DIS1: Statistics Lecture

- Please download the data set from (updated: last night)
  hci.rwth-aachen.de/stats4dis

Are you…Left-Handed?

- We need you for a user study on interactive tabletop!
- The study takes half an hour to complete.
- Leave me your name or email to

  Norbert Dumont norbert.dumont@gmail.com

Review

- What are four phases of technology lifecycle proposed by David Liddle and Jan Borchers
  - Where is the sweet spot? What is its implication?
- What is "multimodal interface"? Give an example
- What is the difference between virtual reality and augmented reality?
- Three classes of devices in an ubiquitous computing environment?

Theory

- ✓ Models of interaction
  - ✓ Affordances, mappings, constraints, types of knowledge, errors
- ✓ Design principles
- ✓ Human cognition and performance
- ✓ Interaction design notation
- ✓ History and vision of HCI

Practice

- ✓ Sketching
- ✓ User observation
- ✓ Iterative design
- ✓ Prototyping
- ✓ Ideation
- ⇒ User studies and evaluation

A Rough Guide to Research

- A hunch or a *research question*: ideas or problem that you are interested in
- Literature review: How does existing research address these questions?
- Qualitative findings: observing users, testing prototypes, surveys
  - *Descriptive results:* explain what happened, and what users said
  - *Correlational results*: numerical, indicate if there is a correlation
- Experiments: controlled environment, verify *causal relationship*
- Analysis, discussion, and conclusion
- Publication: Share your knowledge; contribute to the science

Review: Controlled Experiments

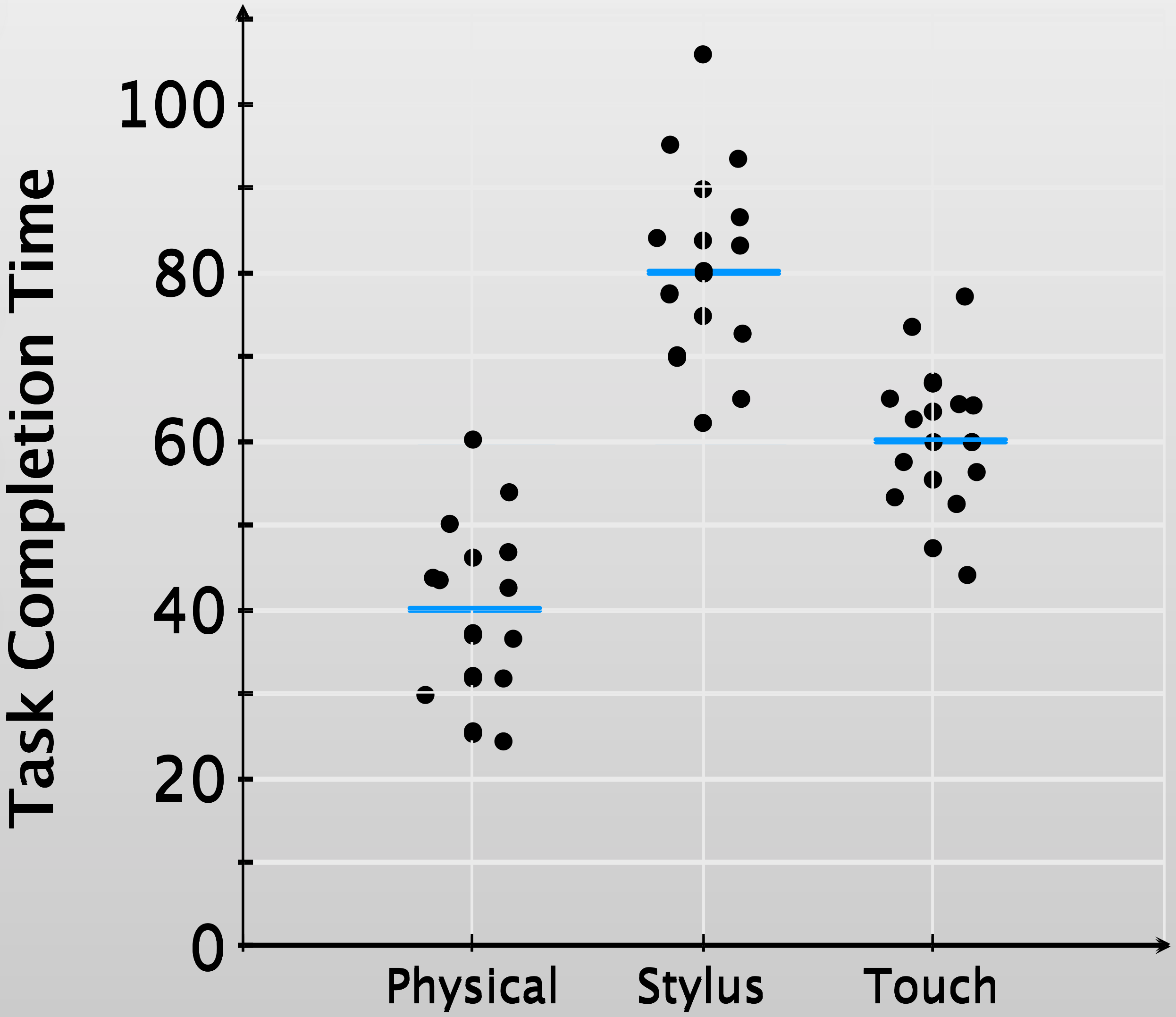
- Research question: On a mobile phone, is typing faster using *physical keys* compared to using a touchscreen and your *fingers* or a *stylus*?
  - Research hypothesis?
  - Variables?
  - Experimental design?
  - Expected data?

Mobile Phone Text Input Example

- Research question: On a mobile phone, is typing faster using *physical keys* compared to using a touchscreen and your *fingers* or a *stylus*?
- IV: keyboard types: {physical, stylus, touch}
- DV: time in seconds for typing a specified sentence.
  - Begin: when the user presses the first key
  - End: when the user presses Enter
- Design: between-groups
  - Each keyboard is tested by 20 participants
  - Each participant types the sentence only one time (one trial)

Variance of Real Data

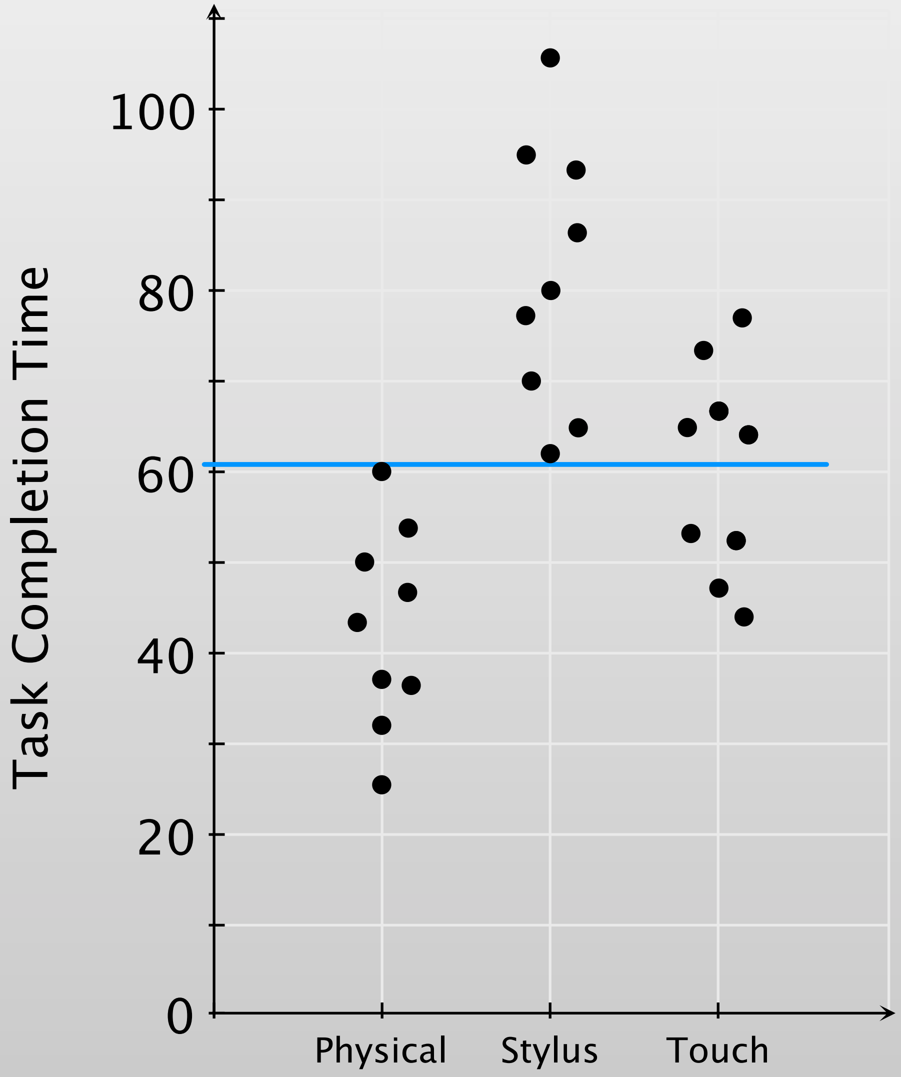
- Data from experiments is noisy
- Effect: Variance caused by the different levels of our IV
- Confound: Variance caused by uncontrolled factors ("confounding variables")



