Statistical Analysis in HCI Research

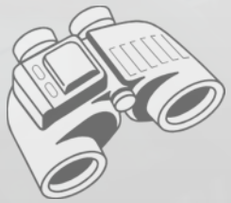
Krishna Subramanian
Media Computing Group
RWTH Aachen University

SS 2015

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

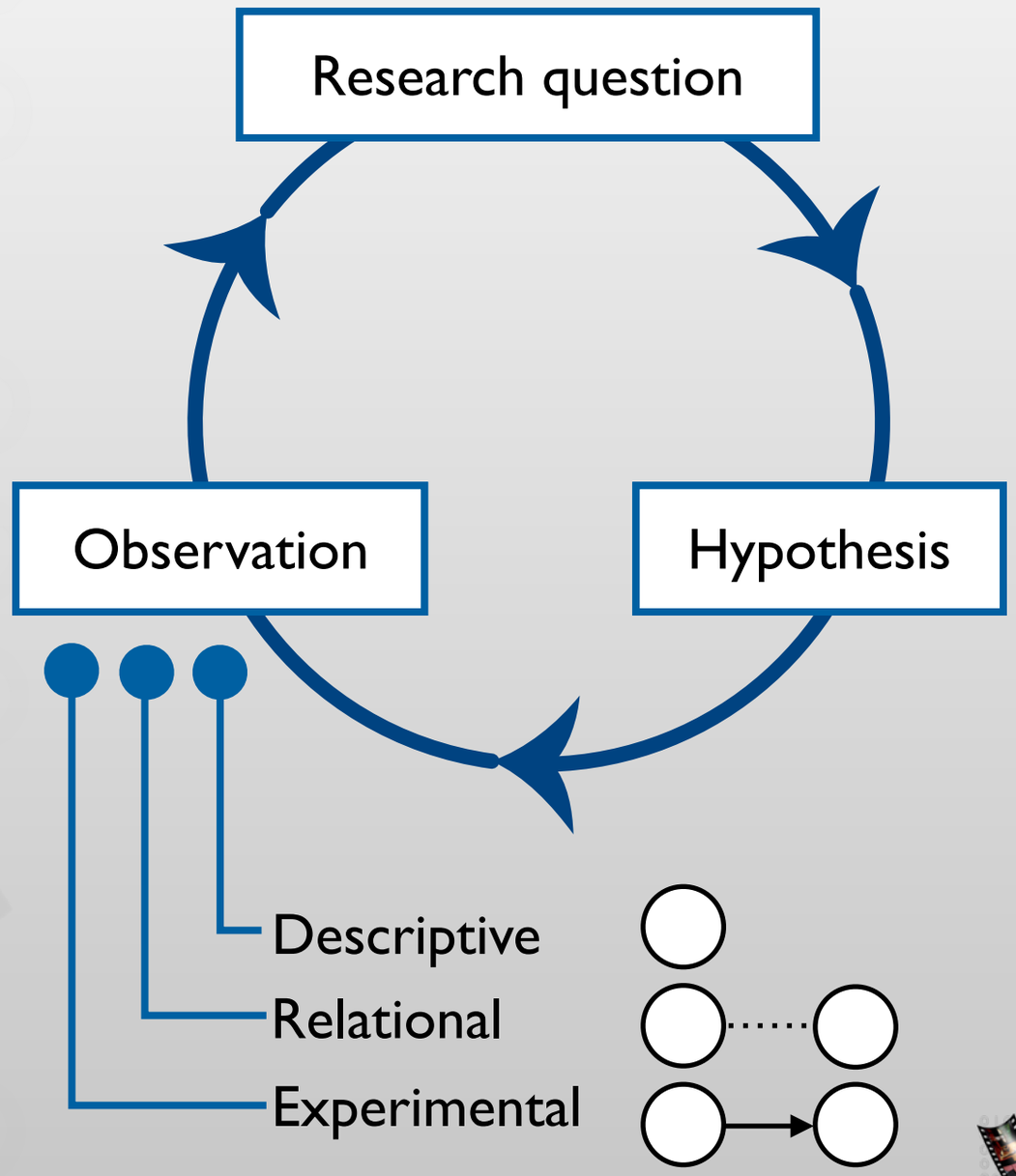
Way Back in Current Topics…

Empirical science

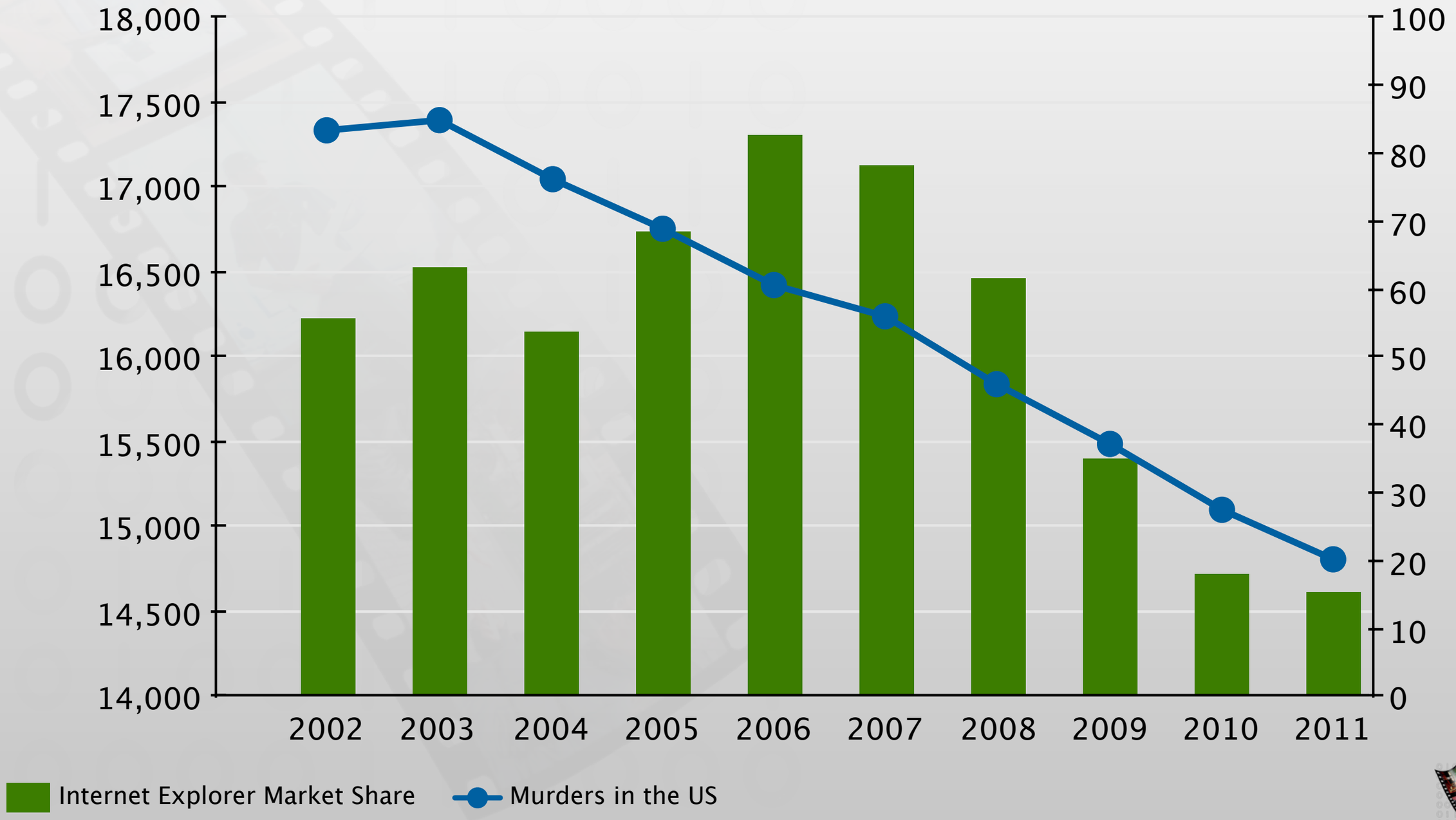


Ethnography

Engineering and design



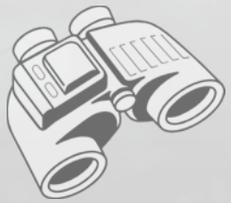
Correlation Does Not Imply Causation!



