Current Topics in Media Computing and HCI

Understanding Statistics in HCI Research

Krishna Subramanian
Media Computing Group
RWTH Aachen University

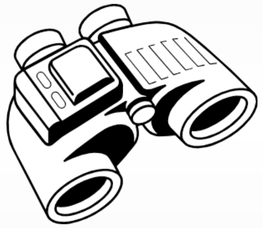
Summer Semester 2018

http://hci.rwth-aachen.de/cthci

RWTH AACHEN UNIVERSITY

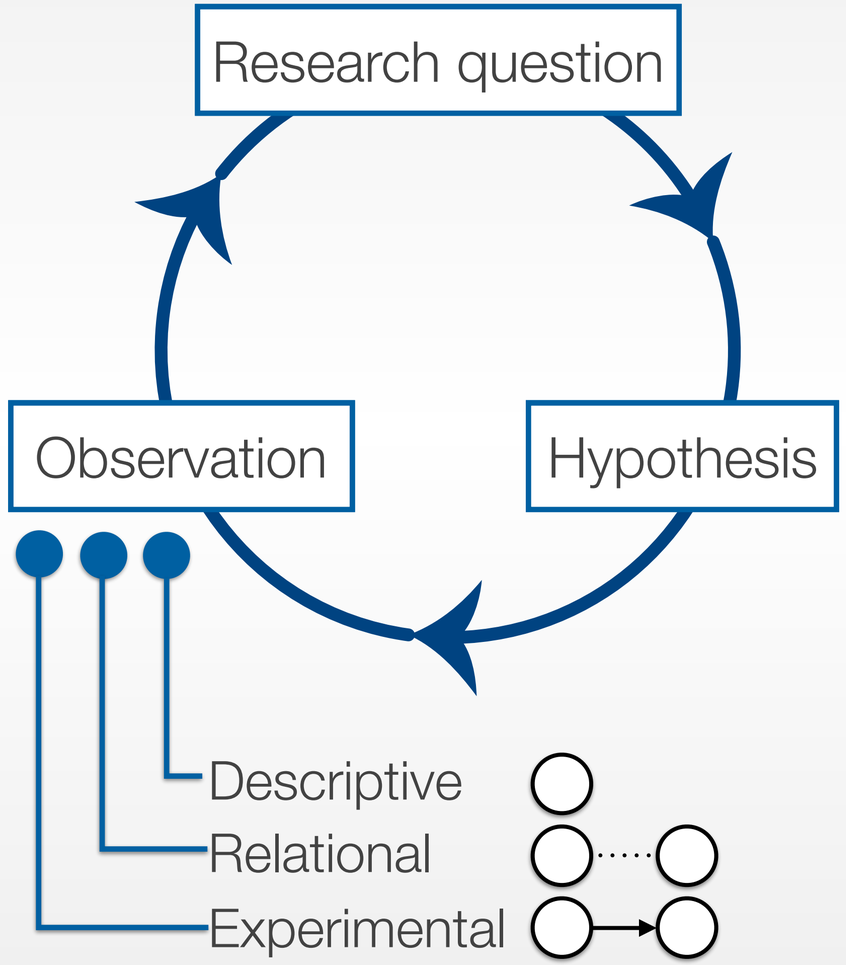
Way Back in Current Topics…

Empirical science



Ethnography

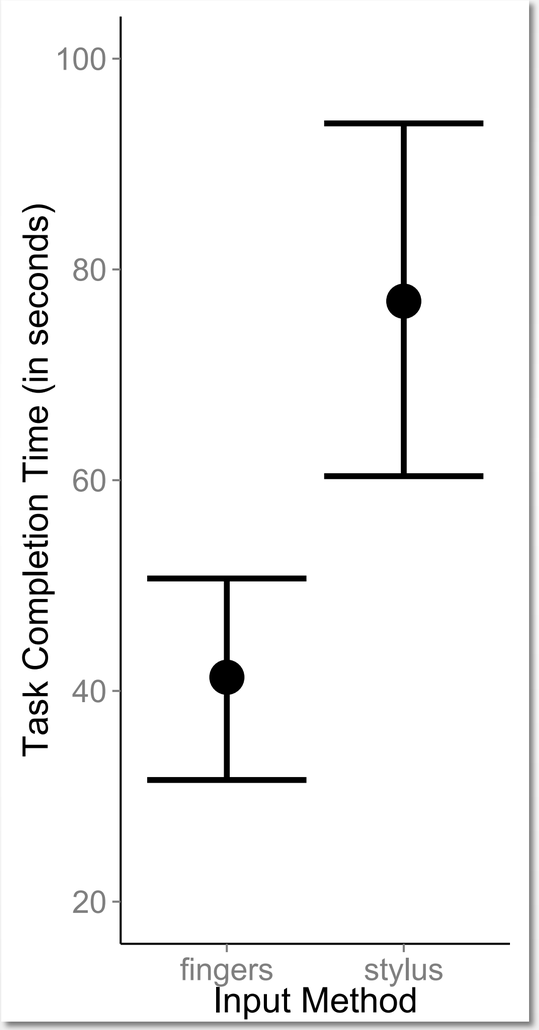
Engineering and design



RWTH AACHEN UNIVERSITY

In A Research Paper…

- The input method (fingers, stylus) had a significant effect on the task completion time, $t(20) = 4.03$, $p < .001$.
- Finger ($M = 42.03$ s; 95% CI [31.78, 52.22]) is faster than Stylus ($M = 76.21$ s; 95% CI [59.40, 93.02]). Difference between the means is 34.18 s.



Scenario: Comparing Input Methods for Typing

Fingers

Stylus

RWTH AACHEN UNIVERSITY

Steps in Experimental Research

1. Formulate hypothesis
2. Design experiment, pick dependent & independent variables, and limit extraneous variables
3. Recruit subjects
4. Run experiment (to collect data which you will analyze)
5. Perform statistical analysis on the collected data to accept or reject hypothesis

RWTH AACHEN UNIVERSITY

1. Formulate hypothesis
2. Design experiment, pick dependent & independent variables, and limit extraneous variables
3. Recruit subjects
4. Run experiment (to collect data which you will analyze)
5. Perform statistical analysis on the collected data to accept or reject hypothesis

- **Null hypothesis (H_0):** The typing speed when using fingers is not different from the typing speed when using a stylus.
- **Alternative hypothesis (H_1):** The typing speed when using fingers is different from the typing speed when using a stylus.

RWTH AACHEN UNIVERSITY

1. Formulate hypothesis
2. Design experiment, pick dependent & independent variables, and limit extraneous variables
3. Recruit subjects
4. Run experiment (to collect data which you will analyze)
5. Perform statistical analysis on the collected data to accept or reject hypothesis

- Experimental design: **Between-subjects design**
- Variables
  - Independent variable (IV): Input method with levels *fingers* and *stylus*
  - Dependent variable (DV): Task completion time (in seconds)
- Control other variables (user experience, model of the smartphone/tablet, etc.)

