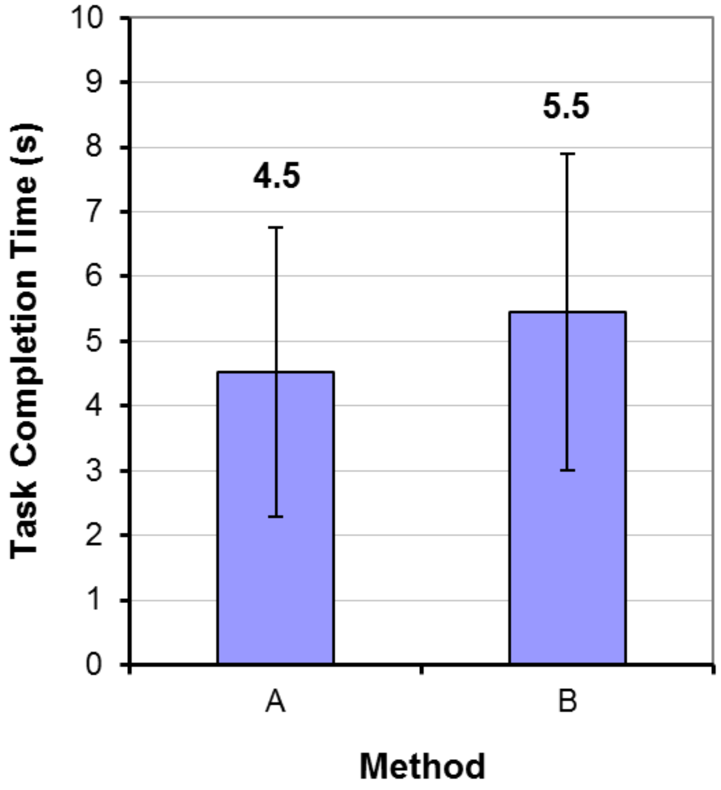
**ANOVA Table for Task Completion Time (s)**

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Subject | 9 | 5.080 | .564 | | | | |
| Method | 1 | 4.232 | 4.232 | 9.796 | .0121 | 9.796 | .804 |
| Method * Subject | 9 | 3.888 | .432 | | | | |

**Probability of obtaining the observed data if the null hypothesis is true**

The mean task completion time for Method A was 4.5s. This was 20.1% less than the mean of 5.5s observed for Method B. The difference was statistically significant ($F_{1,9} = 9.80$, $p < 0.05$)

# Example #2: ANOVA Analysis




| Participant | Method | |
|---|---|---|
| | A | B |
| 1 | 2.4 | 6.9 |
| 2 | 2.7 | 7.2 |
| 3 | 3.4 | 2.6 |
| 4 | 6.1 | 1.8 |
| 5 | 6.4 | 7.8 |
| 6 | 5.4 | 9.2 |
| 7 | 7.9 | 4.4 |
| 8 | 1.2 | 6.6 |
| 9 | 3.0 | 4.8 |
| 10 | 6.6 | 3.1 |
| *Mean* | 4.5 | 5.5 |
| *SD* | 2.23 | 2.45 |

# Example #2: ANOVA Analysis

**ANOVA Table for Task Completion Time (s)**

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Subject | 9 | 37.372 | 4.152 | | | | |
| Method | 1 | 4.324 | 4.324 | .626 | .4491 | .626 | .107 |
| Method * Subject | 9 | 62.140 | 6.904 | | | | |

**Probability of obtaining the observed data if the null hypothesis is true**

Reported as…

$F_{1,9}$ = 0.626, ns

Note: For non-significant effects, use "ns" if $F < 1.0$, or "$p > .05$" if $F > 1.0$.