NHST: Null Hypothesis Significance Testing

- Assuming that there is *no* effect of IV (i.e., null hypothesis is true)
  - E.g., keyboard type does *not* affect completion time
- Then what is the probability that our measurements would occur? ⇒ *p* value
  - E.g., *p* = 0.023:

    "If keyboard type does *not* affect compleation time, then there would be a 2.3% probability that our measurement turns out as it did."
- 0.05 is generally considered the *de facto cutoff level* of *p* for statistical significance

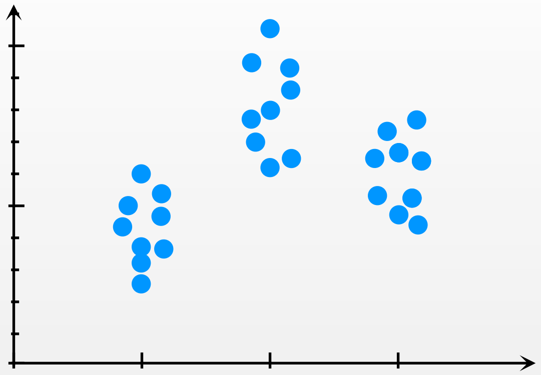


ANOVA:
Analysis of Variance

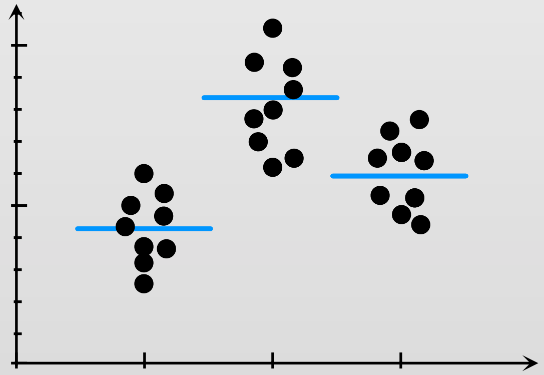
- Goal: partition the variance from different sources

- Method: fit different models and determine how good the models explain the data
    - One extreme: explain each data point with one parameter
    - Another extreme: all data can be represented by a single mean ⇒ no effect
    - Determine just adequate model that fits the data

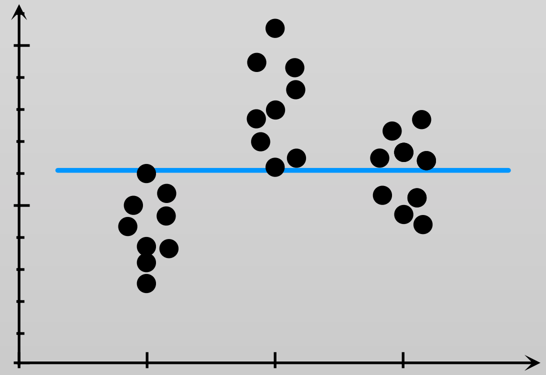
- One-Way ANOVA: one IV, between-groups



Maximal model (each data point is one parameter)

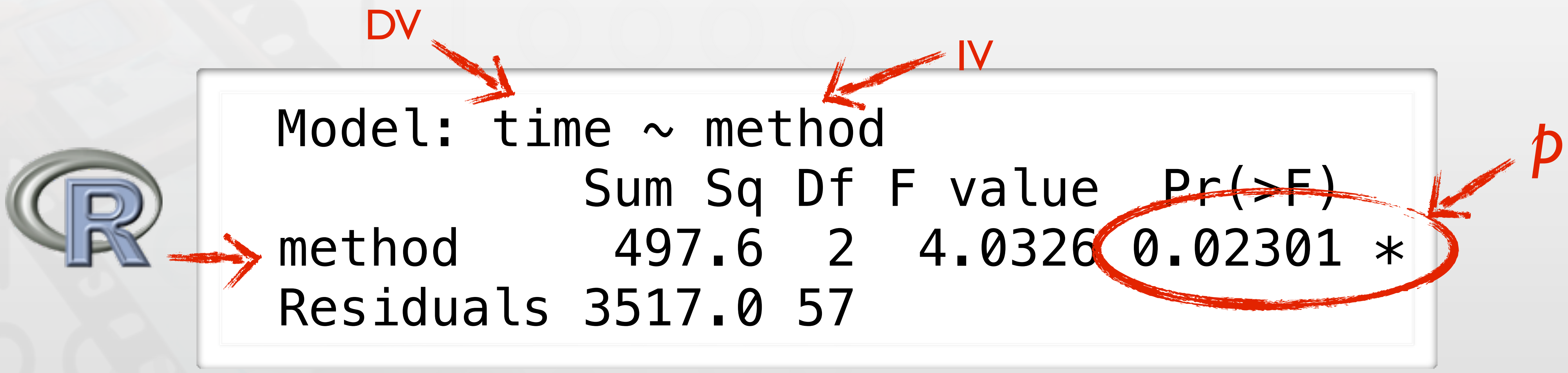


A candidate model



Null model (one mean)

One-Way ANOVA Output



```
Model: time ~ method
             Sum Sq Df F value   Pr(>F)
method        497.6  2  4.0326 0.02301 *
Residuals    3517.0 57
```

- Each line shows variance for one IV
    - Significant *p*-values are indicated by one or more stars (*)
- Report: "The choice of method had a significant effect on completion time, $F(2,57) = 4.03$, $p = 0.02301$."
    - Implies that there is a very low chance (2.3%) that the data would be like this if the method did *not* affect completion time.
- But: we do *not* know *which* method differs yet!

Post-hoc Test: Tukey's Test

- Compares means of data from each level against each other level simultaneously using *t*-tests
- Determines whether the differences between means are more than what the standard error allows
- Output: one *p*-value for each pair
- Below: significant differences between physical and other types, but not between stylus and touch

"comparison of means between stylus and physical"

```
                              Pr(>|t|)
stylus - physical == 0  ...    0.0427 *
touch - physical == 0   ...    0.0087 **
touch - stylus == 0     ...    0.5221
```

Demo: One-Way ANOVA

mobileTextInput.csv

Please follow along on your laptop!