RWTH AACHEN UNIVERSITY

1. Formulate hypothesis
2. Design experiment, pick dependent & independent variables, and limit extraneous variables
3. Recruit subjects
4. Run experiment (to collect data which you will analyze)
5. Perform statistical analysis on the collected data to accept or reject hypothesis

RWTH AACHEN UNIVERSITY

Raw Data

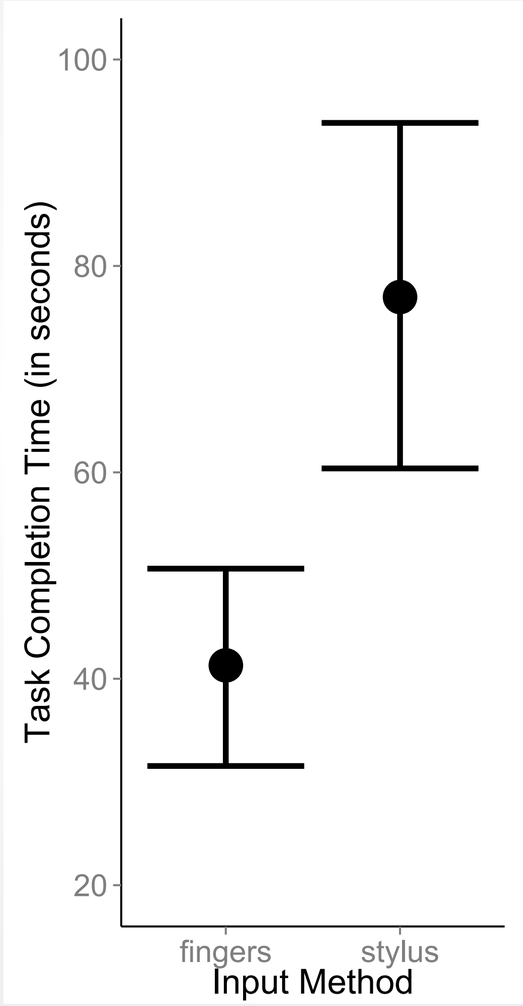
participant_ID, input_type, typing_speed
1, fingers, 70
2, stylus, 90
3, fingers, 50
4, stylus, 60
5, fingers, 90
6, stylus, 85
…

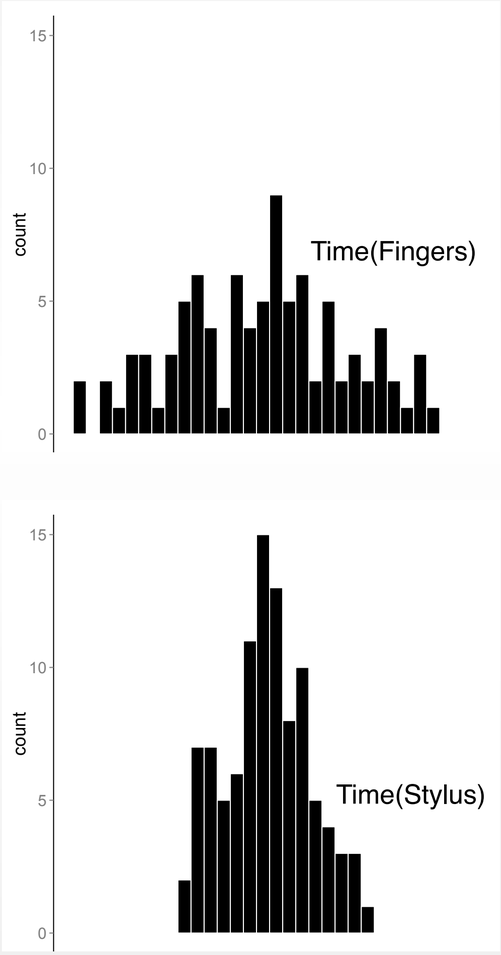
Hard to analyze!

RWTH AACHEN UNIVERSITY

Descriptive Statistics & Visualizations

- Measures of central tendency
  - Mean, median, and mode
- Measures of spread
  - Variance and standard deviation

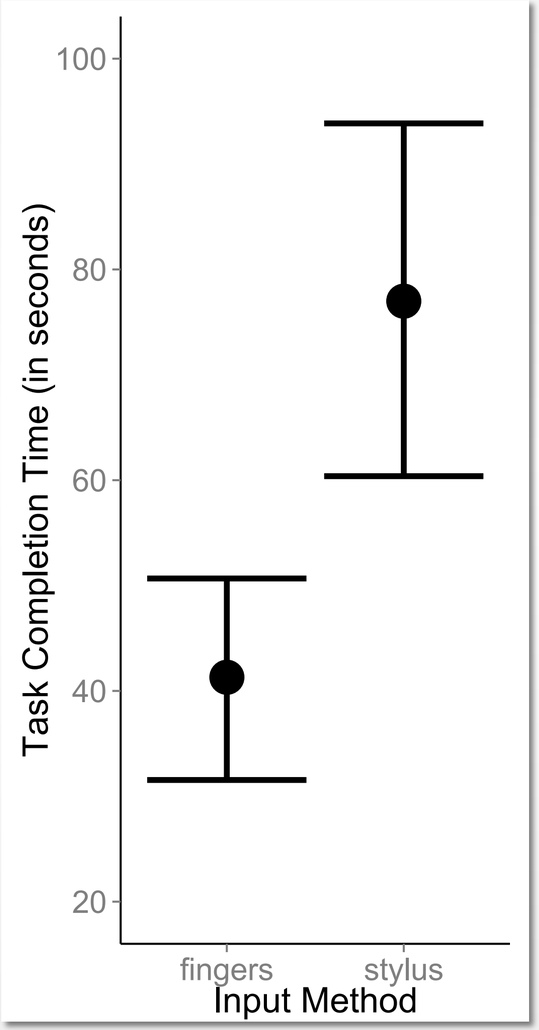




RWTH AACHEN UNIVERSITY

Result of Statistical Analysis

- The input method (fingers, stylus) had a significant effect on the task completion time, $t(20) = 4.03$, $p < .001$.
- Finger (**M = 42.03 s**; 95% CI [31.78, 52.22]) is faster than Stylus (**M = 76.21 s**; 95% CI [59.40, 93.02]). Difference between the means is 34.18 s.



RWTH AACHEN UNIVERSITY

Descriptive Statistics & Visualizations

+ Get a summary of data
+ Detect patterns in data

— Findings valid only for sample, not for the population

RWTH AACHEN UNIVERSITY

Statistical Significance Testing

RWTH AACHEN UNIVERSITY

Statistical Significance Testing

Is there a difference between the distributions at the population level?