Adapted from a tweet of @altonncf with data from FBI and W3Schools

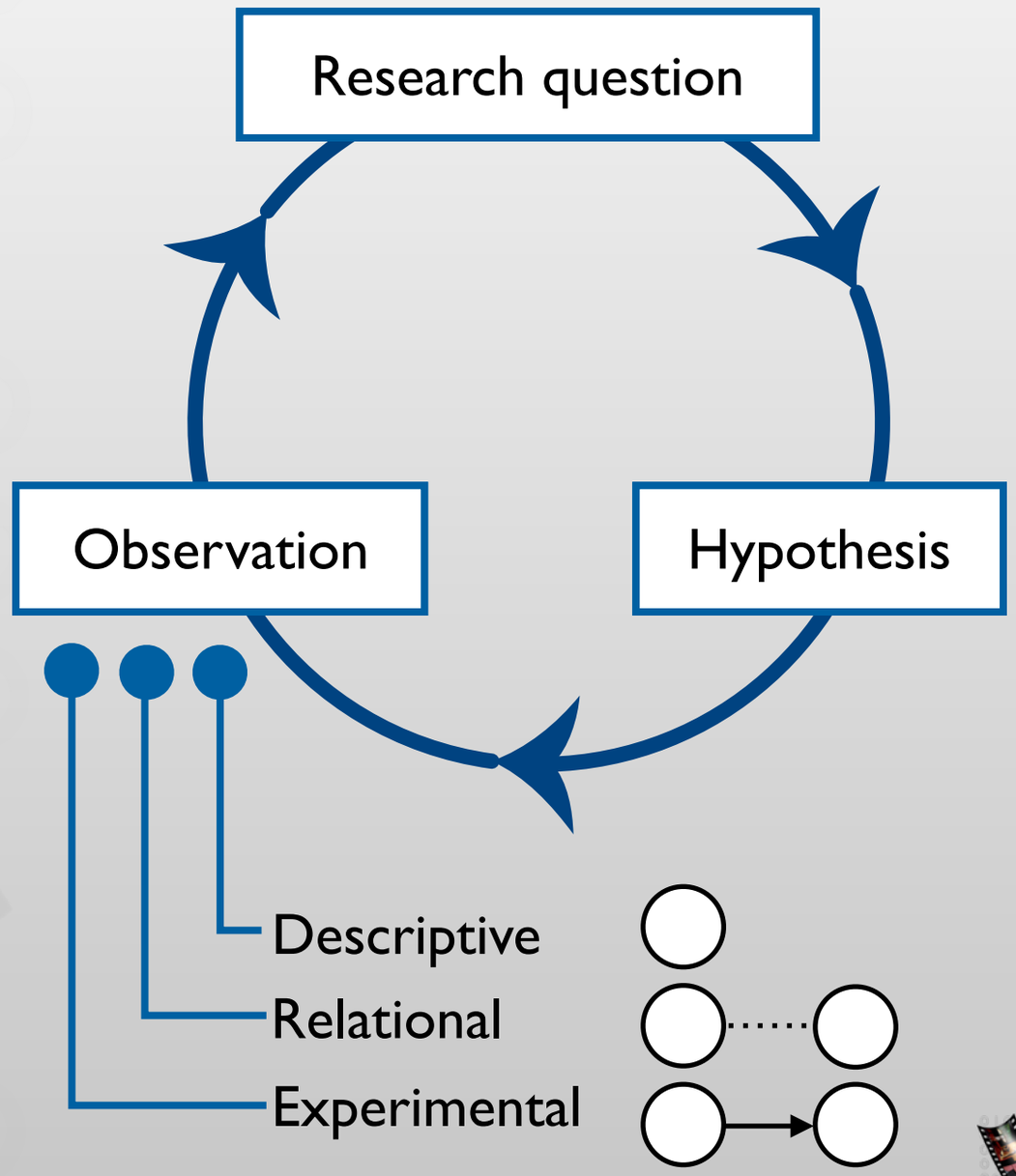
Way Back in Current Topics…

Empirical science



Ethnography

Engineering and design



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. Interpret results to accept or reject hypothesis by using statistical analysis

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. Interpret results to accept or reject hypothesis by using statistical analysis

- *"Users type on a touchscreen mobile phone faster using fingers than using a stylus."*
- Note that it is a binary statement – it can either be true or false.

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. Interpret results to accept or reject hypothesis by using statistical analysis

- Pick experimental design: between-subjects design
- Pick variables
  - Independent variable: input method (fingers, stylus)
  - Dependent variable: task completion time (in seconds)
- Design experiment so that other variables (user experience, model of mobile phone, …) are controlled

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. Interpret results to accept or reject hypothesis by using statistical analysis

- 11 participants for each condition: fingers, stylus
  - remember that we use between-subjects design

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. Interpret results to accept or reject hypothesis by using statistical analysis

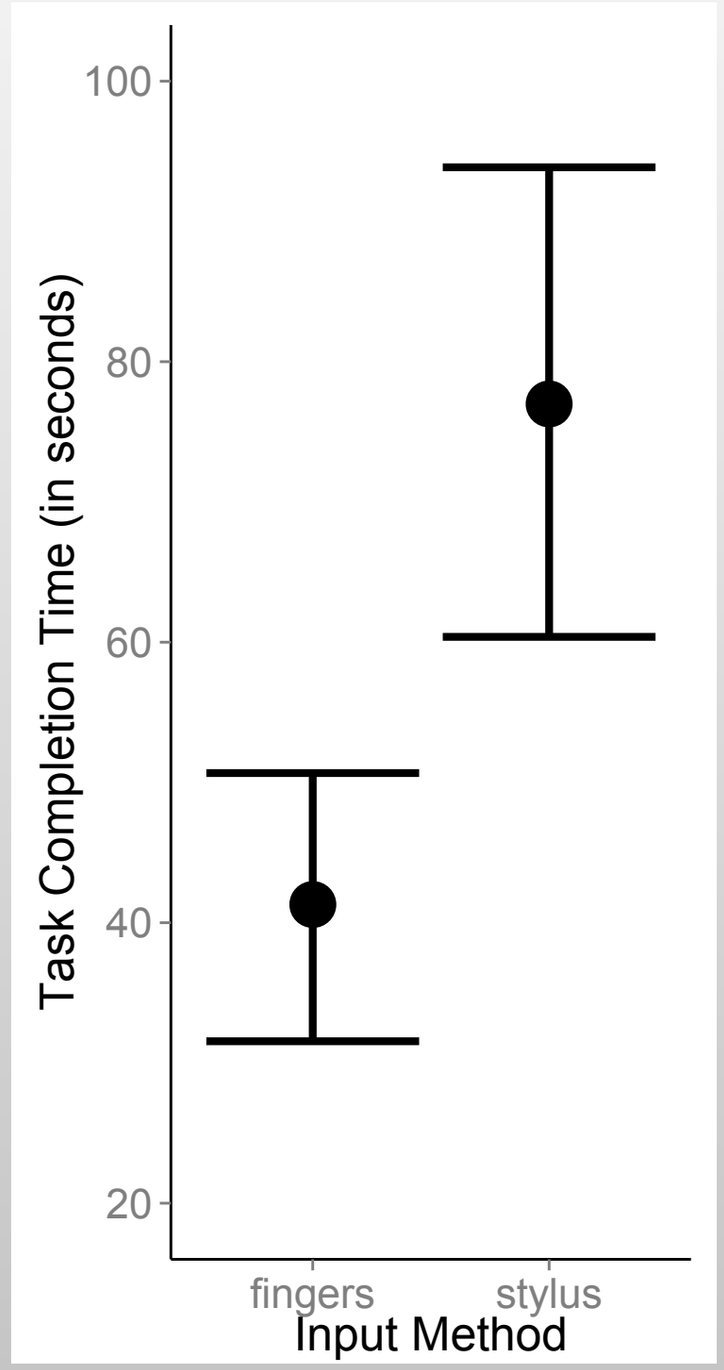
- Collect data from the study for analysis