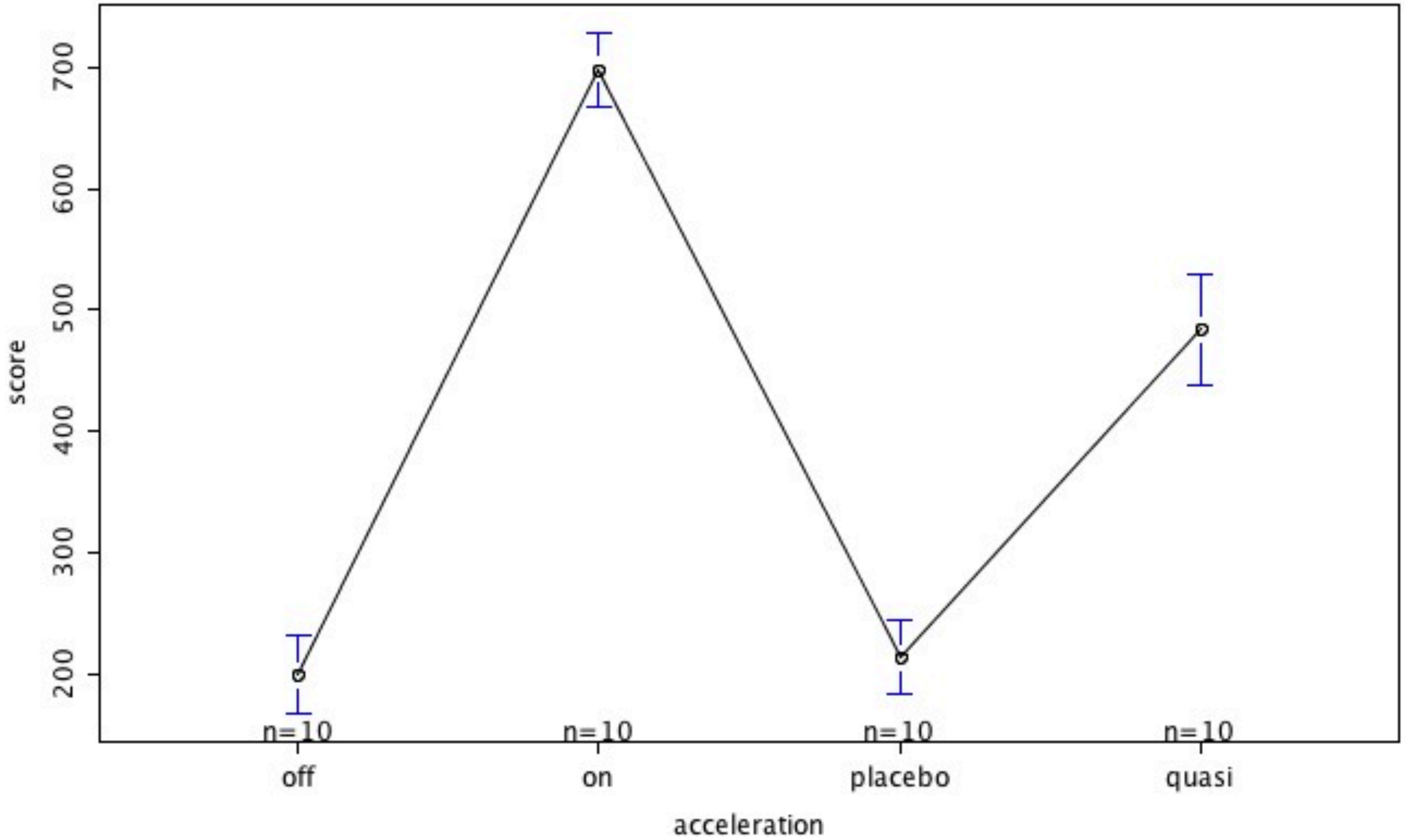
In-class Exercise: One-Way ANOVA

shootingGame.csv

- When people play a first-person shooter, does their mouse acceleration influence the score they get?
  - What are IV and DV?
  - If we use between-group design, how should the data table look like?
  - Visualize data in a plot

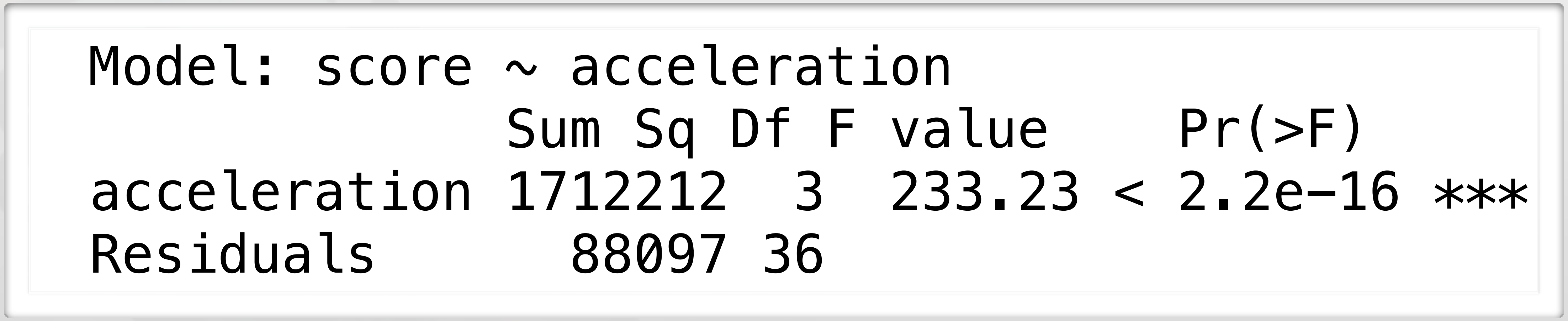
    What should be on x-axis, y-axis?

In-class Exercise: One-Way ANOVA



- Estimate the result from the graph
- Run One-Way ANOVA

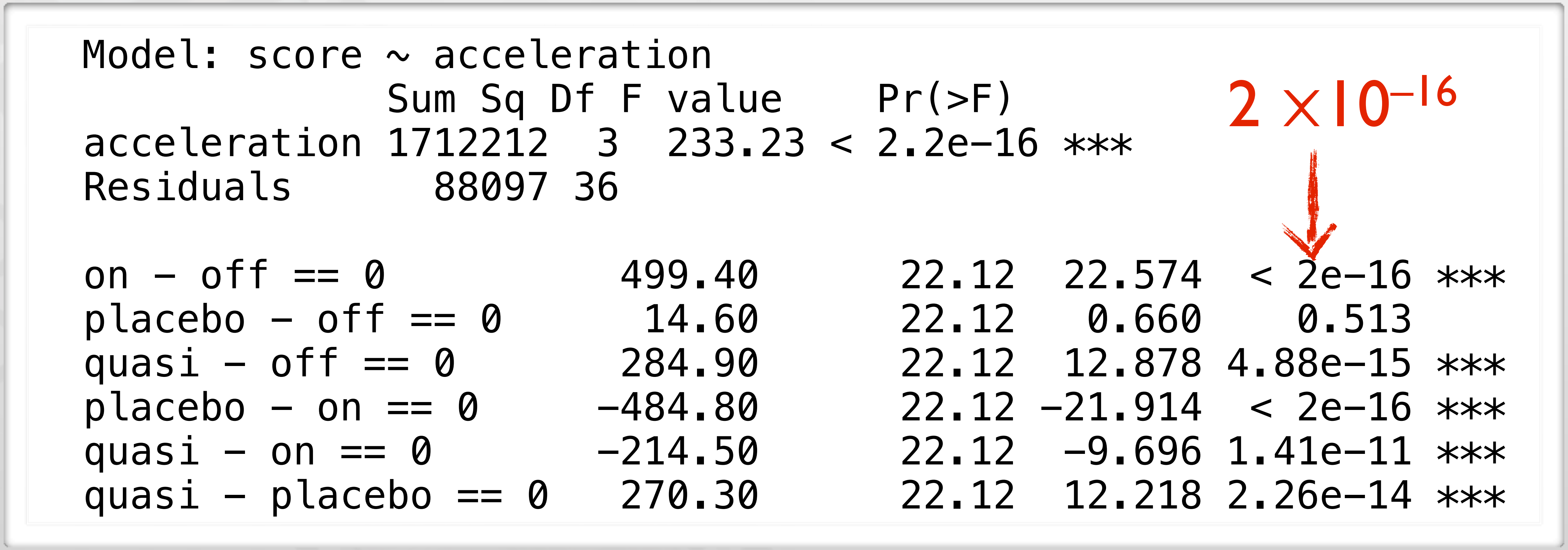
In-class Exercise: One-Way ANOVA



```
Model: score ~ acceleration
             Sum Sq Df F value    Pr(>F)    
acceleration 1712212  3  233.23 < 2.2e-16 ***
Residuals      88097 36
```

- Is the result significant?
- Run Tukey's test. Which pairs of means are significantly different?