RWTH AACHEN UNIVERSITY

Null Hypothesis Significance Testing (NHST)

- Commonly used method for significance testing
- Difference in means between sampled distributions

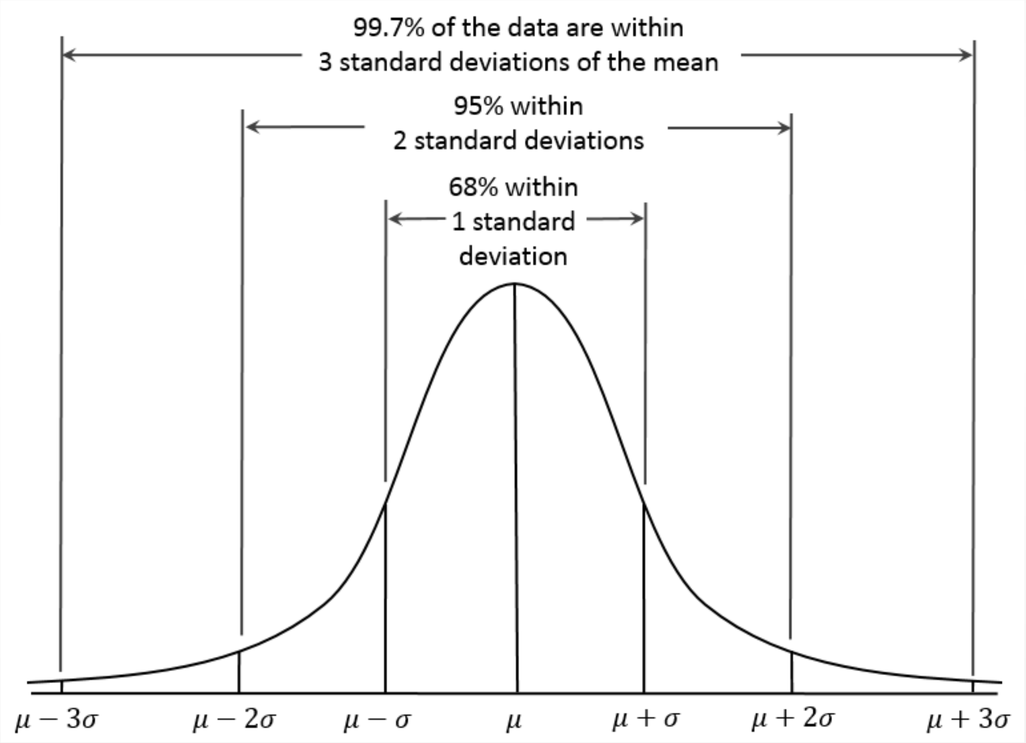
  => difference in the populations (significant difference)

  => no difference in populations, difference is due to random chance (sampling)
- Purpose of NHST: To tell these two differences apart

RWTH AACHEN UNIVERSITY

Normal Distributions

- Characteristic “bell-shape” of the distribution
- Central Limit Theorem
  - “mean of a sample will be closer to the mean of the population as the sample size increases”
  - “means of various samples of the population will follow a normal distribution”
  - Usually, a sample size of 30 is adequate



RWTH AACHEN UNIVERSITY

Null Hypothesis Significance Testing (NHST)

- Assume H_0 to be true (i.e., no difference at the population level)
- Conduct the experiment and collect data
- Fit a statistical model (e.g., a normal distribution) to the data
- Compute *p*-value, which is defined as:
  - “The chances of obtaining the experimental data we’ve collected assuming the null hypothesis is true”

RWTH AACHEN UNIVERSITY

Null Hypothesis Significance Testing (NHST)

- *De facto* cutoff level of $p = 0.05$ for statistical significance
  - **$p \leq 0.05$** => **reject** H_0 (and accept H_1)
  - **$p > 0.05$** => **accept** H_0

RWTH AACHEN UNIVERSITY

In-Class Exercise: *p*-value

- Which of the following statements are correct?

A. There is a 3% probability that school students watch TV more than college students

B. There is a 3% probability that school students watch TV in a different amount than college students

C. Assuming that school students watch TV in different amount than college students, there is a 3% probability that this result occurs

D. Assuming that school students and college students watch TV in the same amount, there is a 3% probability that this result occurs

RWTH AACHEN UNIVERSITY

In-Class Exercise: *p*-value

- Which of the following statements are correct?

A. There is a 3% probability that school students watch TV more than college students

B. There is a 3% probability that school students watch TV in a different amount than college students

C. Assuming that school students watch TV in different amount than college students, there is a 3% probability that this result occurs

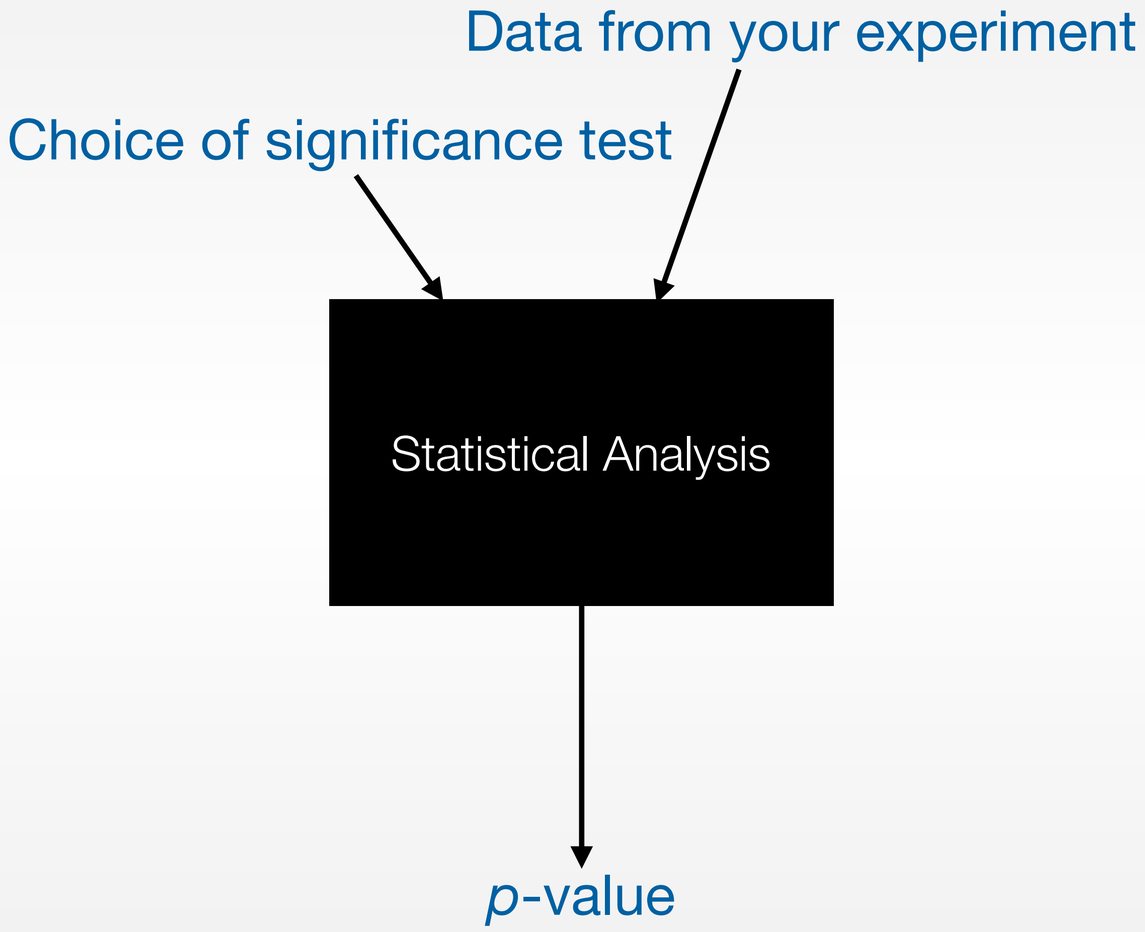
D. Assuming that school students and college students watch TV in the same amount, there is a 3% probability that this result occurs -> **Correct**