# Example #2: ANOVA Analysis

**ANOVA Table for Task Completion Time (s)**

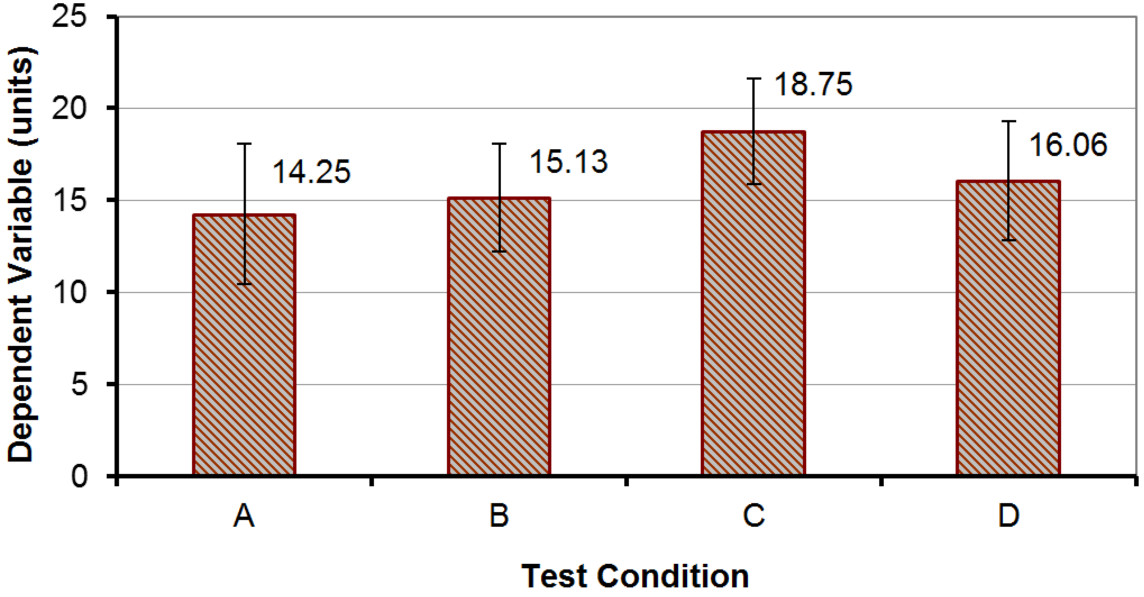
| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Subject | 9 | 37.372 | 4.152 | | | | |
| Method | 1 | 4.324 | 4.324 | .626 | .4491 | .626 | .107 |
| Method * Subject | 9 | 62.140 | 6.904 | | | | |

**Probability of obtaining the observed data if the null hypothesis is true**

The mean task completion times were 4.5s for Method A and 5.5s for Method B. As there was substantial variation in the observations across participants, the difference was not statistically significant as revealed in an analysis of variance ($F_{1,9}$ = 0.626, ns).

# More Than Two Test Conditions

| Participant | Test Condition | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 1 | 11 | 11 | 21 | 16 |
| 2 | 18 | 11 | 22 | 15 |
| 3 | 17 | 10 | 18 | 13 |
| 4 | 19 | 15 | 21 | 20 |
| 5 | 13 | 17 | 23 | 10 |
| 6 | 10 | 15 | 15 | 20 |
| 7 | 14 | 14 | 15 | 13 |
| 8 | 13 | 14 | 19 | 18 |
| 9 | 19 | 18 | 16 | 12 |
| 10 | 10 | 17 | 21 | 18 |
| 11 | 10 | 19 | 22 | 13 |
| 12 | 16 | 14 | 18 | 20 |
| 13 | 10 | 20 | 17 | 19 |
| 14 | 10 | 13 | 21 | 18 |
| 15 | 20 | 17 | 14 | 18 |
| 16 | 18 | 17 | 17 | 14 |
| *Mean* | 14.25 | 15.13 | 18.75 | 16.06 |
| *SD* | 3.84 | 2.94 | 2.89 | 3.23 |



# ANOVA

**ANOVA Table for Dependent Variable (units)**

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Subject | 15 | 81.109 | 5.407 | | | | |
| Test Condition | 3 | 182.172 | 60.724 | 4.954 | .0047 | 14.862 | .896 |
| Test Condition * Subject | 45 | 551.578 | 12.257 | | | | |