In-class Exercise: One-Way ANOVA

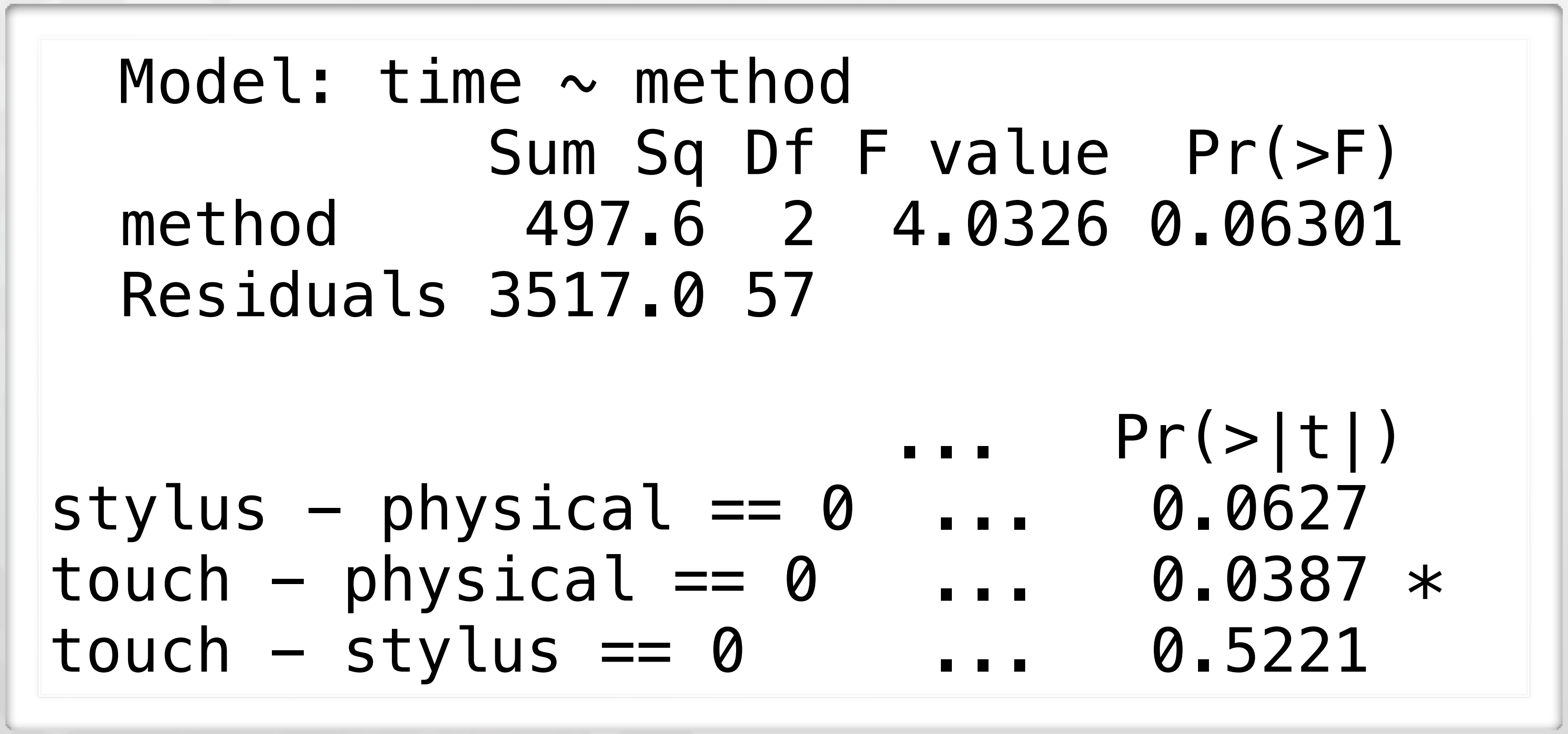


```
Model: score ~ acceleration
             Sum Sq Df F value    Pr(>F)
acceleration 1712212  3  233.23 < 2.2e-16 ***
Residuals      88097 36

on - off == 0          499.40      22.12  22.574  < 2e-16 ***
placebo - off == 0      14.60      22.12   0.660    0.513
quasi - off == 0       284.90      22.12  12.878 4.88e-15 ***
placebo - on == 0     -484.80      22.12 -21.914  < 2e-16 ***
quasi - on == 0       -214.50      22.12  -9.696 1.41e-11 ***
quasi - placebo == 0   270.30      22.12  12.218 2.26e-14 ***
```

2×10^{-16}

- What would you conclude from your results?

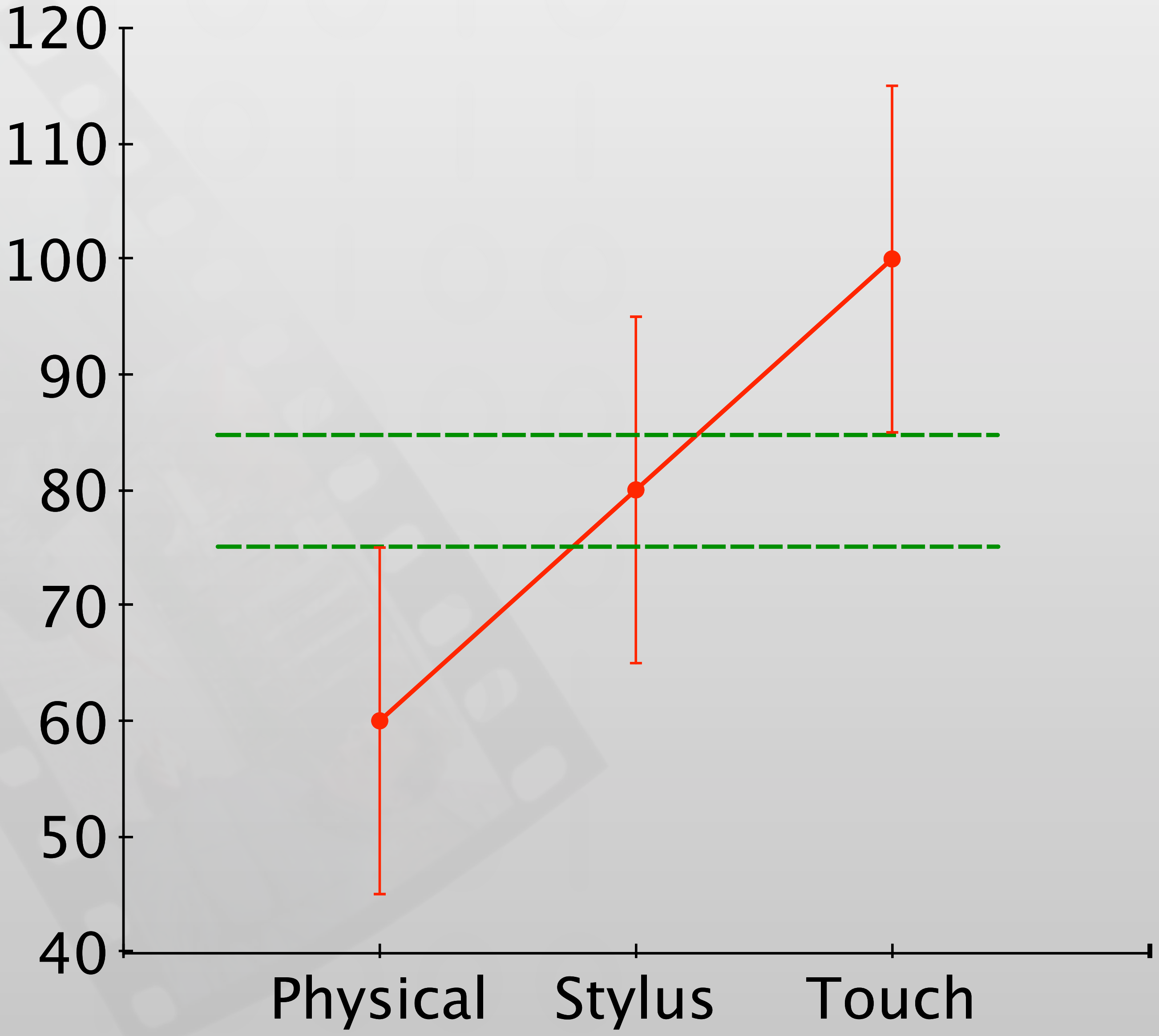
Help! Non-Significant *p*-value



```
  Model: time ~ method
            Sum Sq Df F value  Pr(>F)
  method     497.6  2  4.0326 0.06301
  Residuals 3517.0 57

                              ...   Pr(>|t|)
stylus - physical == 0     ...    0.0627
touch - physical == 0      ...    0.0387 *
touch - stylus == 0        ...    0.5221
```