RWTH AACHEN UNIVERSITY

A Few Words on NHST

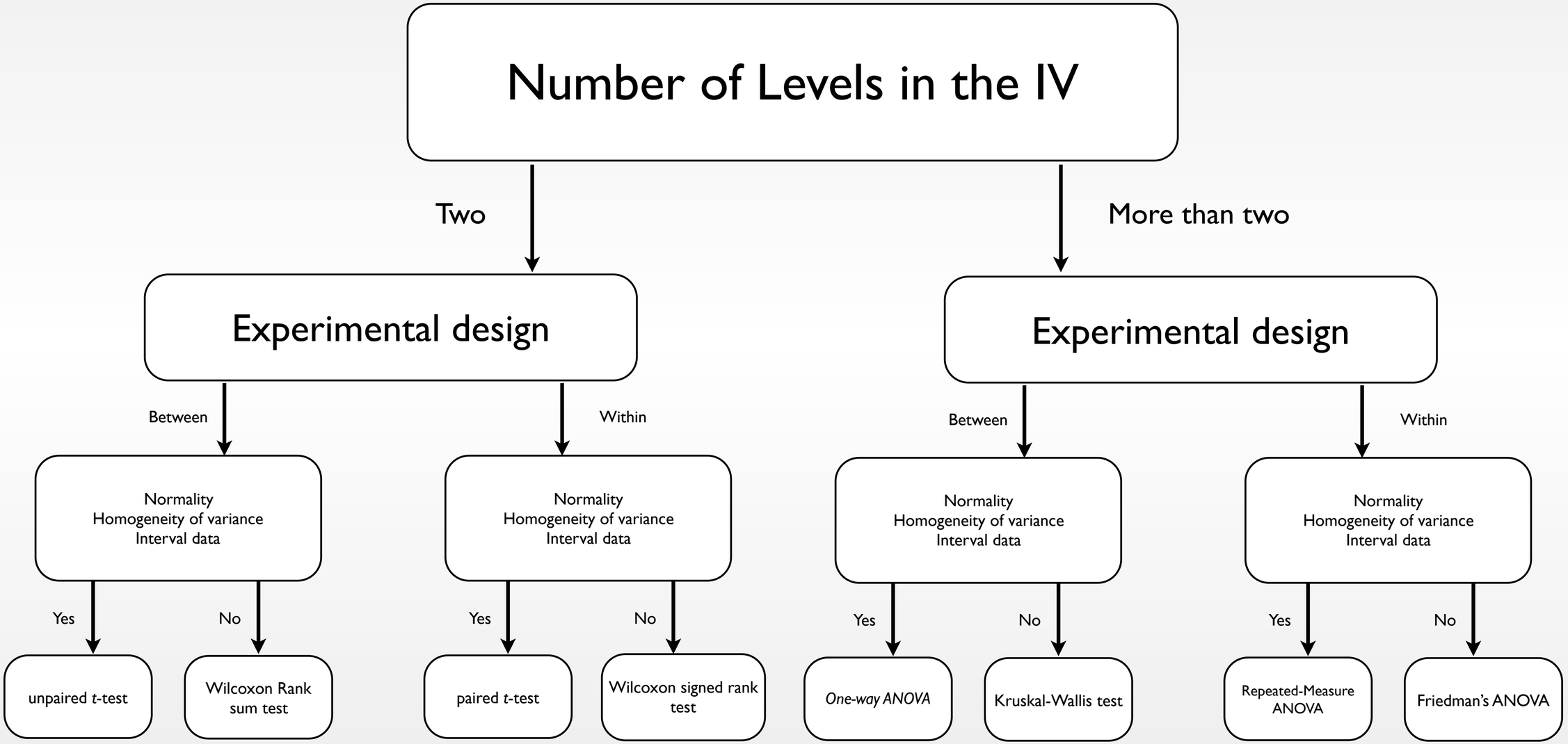
- **Test statistic:** A measure of how well our data fits a statistical model (e.g., *t*-distribution, *F*-distribution, etc.)
- *p*-value is computed from test statistic
- *p*-value is sensitive to sample sizes

RWTH AACHEN UNIVERSITY



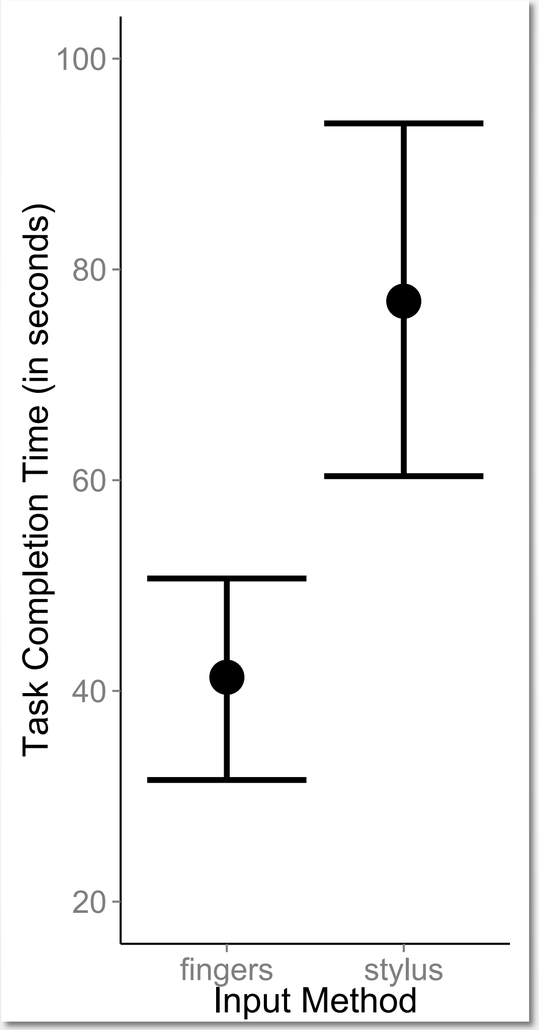
RWTH AACHEN UNIVERSITY

Do the Authors Use the Correct Test?