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. Interpret results to accept or reject hypothesis by using statistical analysis

- We perform statistical analysis to show whether our hypothesis (step 1) is true or false
  - *"Users type on a touchscreen mobile phone faster using fingers than using a stylus"*
- We compare two distributions
  - Task completion time when using fingers vs. task completion time when using stylus
- How do we compare distributions?
  - By approximating them using models

Result of Statistical Analysis

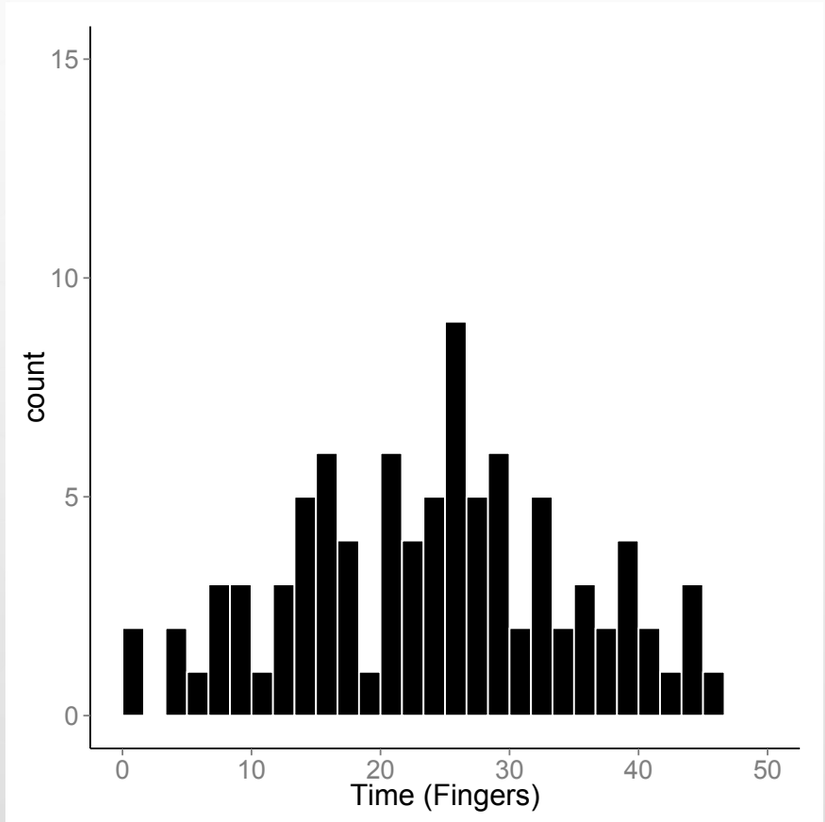
- The input method (fingers, stylus) had a significant effect on the task completion time, $t(20) = 4.03$, $p < .001$.
- Finger (M = 42.03 seconds 95% CI [31.78, 52.22]) is faster than Stylus (M =76.21 seconds [59.40, 93.02]). Difference between means (effect size) = 34.18 seconds.



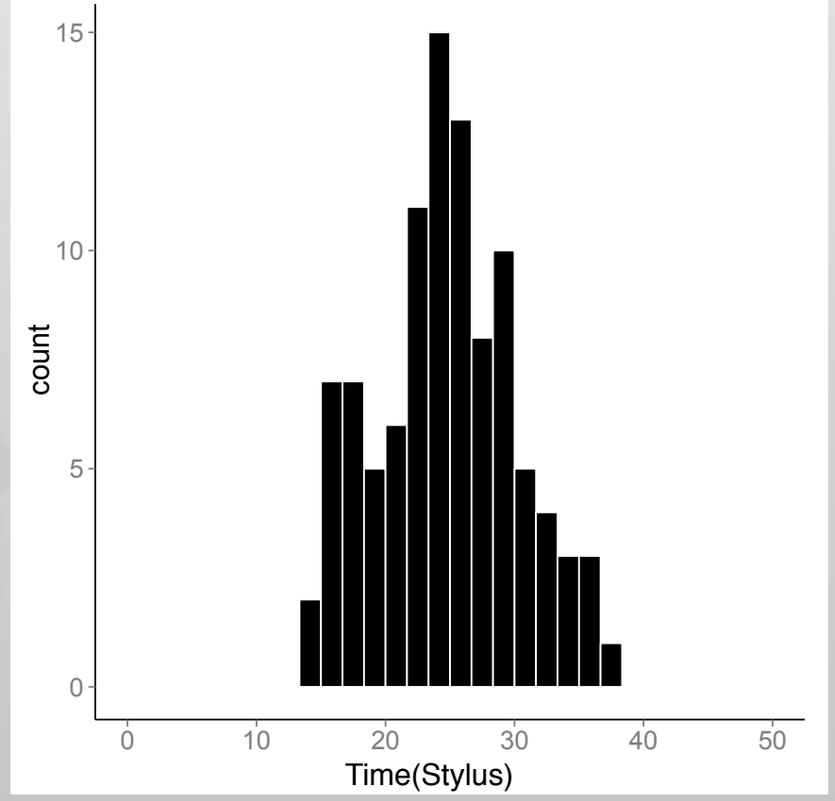
Descriptive Statistics

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

Example



mean = 25s



mean = 25s

Note: different data from previous slides

Descriptive Statistics

- Measures of central tendency
  - Mean: "**average**"
  - Median: the **middle point** of the sorted data
- Measures of spread (or variability)
  - SD: Standard deviation (square-root of variance)

Sample vs. Population

- Population: all your target users (e.g., all human-beings)
- Sample: the participants in your study
- Descriptive statistics (mean, SD, etc.) refer **only** to the sample and **not** the population
- How do we make statements about population?

Null Hypothesis Significance Testing

Null Hypothesis Significance Testing (NHST)

- When we have a difference in means between sample distributions, it can be due to following reasons:
  - There is indeed a true difference in the populations
  - There is no difference in populations, this difference is due to random chance
- NHST is used to tell these two apart
- We cannot be 100% sure that there is a difference, but we can estimate the chances (i.e., probability) that we are wrong.
  - If this probability is sufficiently small, then we say we have a statistically significant finding.

Null Hypothesis Significance Testing (NHST)

- We first assume that there is no effect of independent variable on the dependent variable (this is called **null hypothesis**)
- Under this assumption, we conduct the experiment and collect data
- Then, *p*-value gives the probability that our experiment produced this data.
  - E.g., $p = 0.05$: "Assuming input type (fingers, stylus) does *not* influence task completion time, then there is a 5% probability that we got this data from our experiment."
- *De facto* cutoff level of $p = 0.05$ for statistical significance
  - If $p < 0.05$, the null hypothesis is probably wrong (because our experiment produced this data)

In-class Exercise: *p* value

- Suppose your want 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?

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

B. There are 3% probability that school students watch TV in different amount that college students

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

D. 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

Easily Confusable!

- Which of the following statements are correct?