- If ANOVA doesn't report significance, post-hoc test is *not* enough to support your hypothesis
  - Post-hoc test does not account for the variance caused between different conditions
- Increase sample size, or do Power Analysis (not covered here)

Non-Significant ANOVA but Significant Post-hoc



Data Types

- Interval variables: there is a fixed magnitude of difference between two values
  - Can meaningfully add two values
  - E.g., task completion time, distance from the center of target
- One assumption of ANOVA is that the data is interval variables
  - We often get non-interval variables, e.g., answers on Likert scales
- Ordinal variables: order is significant, but no meaningful arithmetic operations can be performed
  - E.g., "How easy do you think this statistics lecture is?"

    ❍ Very easy ❍ Easy ❍ Hard ❍ Very hard

Non-parametric Tests

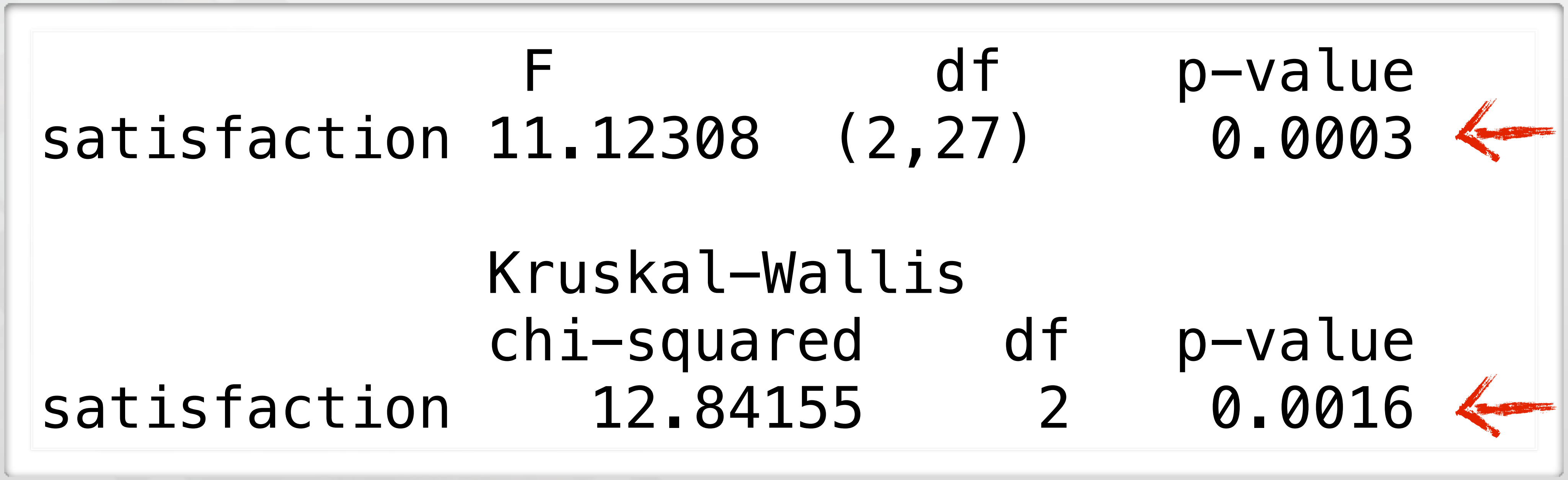
- Assumptions are less restricted than ANOVA (parametric)
- Less powerful: if the effect is small, you might not be able to detect significance
- Kruskal-Wallis test: non-parametric counterpart of ANOVA
  - Wilcoxon rank sum test: counterpart of *t*-test for comparing each pair

Demo: Non-parametric Test

sus1.csv

Please follow along on your laptop!

One-Way ANOVA vs. Kruskal-Wallis



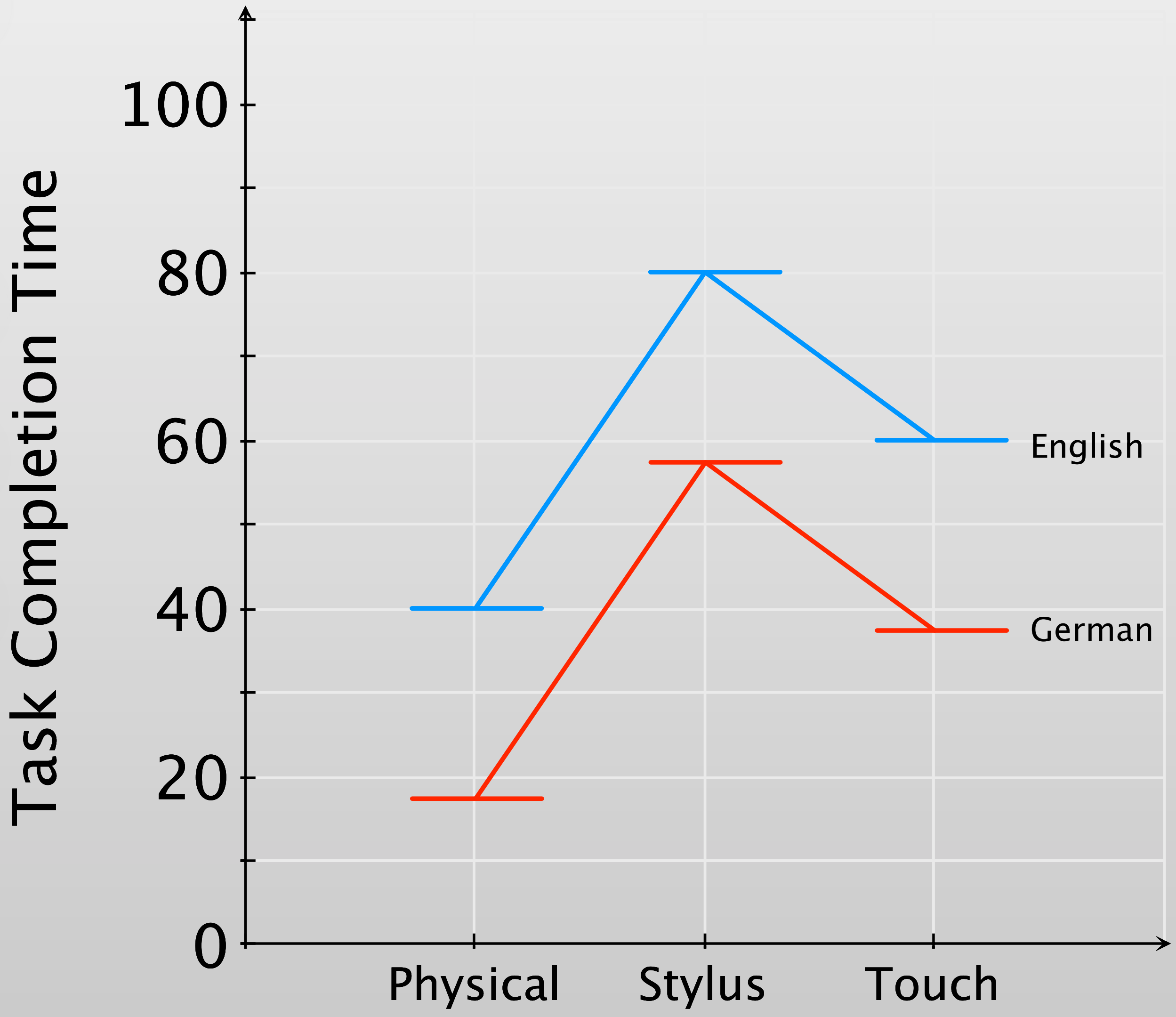
```
                  F           df      p-value
satisfaction 11.12308   (2,27)      0.0003

             Kruskal-Wallis
             chi-squared    df    p-value
satisfaction    12.84155     2     0.0016
```