- There was a significant effect of Test Condition on the dependent variable ($F_{3,45} = 4.95$, $p < .005$)
- Degrees of freedom
  - If $n$ is the number of test conditions and $m$ is the number of participants, the degrees of freedom are…
  - Effect → $(n - 1)$
  - Residual → $(n - 1)(m - 1)$
  - Note: single-factor, within-subjects design

# Post Hoc Comparisons Tests

- A significant *F*-test means that at least one of the test conditions differed significantly from one other test condition
- Does not indicate which test conditions differed significantly from one another
- To determine which pairs differ significantly, a post hoc comparisons tests is used
- Many post hoc tests exist: Fisher PLSD, Bonferroni/Dunn, Dunnett, Tukey/Kramer, Games/Howell, Student-Newman-Keuls, orthogonal contrasts, Scheffé, etc.

# Scheffé Post Hoc Comparisons

**Scheffe for Dependent Variable (units)**
**Effect: Test Condition**
**Significance Level: 5 %**

| | Mean Diff. | Crit. Diff. | P-Value | |
|---|---|---|---|---|
| A, B | -.875 | 3.302 | .9003 | |
| A, C | -4.500 | 3.302 | .0032 | S |
| A, D | -1.813 | 3.302 | .4822 | |
| B, C | -3.625 | 3.302 | .0256 | S |
| B, D | -.938 | 3.302 | .8806 | |
| C, D | 2.688 | 3.302 | .1520 | |

**Test conditions A:C and B:C differ significantly**

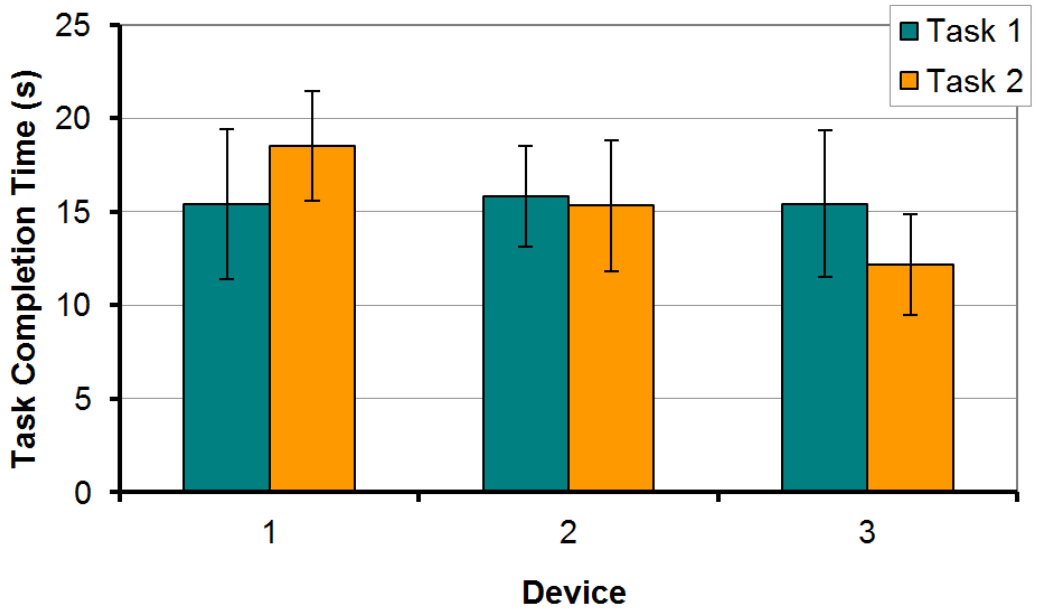
# Two-way ANOVA

- An experiment with two independent variables is a *two-way design*
- ANOVA tests for
  - Two main effects + one interaction effect
- Example
  - Independent variables
    - Device → D1, D2, D3 (e.g., mouse, stylus, touchpad)
    - Task → T1, T2 (e.g., point-select, drag-select)
  - Dependent variable
    - Task completion time
  - Both IVs assigned within-subjects
  - Participants: 12