Result of Statistical Analysis

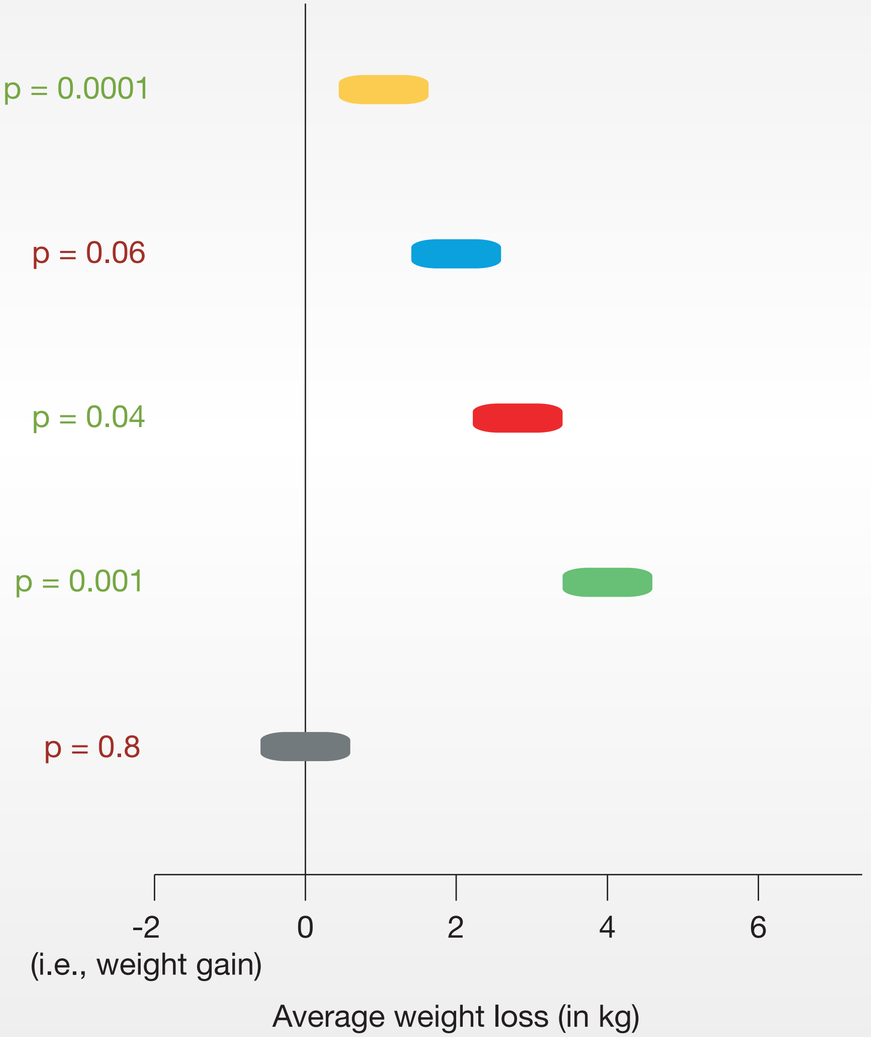
- The input method (fingers, stylus) had a significant effect on the task completion time, ***t*(20) = 4.03, *p* < .001**.
- Finger (*M* = 42.03 s; 95% CI [31.78, 52.22]) is faster than Stylus (*M* = 76.21 s; 95% CI [59.40, 93.02]). Difference between the means is 34.18 s.



Statistically Significant = Practically Significant?

RWTH AACHEN UNIVERSITY

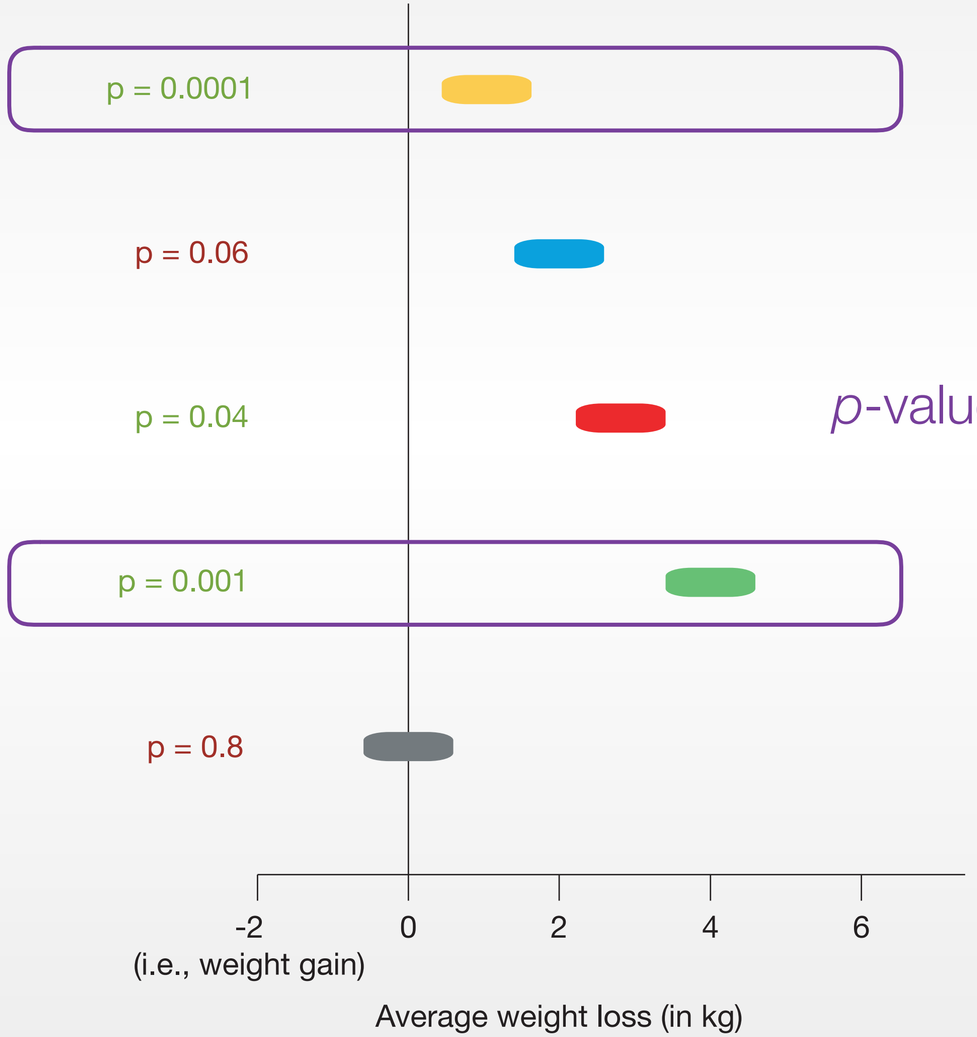
Scenario: Weight Loss via Pills



Adopted from Ziliak and Mccloskey, 2009

RWTH AACHEN UNIVERSITY

Scenario: Weight Loss via Pills



p-values do not help with interpretation!

Adopted from Ziliak and Mccloskey, 2009

RWTH AACHEN UNIVERSITY

Effect Size

- *p*-value: Is there a difference between distributions at the population level?
  - **But:** Statistically significant ($p < 0.05$) != practically significant
- Need a measure of how the big the difference is (= effect size)

RWTH AACHEN UNIVERSITY

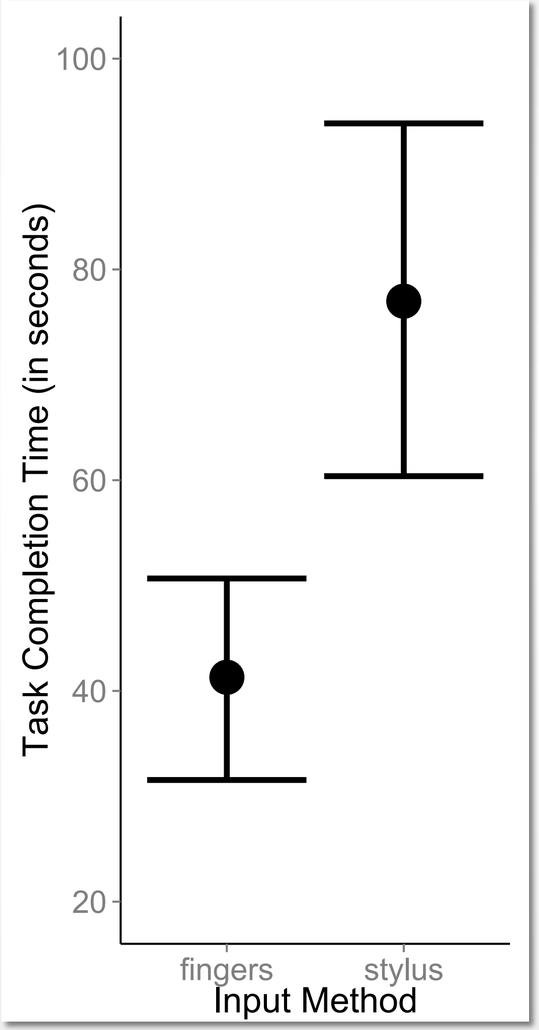
Effect Size: Examples

- 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
- Difference between means has a measurement unit (e.g., seconds, points, etc.) and therefore requires domain knowledge