- *p*-value of Kruskal-Wallis test is higher ⇒ easier to be non-significant

- Parametric method has more power to discover the significance

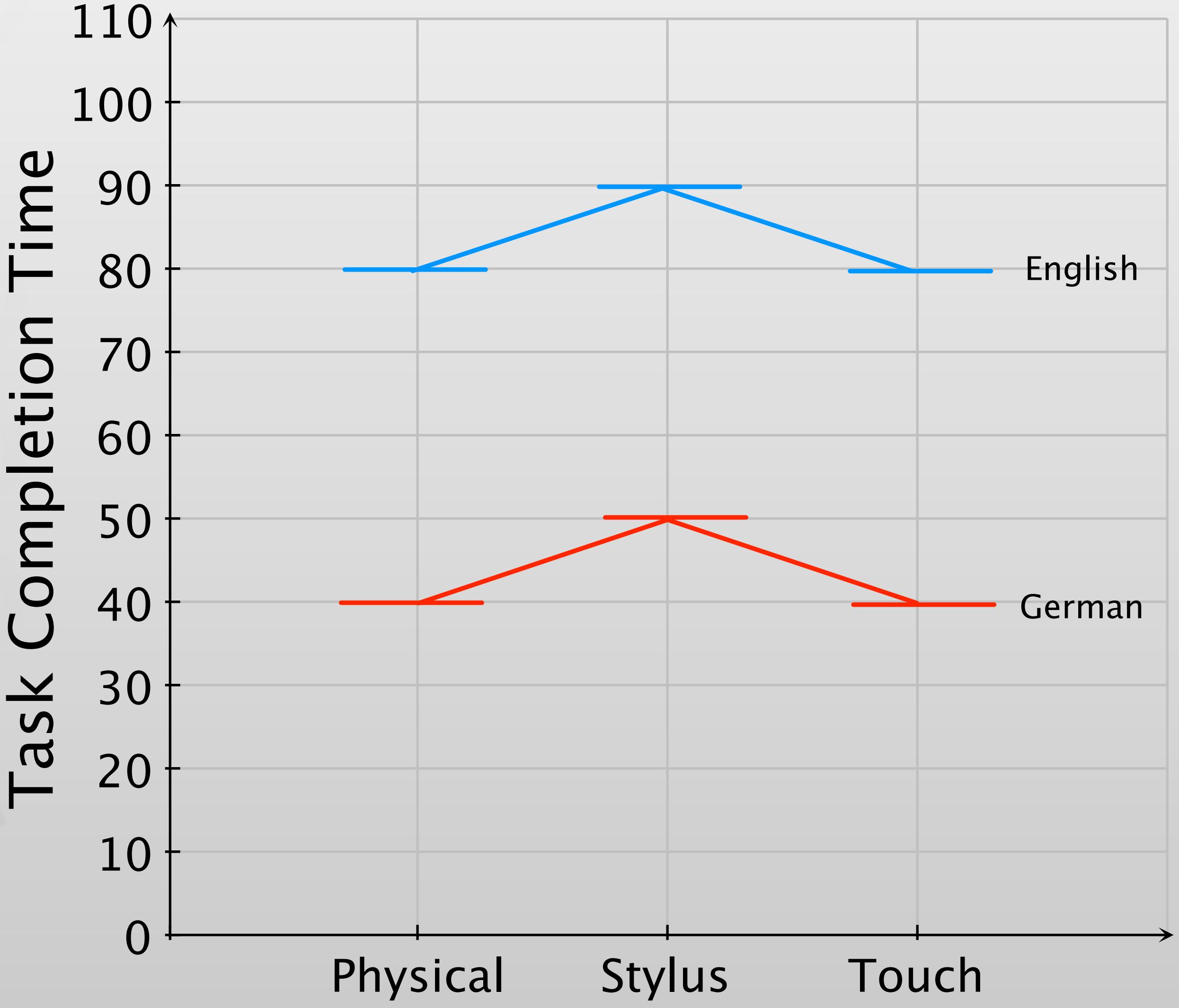
N-Way ANOVA

- For more than one IV, between groups
  - Often found in research
- Example: Does typing time for different input methods differ in different languages?