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

Incorrect: not the definition of p-value, specifying direction of the comparison

B. There are 3% probability that school students watch TV in different amount that college students

Incorrect: not the definition of p-value, specifying direction of the comparison

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

Incorrect: assuming the difference in population

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

Correct: assuming no difference in the population and does not specify the direction

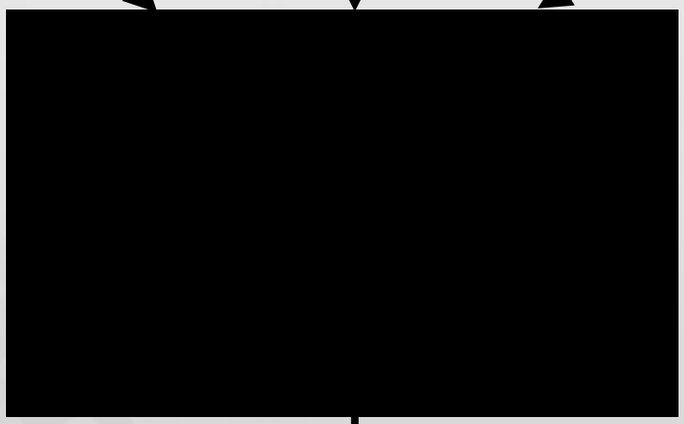
Test Statistics

- We try to fit a statistical model (that represents alternate hypothesis) to our data
- Test statistics (t, F, χ^2, etc.) tell us whether the model is a good fit for our data or not
  - Good fit: p-value is low (and we accept alternate hypothesis)
  - Bad fit: p-value is high (and we reject alternate hypothesis)
- Theoretical probability distribution of test statistics depends on degrees of freedom (DOF)
  - Therefore, report DOF with your test statistic.

DOF (calculated from sample size)

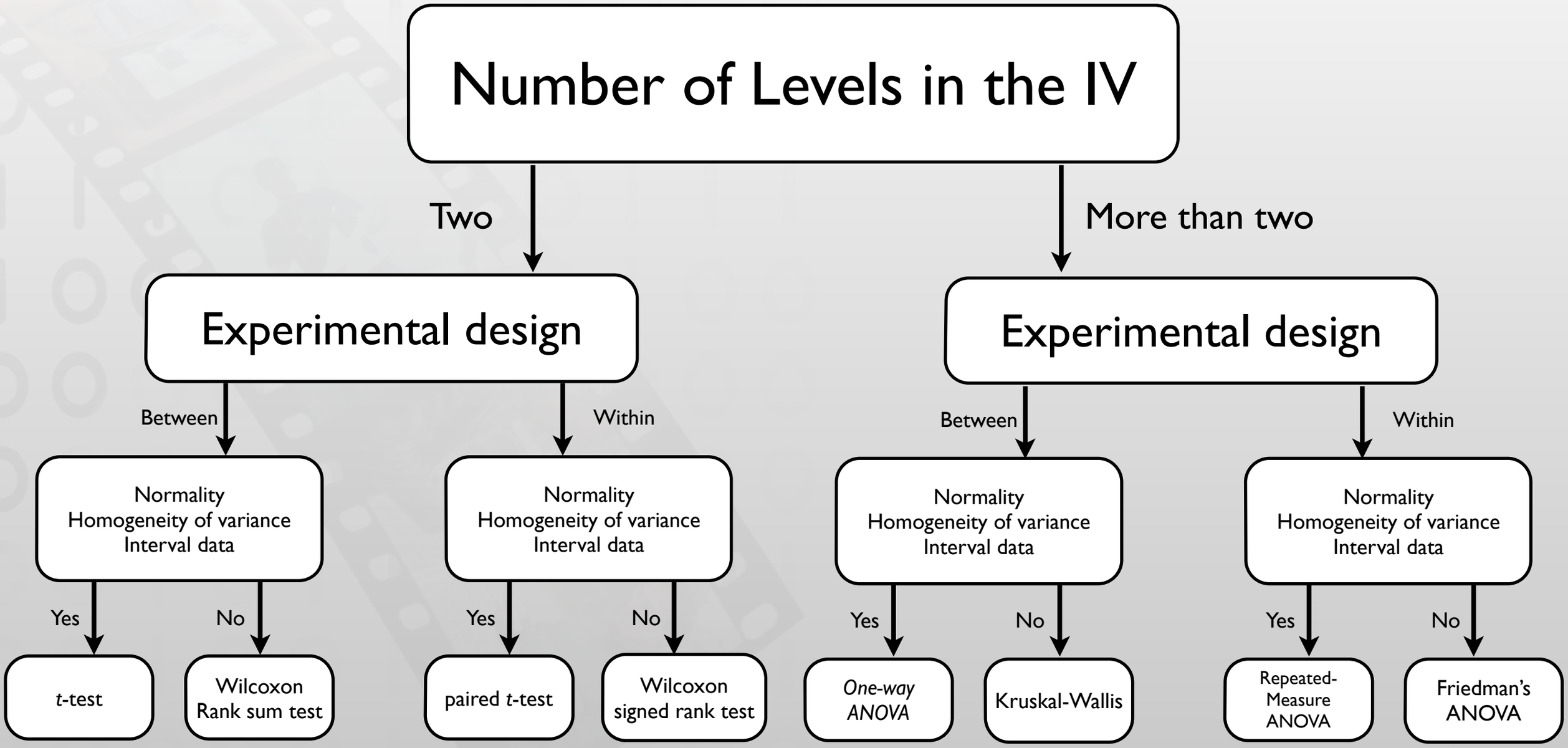
Choice of test statistic (e.g., t or F)

Data from your experiment



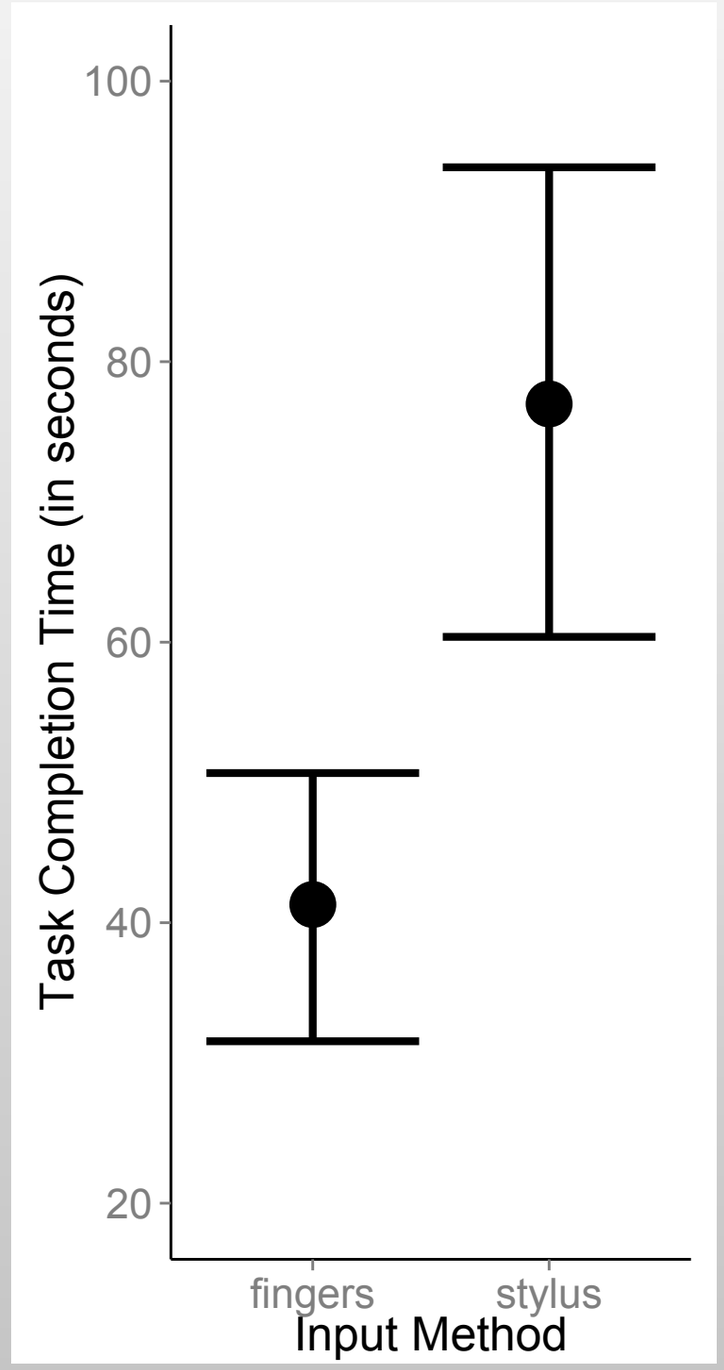
p-value

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 seconds 95% CI [31.78, 52.22]) is faster than Stylus (M =76.21 seconds [59.40, 93.02]). Difference between means (effect size) = 34.18 seconds.