RWTH AACHEN UNIVERSITY

Result of Statistical Analysis

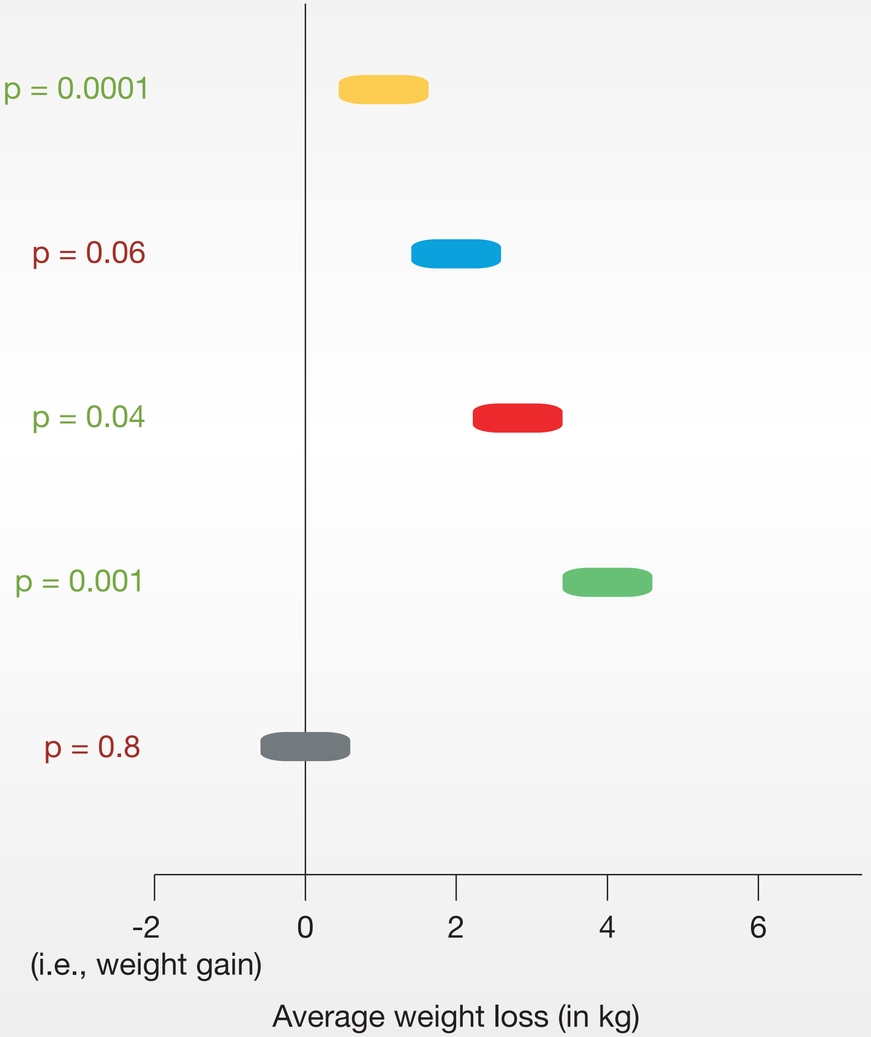
- The input method (fingers, stylus) had a significant effect on the task completion time, $t(20) = 4.03$, $p < .001$.
- Finger (*M* = 42.03 s; 95% CI [31.78, 52.22]) is faster than Stylus (*M* = 76.21 s; 95% CI [59.40, 93.02]). **Difference between the means is 34.18 s**.



How Confident Are We with Our Findings?