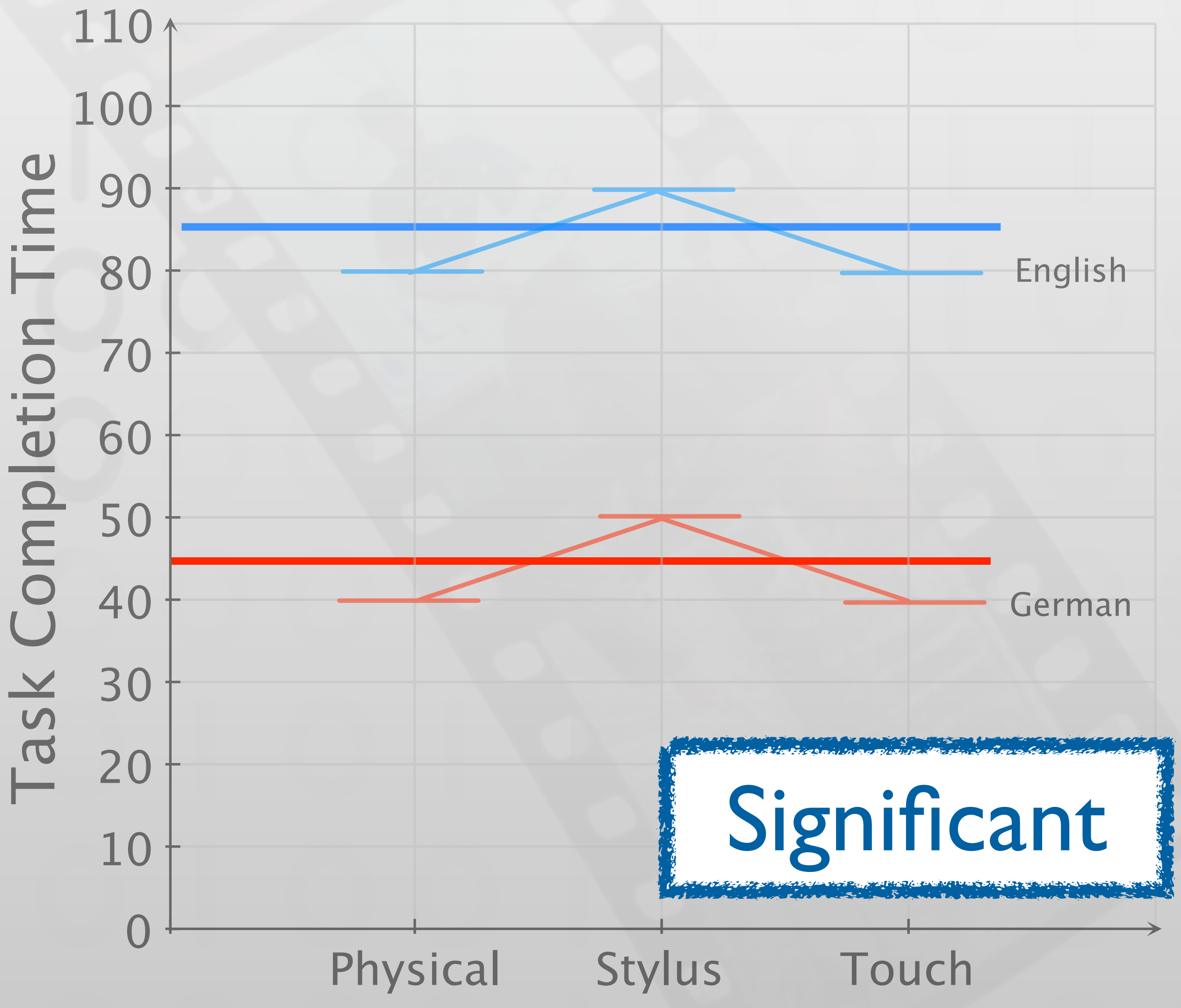


Main Effect

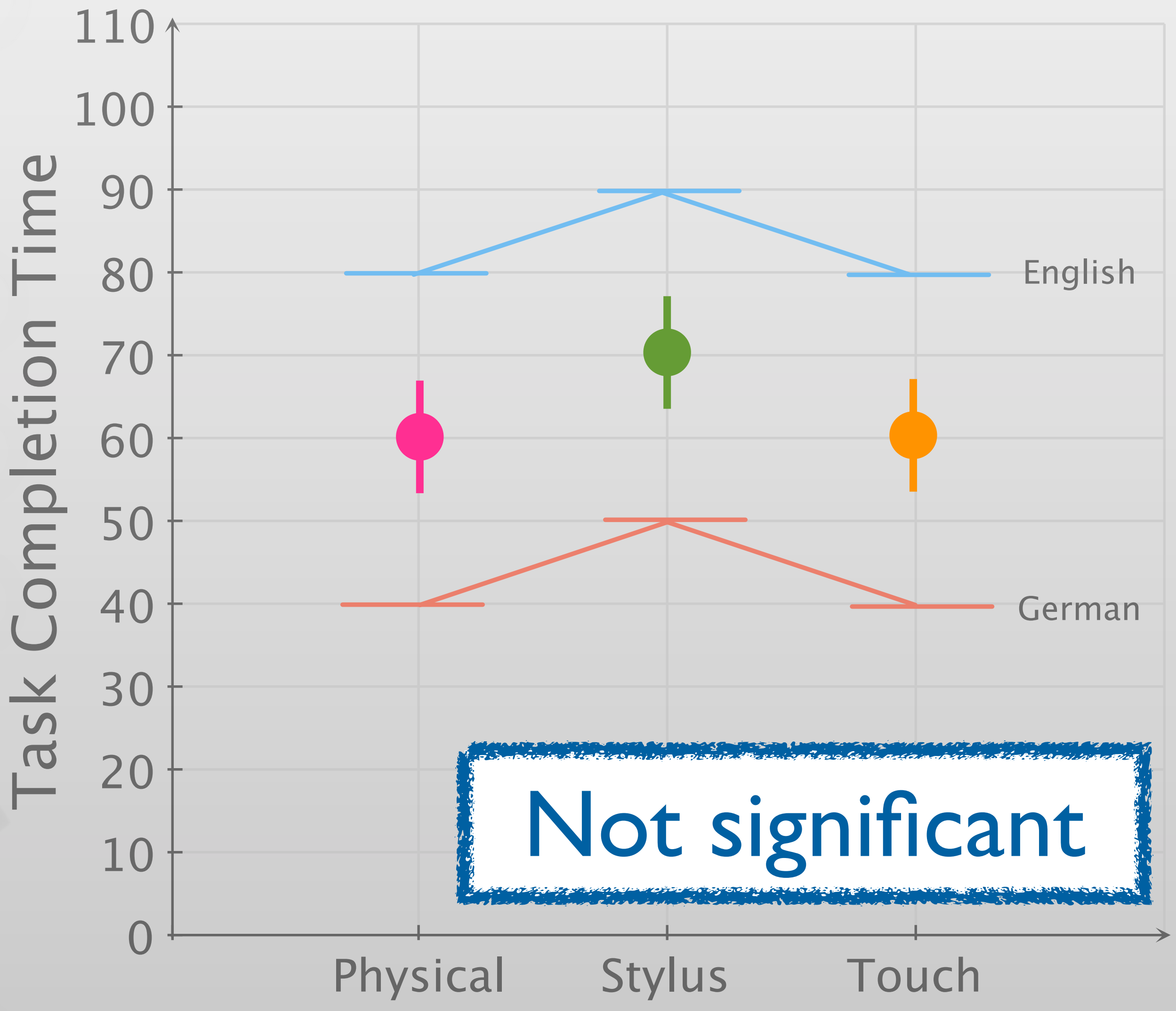
- Effect that each independent variable has by itself
- This graph: language has a main effect
  - Language changes task completion time, when averaged across all input methods
- Input method does *not* have a main effect
  - Input method does *not* change task completion time, when averaged across both languages