Effect Size

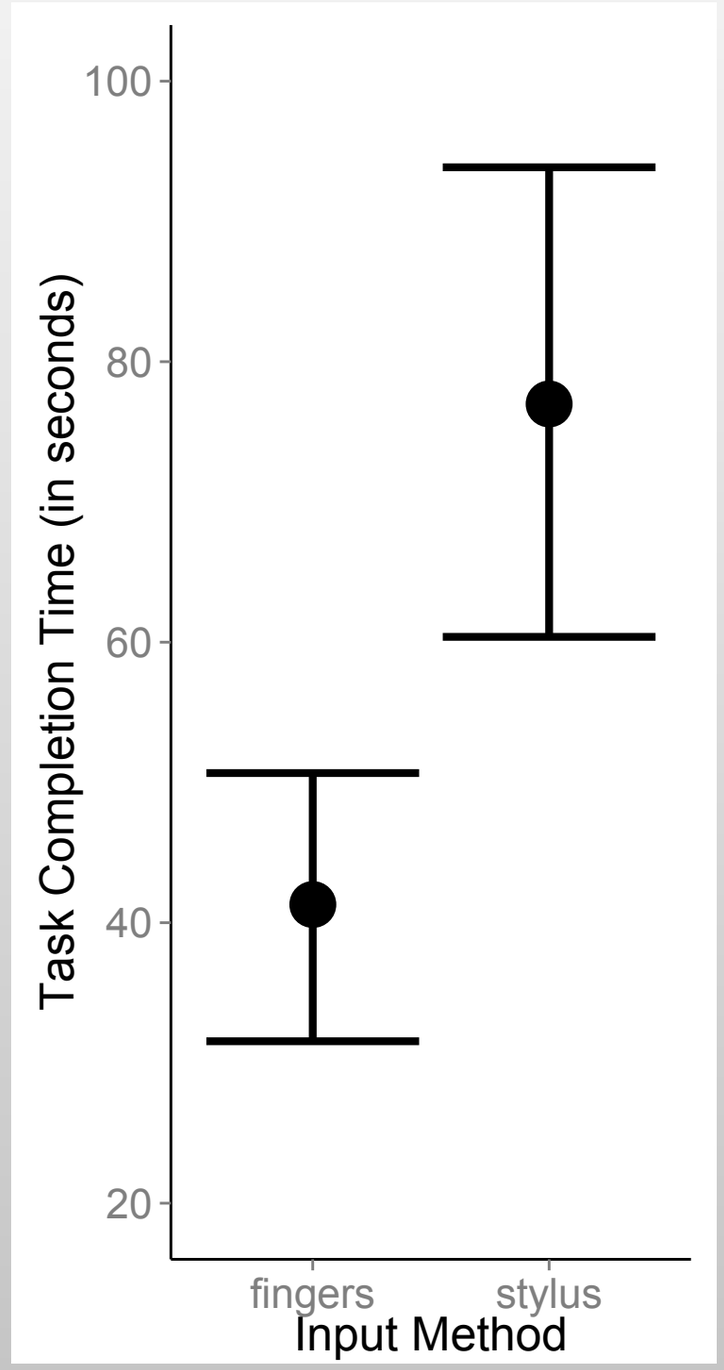
- *p*-value only gives us the chances our result are significant
- But even if the result is statistically significant ($p < 0.05$), it may not be practically significant
- We use effect size to indicate the magnitude of our result
  - In experimental studies, it indicates how strong the impact of manipulation of independent variables is on the dependent variables.