# Two-way ANOVA Example

| Participant | Device 1 | | Device 2 | | Device 3 | |
|---|---|---|---|---|---|---|
| | Task 1 | Task 2 | Task 1 | Task 2 | Task 1 | Task 2 |
| 1 | 11 | 18 | 15 | 13 | 20 | 14 |
| 2 | 10 | 14 | 17 | 15 | 11 | 13 |
| 3 | 10 | 23 | 13 | 20 | 20 | 16 |
| 4 | 18 | 18 | 11 | 12 | 11 | 10 |
| 5 | 20 | 21 | 19 | 14 | 19 | 8 |
| 6 | 14 | 21 | 20 | 11 | 17 | 13 |
| 7 | 14 | 16 | 15 | 20 | 16 | 12 |
| 8 | 20 | 21 | 18 | 20 | 14 | 12 |
| 9 | 14 | 15 | 13 | 17 | 16 | 14 |
| 10 | 20 | 15 | 18 | 10 | 11 | 16 |
| 11 | 14 | 20 | 15 | 16 | 10 | 9 |
| 12 | 20 | 20 | 16 | 16 | 20 | 9 |
| *Mean* | 15.4 | 18.5 | 15.8 | 15.3 | 15.4 | 12.2 |
| *SD* | 4.01 | 2.94 | 2.69 | 3.50 | 3.92 | 2.69 |

| | Task 1 | Task 2 | *Mean* |
|---|---|---|---|
| Device 1 | 15.4 | 18.5 | 17.0 |
| Device 2 | 15.8 | 15.3 | 15.6 |
| Device 3 | 15.4 | 12.2 | 13.8 |
| *Mean* | 15.6 | 15.3 | 15.4 |



# ANOVA

**ANOVA Table for Task Completion Time (s)**

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Subject | 11 | 134.778 | 12.253 | | | | |
| Device | 2 | 121.028 | 60.514 | 5.865 | .0091 | 11.731 | .831 |
| Device * Subject | 22 | 226.972 | 10.317 | | | | |
| Task | 1 | .889 | .889 | .076 | .7875 | .076 | .057 |
| Task * Subject | 11 | 128.111 | 11.646 | | | | |
| Device * Task | 2 | 121.028 | 60.514 | 5.435 | .0121 | 10.869 | .798 |
| Device * Task * Subject | 22 | 244.972 | 11.135 | | | | |

# ANOVA

The grand mean for task completion time was 15.4 seconds. Device 3 was the fastest at 13.8 seconds, while device 1 was the slowest at 17.0 seconds. The main effect of device on task completion time was statistically significant ($F_{2,22} = 5.865$, $p < .01$). The task effect was modest, however. Task completion time was 15.6 seconds for task 1. Task 2 was slightly faster at 15.3 seconds; however, the difference was not statistically significant ($F_{1,11} = 0.076$, ns). The results by device and task are shown in Figure x. There was a significant Device × Task interaction effect ($F_{2,22} = 5.435$, $p < .05$), which was due solely to the difference between device 1 task 2 and device 3 task 2, as determined by a Scheffé post hoc analysis.

# How to run Statistical Tests?

- Use statistical analysis software
  - Microsoft Excel (Not recommend)
  - IBM SPSS
  - SAS JMP
  - R
  - Matlab
  - Python
  - GoStats (http://www.yorku.ca/mack/GoStats/)