RWTH AACHEN UNIVERSITY

Scenario: Weight Loss via Pills

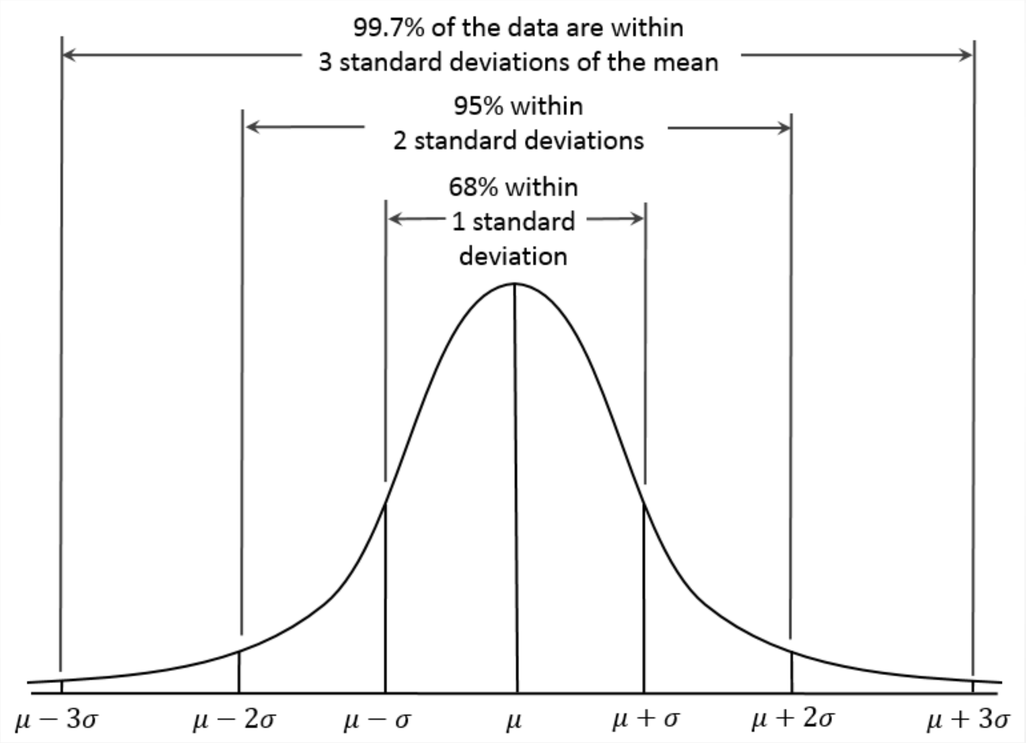


Adopted from Ziliak and Mccloskey, 2009

RWTH AACHEN UNIVERSITY

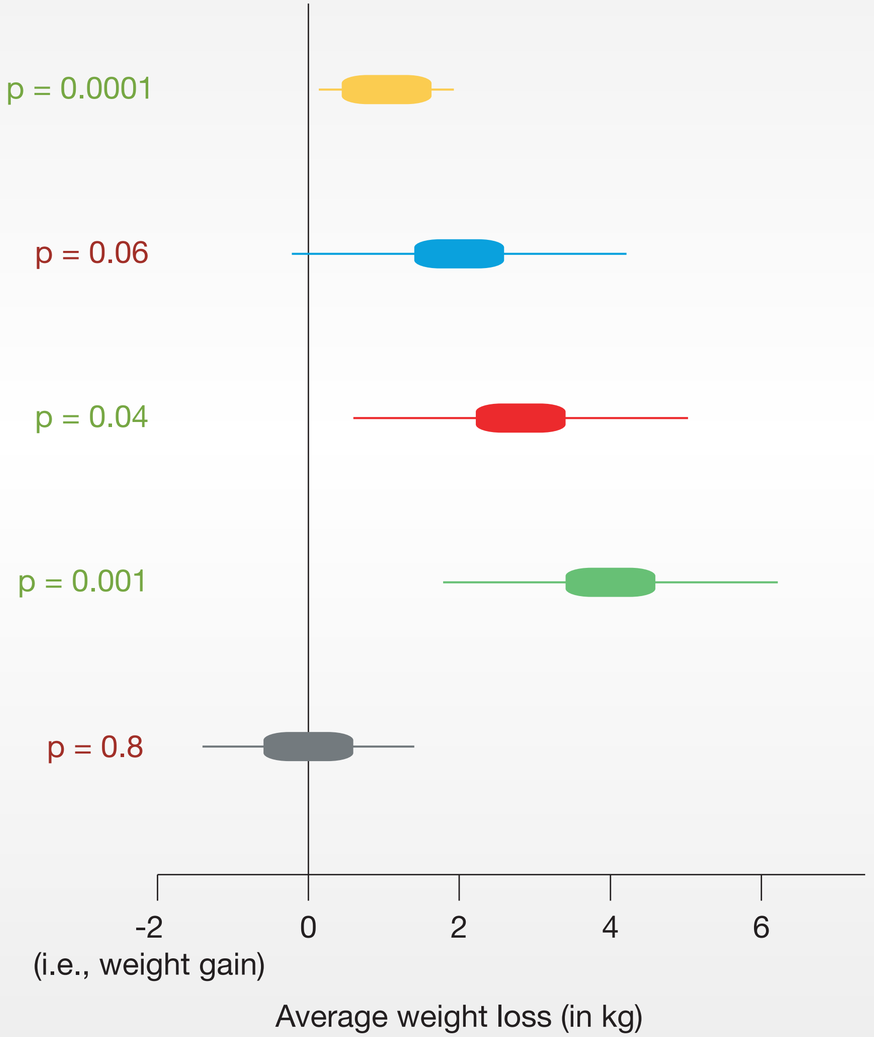
Normal Distributions

- Characteristic “bell-shape” of the distribution
- Central Limit Theorem
  - “mean of a sample will be closer to the mean of the population as the sample size increases”
  - “means of various samples of the population will follow a normal distribution”
  - Usually, a sample size of 30 is adequate



RWTH AACHEN UNIVERSITY

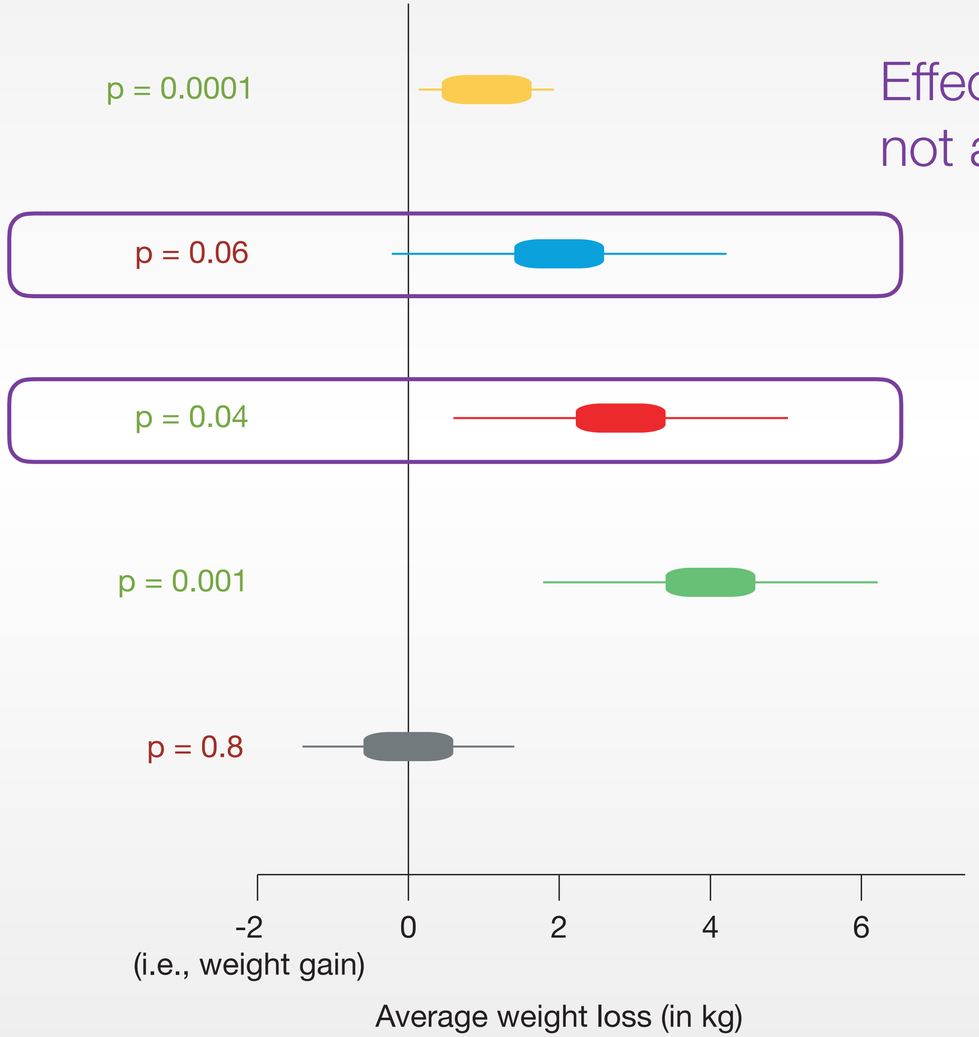
Scenario: Weight Loss via Pills



Adopted from Ziliak and Mccloskey, 2009

RWTH AACHEN UNIVERSITY

Scenario: Weight Loss via Pills



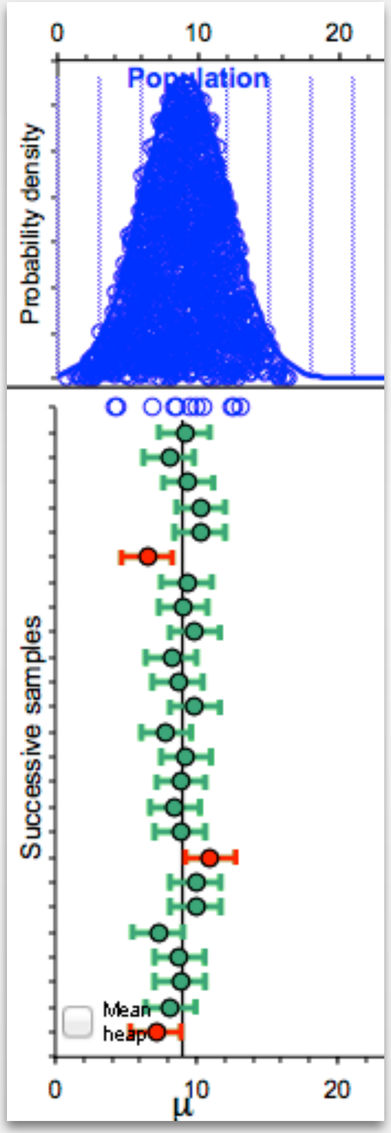
Effect sizes by themselves are not adequate for interpretation!