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

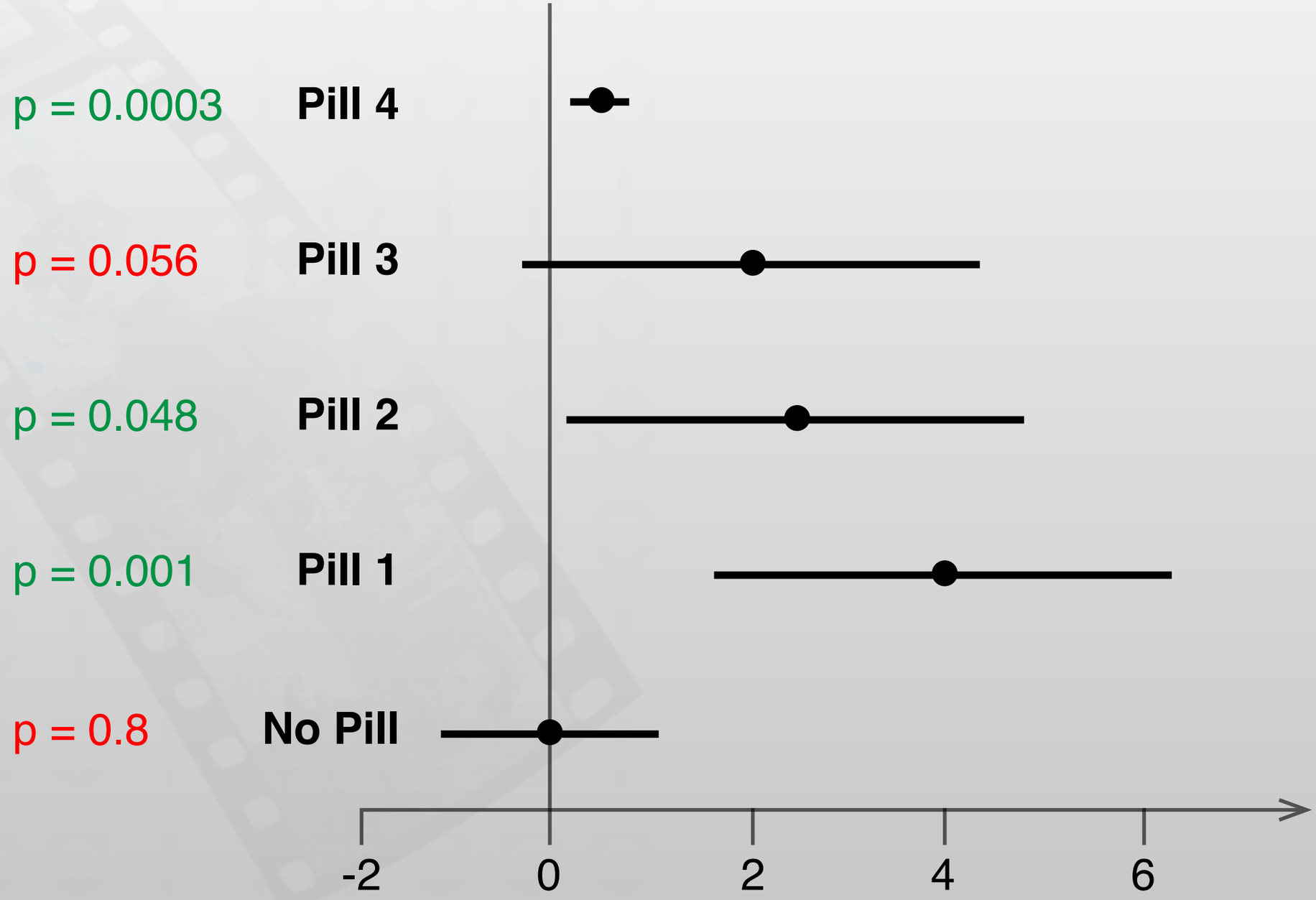
Result of Statistical Analysis

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



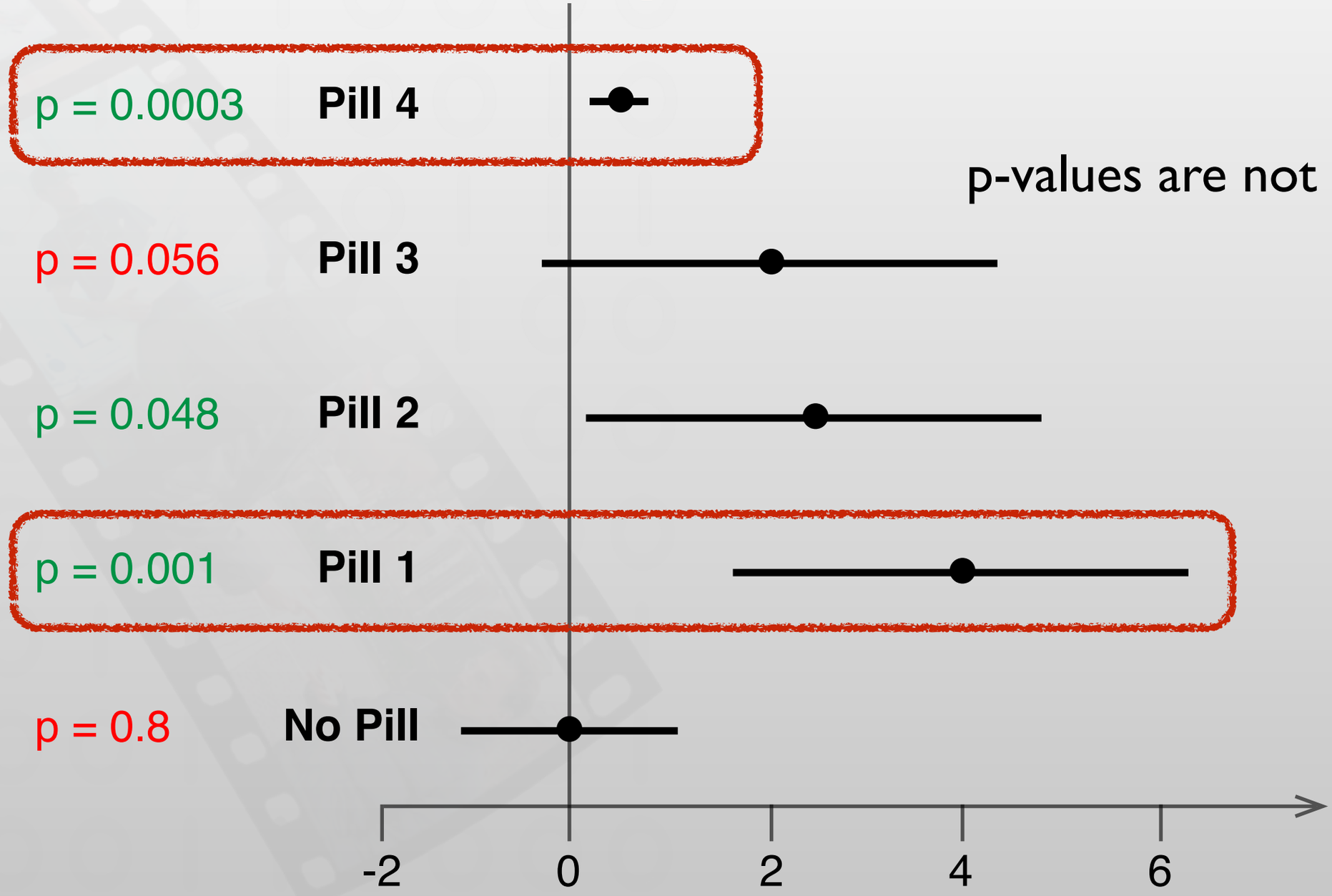
Interpreting Uncertainty in Data

How Uncertainty Influences our Interpretation



Adopted from Ziliak and Mccloskey, 2009

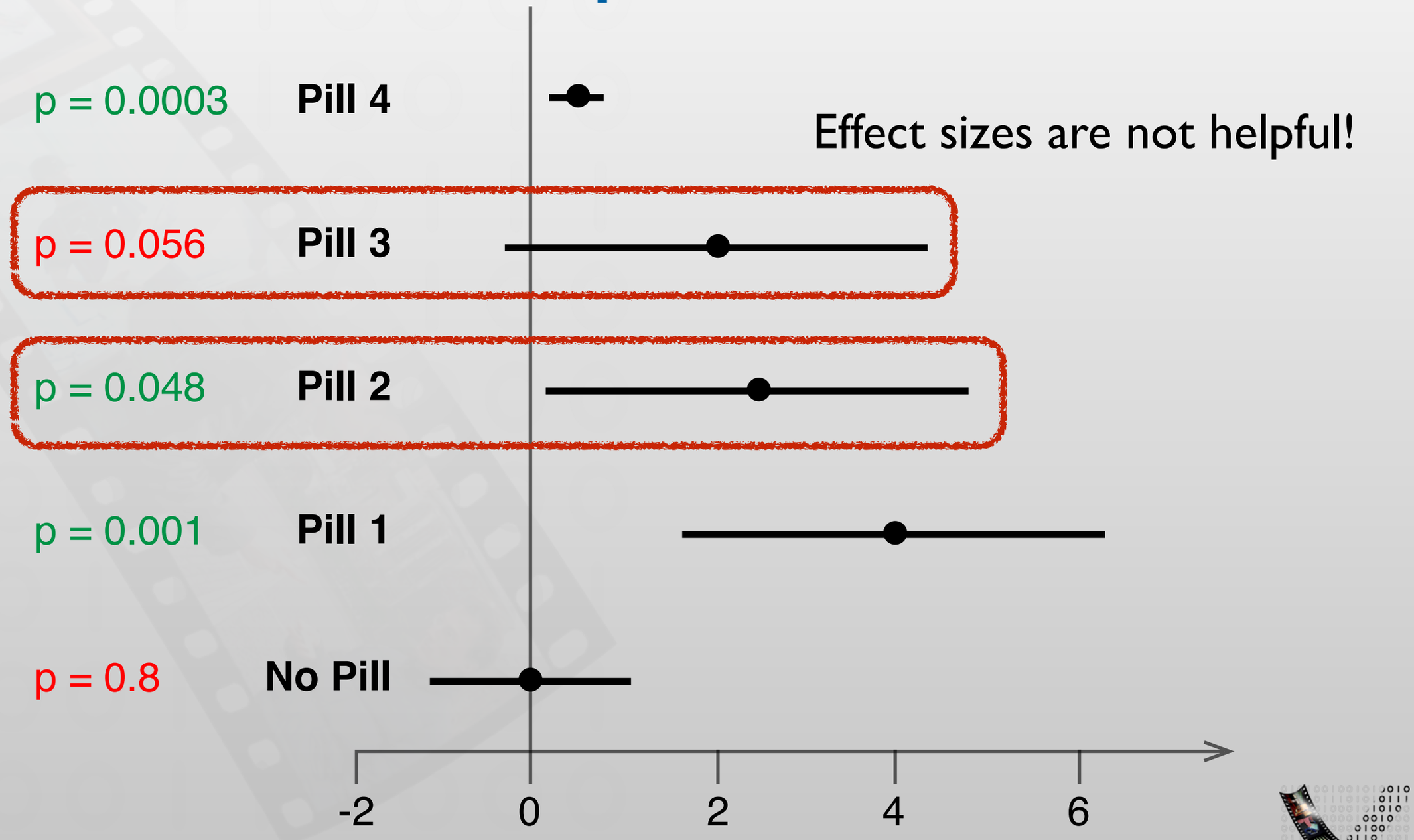
How Uncertainty Influences our Interpretation



p-values are not helpful!