# Same Stats, Different Graphs:

## Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

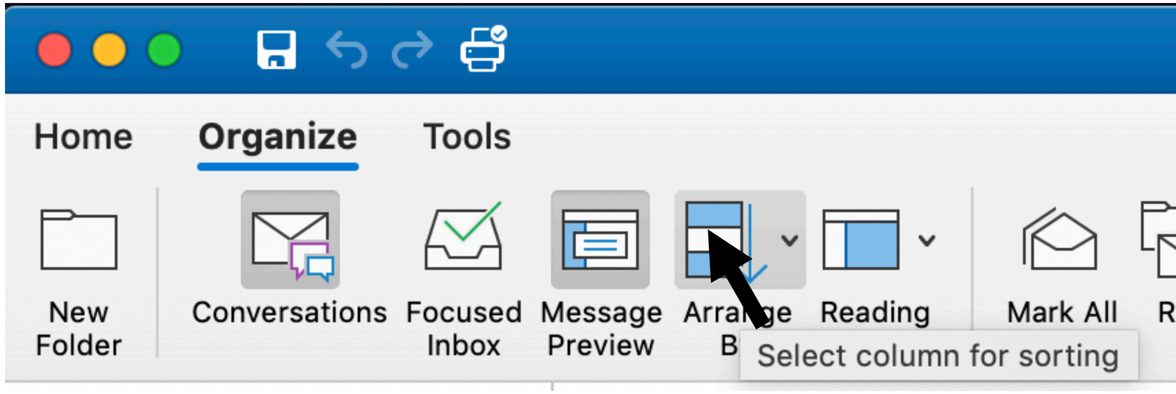
Justin Matejka
George Fitzmaurice

AUTODESK RESEARCH

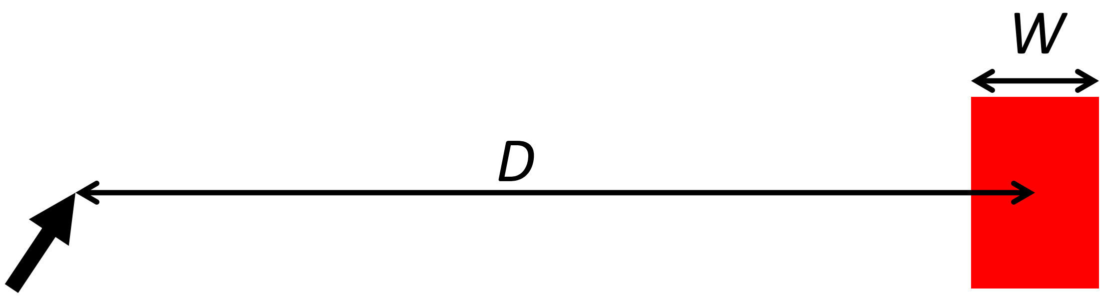
https://www.autodeskresearch.com/publications/samestats

# Pointing Device Evaluation

- Real task: interacting with GUI's
  - Pointing is fundamental

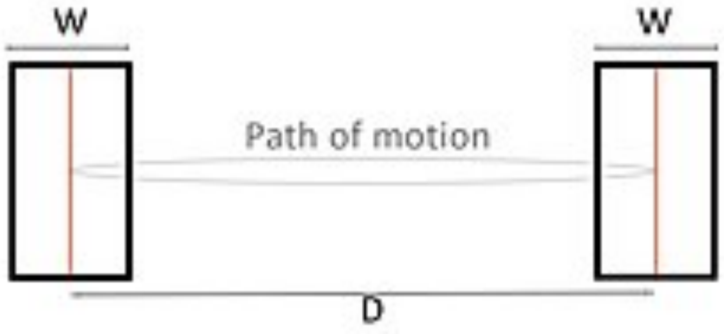



- Experimental task: target acquisition
  - Abstract, elementary, essential



# Fitts' Law

- One of the most widely used models in HCI
- Model for rapid aimed movements
- Explains relationship between the difficulty of selecting a target and the time to complete the selection
- Index of difficulty: $\mathrm{ID} = \log_2 \left( \frac{2D}{W} \right)$




- Visual demonstration: http://simonwallner.at/ext/fitts/

# Assignment #1: Quantitative Evaluation

- Use GoFitts software
  (http://www.yorku.ca/mack/FittsLawSoftware/doc/index.html?GoFitts.html)
- Choose two pointing devices of your choice:
  e.g., Touchpad and Mouse
- Run an experiment with four participants
- Measure the throughput for each device
- Report should include:
  - Experiment design
  - Experiment results
  - Your reflections on the study

**Due Oct 5 (Mon) 23:59 pm**
**Instruction will be on Collab**

Thank you!