Adopted from Ziliak and Mccloskey, 2009

RWTH AACHEN UNIVERSITY

95% Confidence Interval

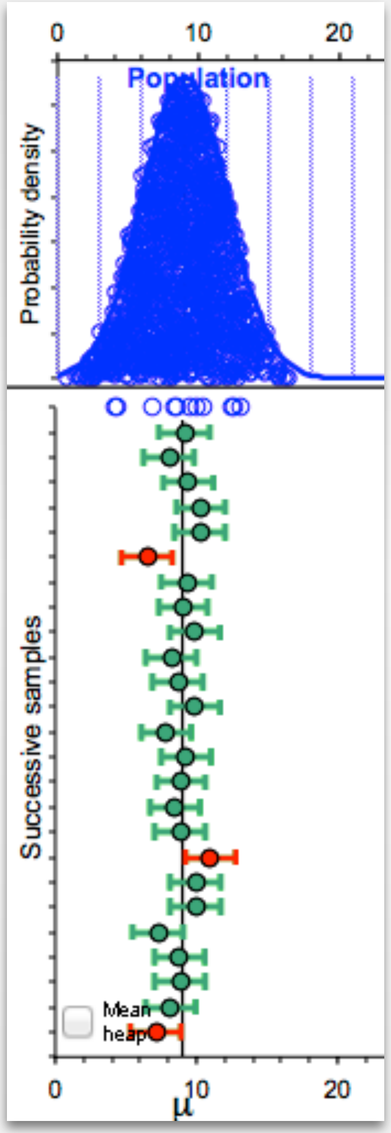
- An interval estimate (i.e., a range) of the population mean
- In an infinite number of experiments, 95% of the time, the 95% CIs will contain the population mean
- 95% is a convention, might vary across domains (e.g., medicine, psychology have different conventions)



RWTH AACHEN UNIVERSITY

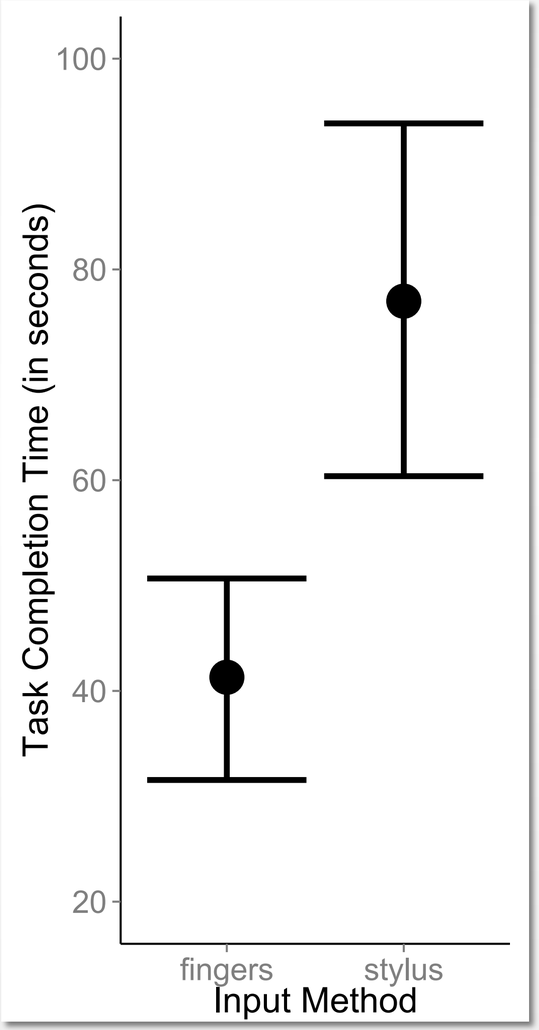
95% Confidence Interval

- Report both mean and confidence interval
  - E.g., M = 39.96 95% CI [25.30, 54.62]



Result of Statistical Analysis

- The input method (fingers, stylus) had a significant effect on the task completion time, $t(20) = 4.03$, $p < .001$.
- Finger (*M* = 42.03 s; **95% CI [31.78, 52.22]**) is faster than Stylus (*M* = 76.21 s; **95% CI [59.40, 93.02]**). Difference between the means is 34.18 s.



Required Reading

- (Cumming and Finch, American Psychologist 2005) *Inference by Eye: Confidence Intervals and How to Read Pictures of Data*
- (Delmas et al., 2005) *Using Assessment Items To Study Students' Difficulty Reading and Interpreting Graphical Representations of Distributions*
  - An exercise sheet on interpreting graphs (named "In-Class Exercise 1 — Interpreting Graphs.pdf") will be uploaded to L2P (not graded).

RWTH AACHEN UNIVERSITY

Recommended Reading (1/2)

- Statistical Methods for HCI Research by Koji Yatani, U. of Tokyo
  - Link: http://yatani.jp/teaching/doku.php?id=hcistats:start
- Practical Statistics for HCI by Jacob O. Wobbrock, U. of Washington
  - Uses SPSS and JMP (trial version available for free download)
  - Link: http://depts.washington.edu/aimgroup/proj/ps4hci/
- In-class demo of CI jumping: http://www.latrobe.edu.au/psychology/research/research-areas/cognitive-and-developmental-psychology/esci/understanding-the-new-statistics
  - Chapters 1-4, CIJumping tab

RWTH AACHEN UNIVERSITY

Recommended Reading (2/2)

- How to compute 95% CI
  http://www.stat.yale.edu/Courses/1997-98/101/confint.htm
- How to report statistics in thesis/research papers:
  http://my.ilstu.edu/~jhkahn/apastats.html (APA style)
- Issues with statistical analysis:
  Dunlop, M. D., & Baillie, M. (2009). Paper rejected (p> 0.05): an introduction to the debate on appropriateness of null-hypothesis testing. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, *1*(3), 86-93.
- Alternative approaches:
  - The New Statistics: Cumming, G. (2013). The New Statistics. *Psychological Science*, *25*(1), 7–29. http://doi.org/10.1177/0956797613504966
  - Bayesian analysis: Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press.
- Last Week Tonight with John Oliver — Scientific Studies
  https://youtu.be/0Rnq1NpHdmw

RWTH AACHEN UNIVERSITY

Summary

- We need statistical analysis to establish causal relationship between our IV and DV
- Raw data is hard to analyze
- Descriptive statistics (central tendency, spread) summarize data, but one can't make statements about the population
- NHST can be used to accept or reject null hypothesis
- Effect size quantifies the effect of IV on DV
- Confidence intervals help deal with uncertainty in data

RWTH AACHEN UNIVERSITY