Adopted from Ziliak and Mccloskey, 2009

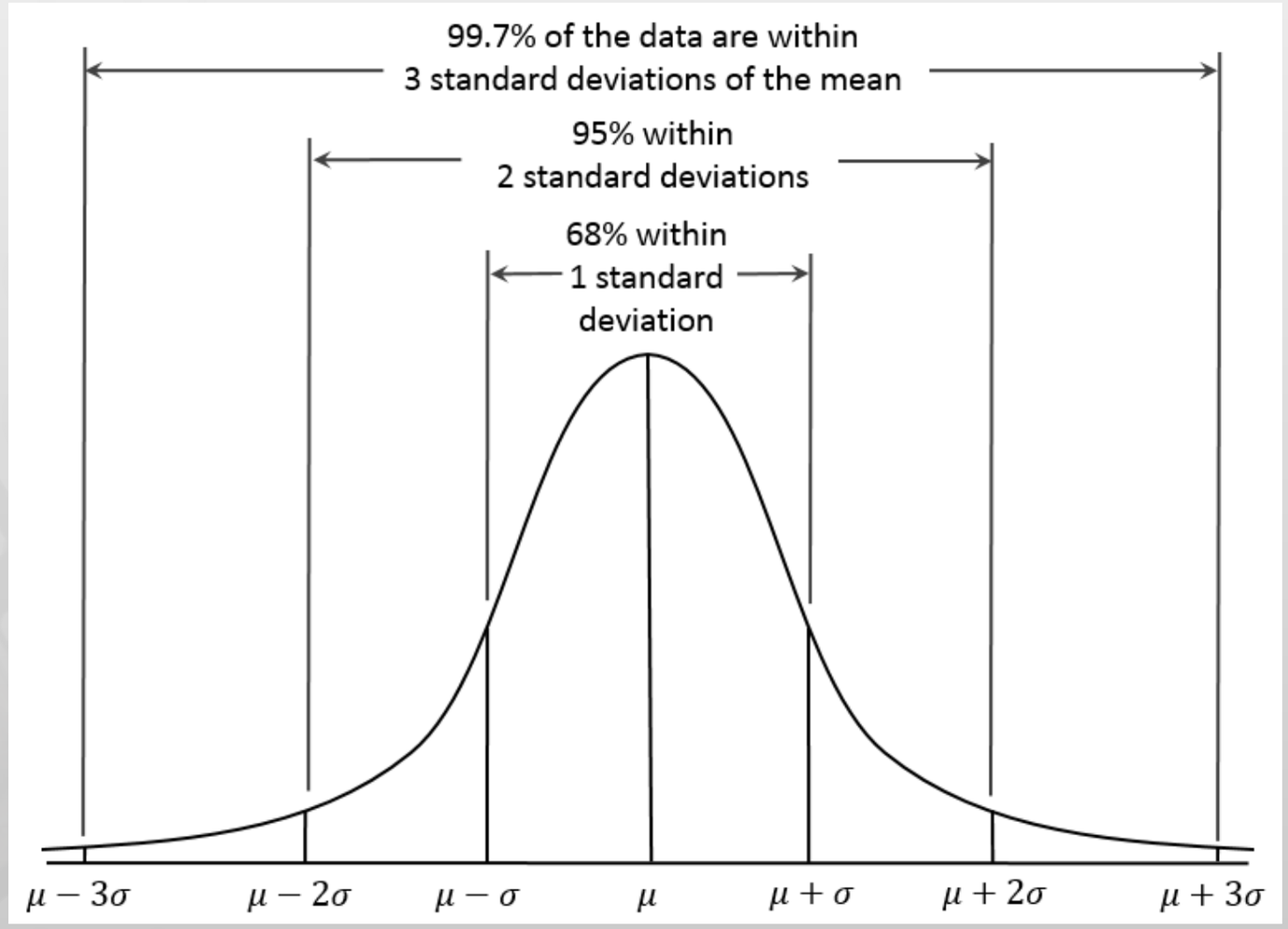
How Uncertainty Influences our Interpretation



Adopted from Ziliak and Mccloskey, 2009

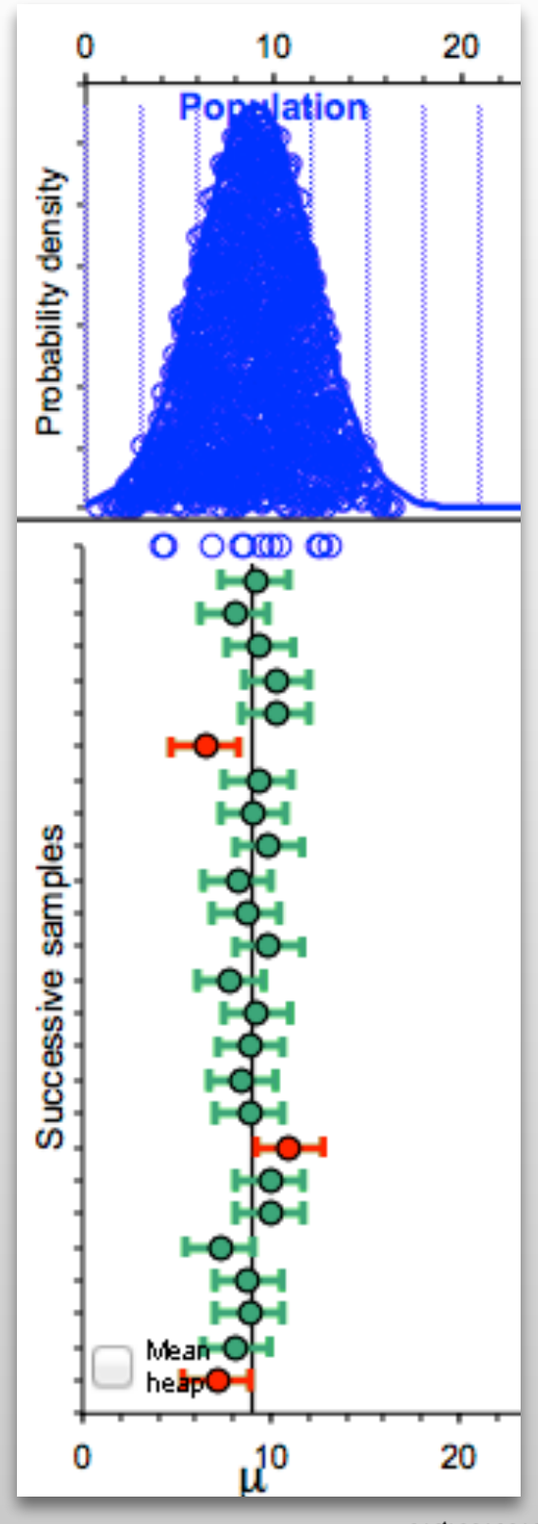
Normal Distributions

- Characteristic "bell-shape" of the distribution
- Central Limit Theorem
  - "Distribution of a large number of independent, identically distributed variables will be approximately normal…"



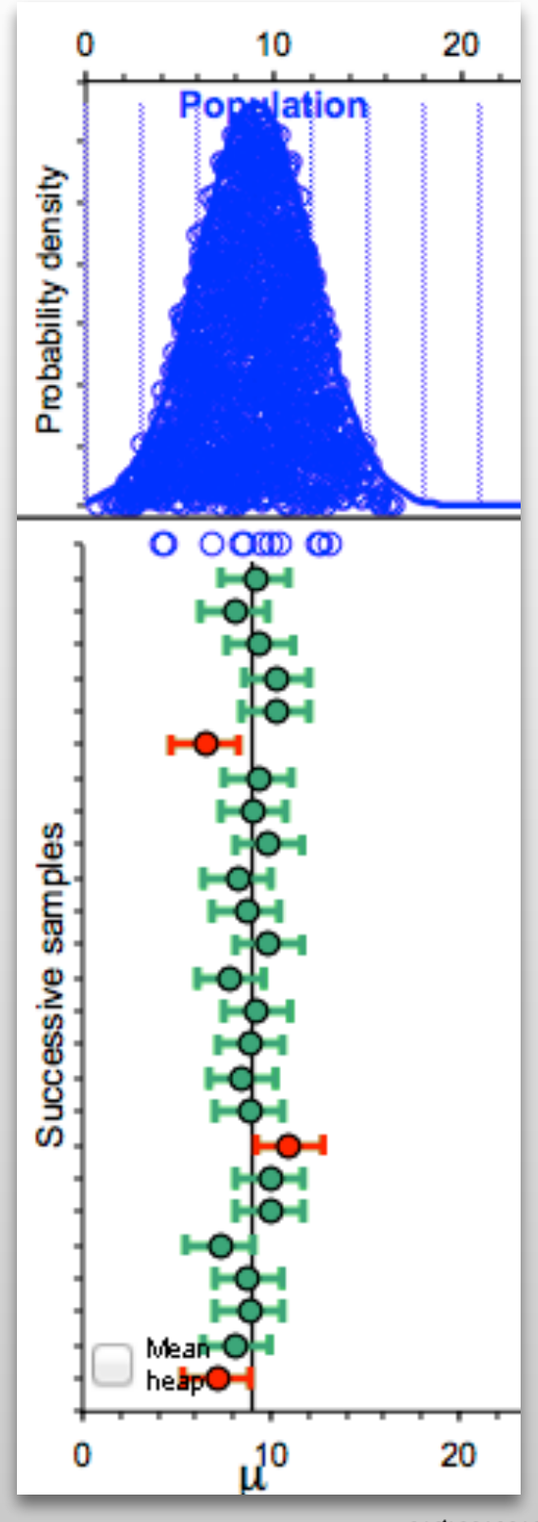
95% Confidence Interval

- An interval estimate (i.e., a range) of the population mean
- In an infinite number of experiments, 95% of the CIs will include the population mean



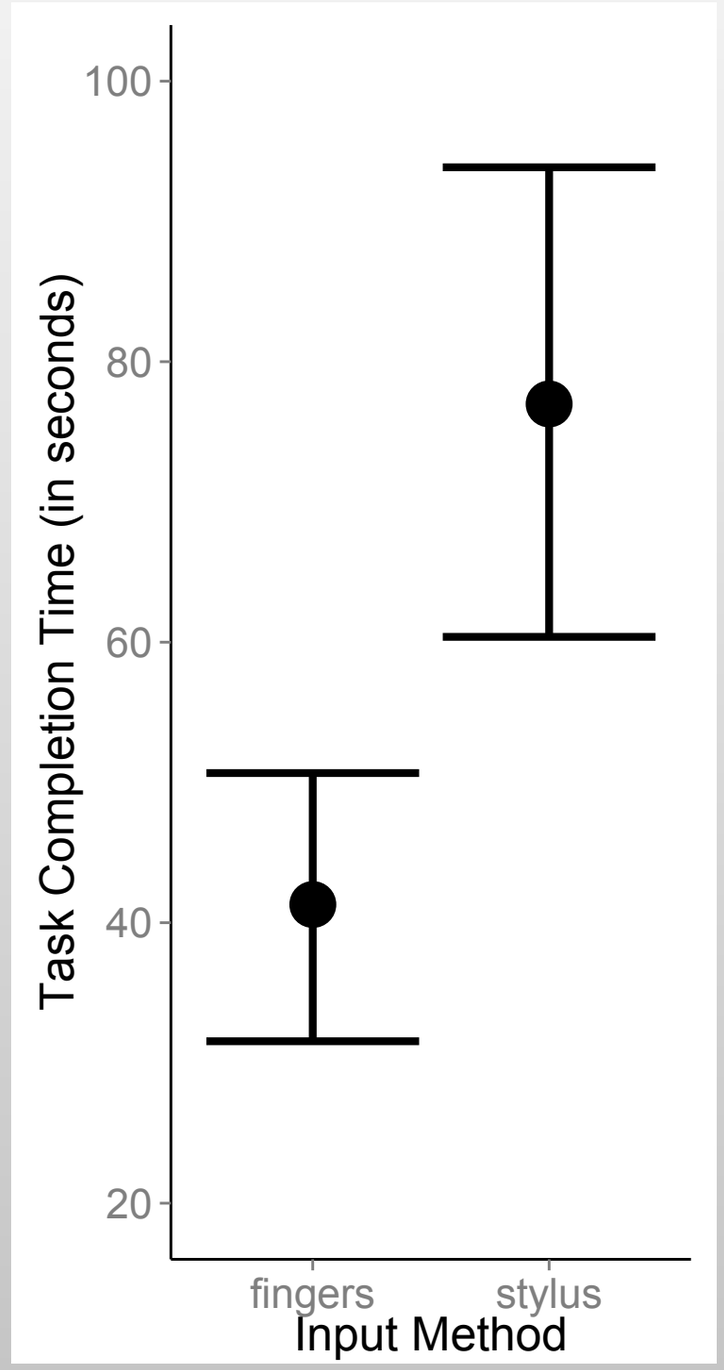
95% Confidence Interval

- Given a sample, the 95% CI tells you the following:
  - If you sample from the same population a large number of times, 95% of the time the population mean will be contained in the 95% CI
- 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 seconds 95% CI [31.78, 52.22]) is faster than Stylus (M =76.21 seconds [59.40, 93.02]). Difference between means (effect size) = 34.18 seconds.



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
  - You don't have to read the whole paper. A PDF document, which contains a set of exercises related to interpreting graphs, will be uploaded to L2P. Your task is try and solve these exercises using online resources.

Recommended Reading

- Statistical Methods for HCI Research by Koji Yatani, U. of Tokyo

  Uses R (free software)

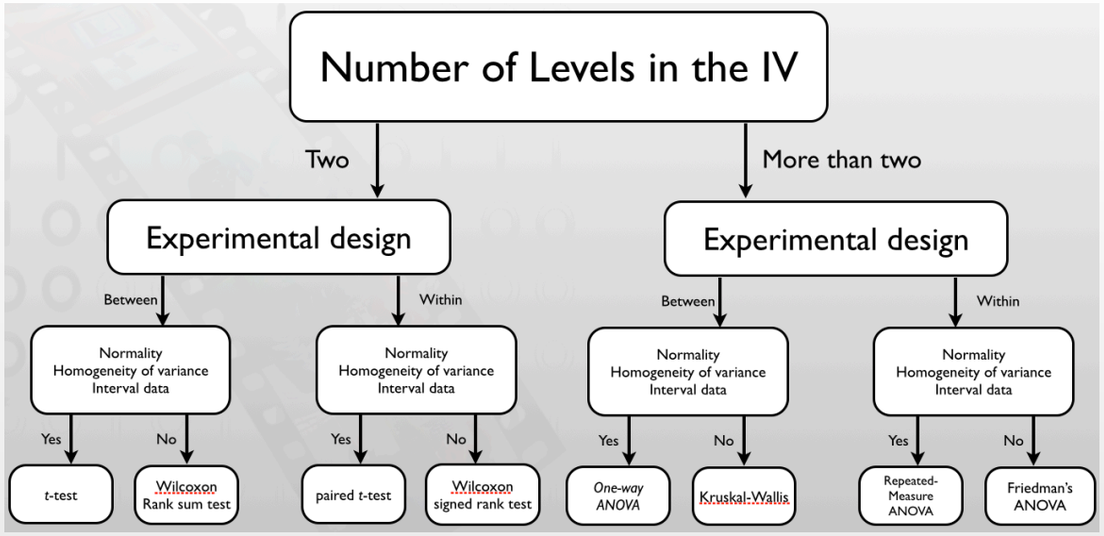
  http://yatani.jp/teaching/doku.php?id=hcistats:start

- Practical Statistics for HCI by Jacob O. Wobbrock, U. of Washington

  Uses SPSS and JMP (trial version available for free download)

  http://depts.washington.edu/aimgroup/proj/ps4hci/

Summary



- Effect size (mean) and their confidence interval describes the data
- Effect sizes quantify the magnitude the effect of IV on DV
- *p*-value is the probability that the result occurs assuming no effect of IV.
- You choose a test based on number of levels in the IV, experimental design, and statistical assumptions