Estimating Main Effect with Marginal Means



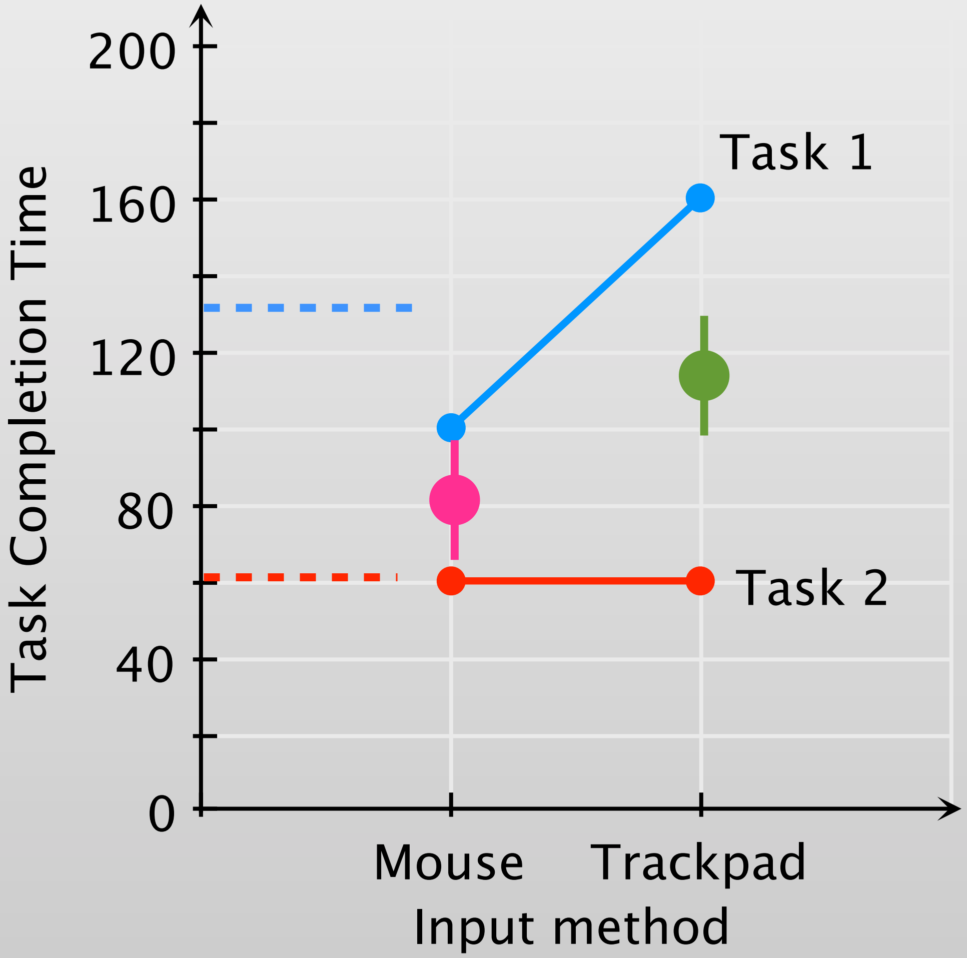
Marginal mean by language



Marginal mean by input method

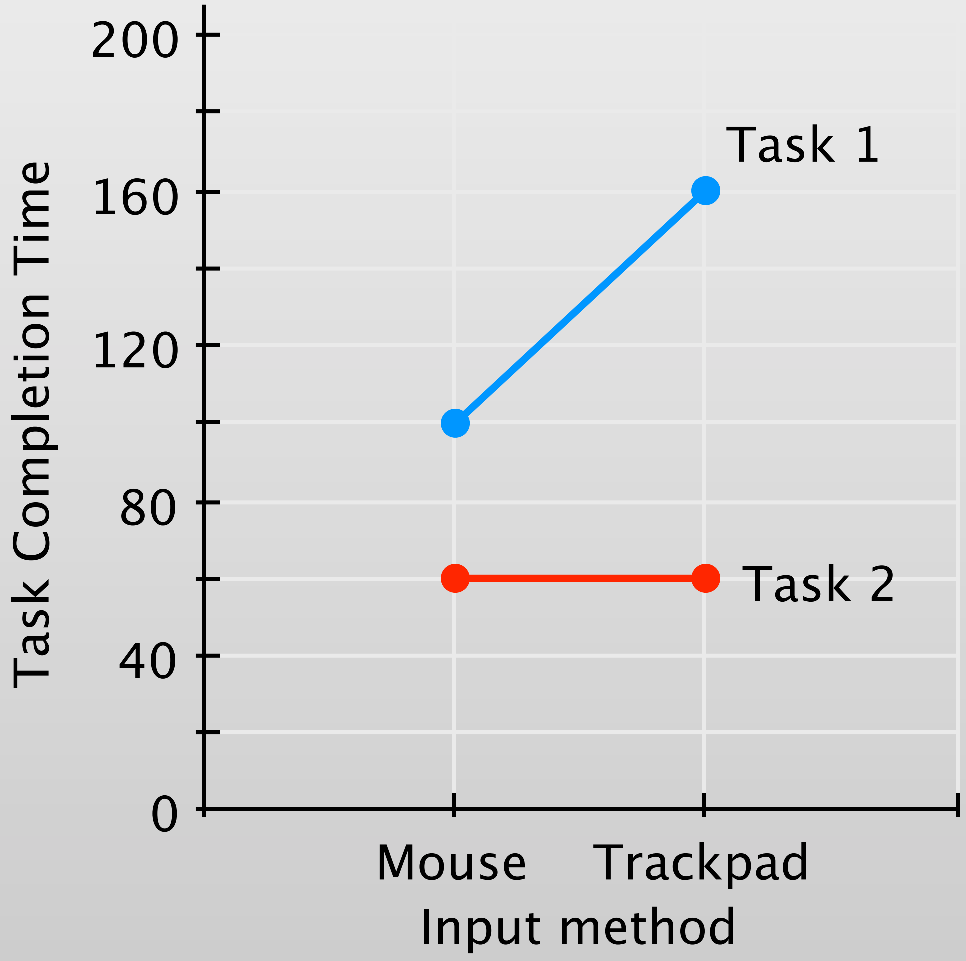
Interaction

- Effect of one independent variable depends on the particular level of another independent variable
  - Cannot conclude the effect of each independent variable overall
- Example: Does input method affect completion time in Task 1 and Task 2?
  - Interaction between task and input method
  - In Task 2, different input methods do not lead to different completion times
  - But in Task 1, they do



Simple Main Effect

- Solution: fix the level of one interacting variable (treat as two separate experiments – with lower *n*)
- In our example:
  - Different input methods do not cause differences in Task 2, but they cause differences in Task 1



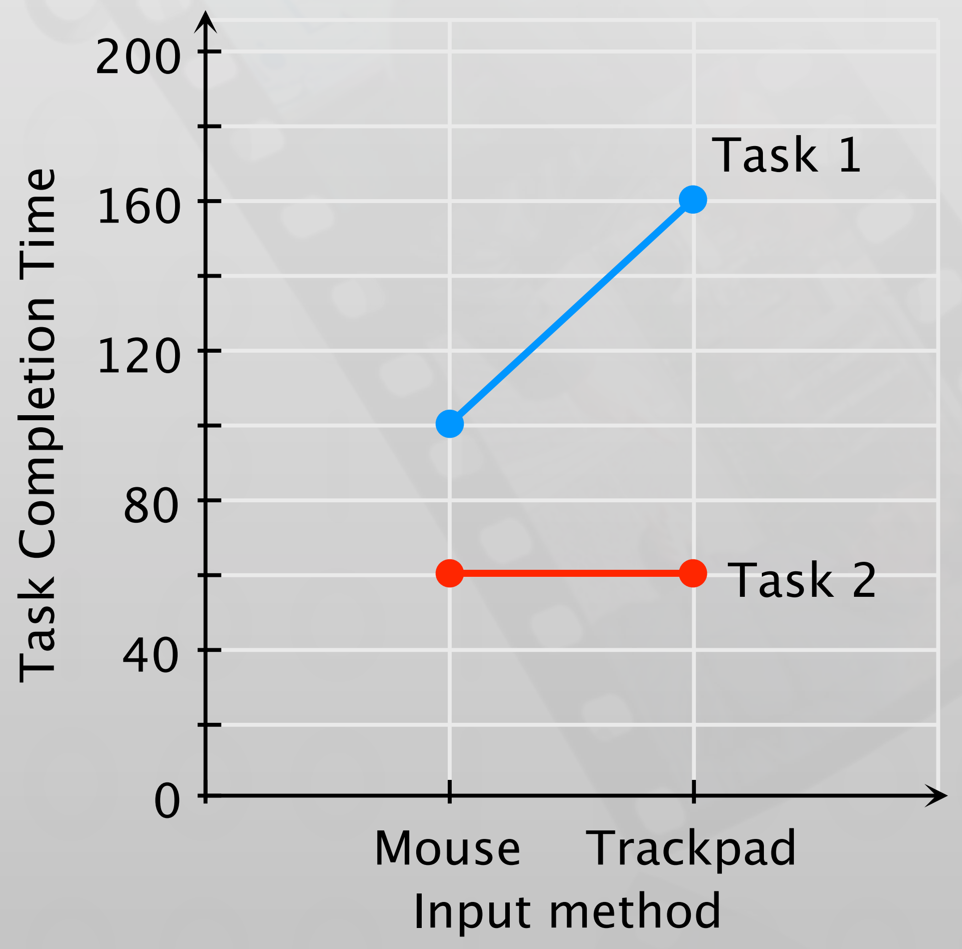
Demo: *N*-Way ANOVA with Interaction

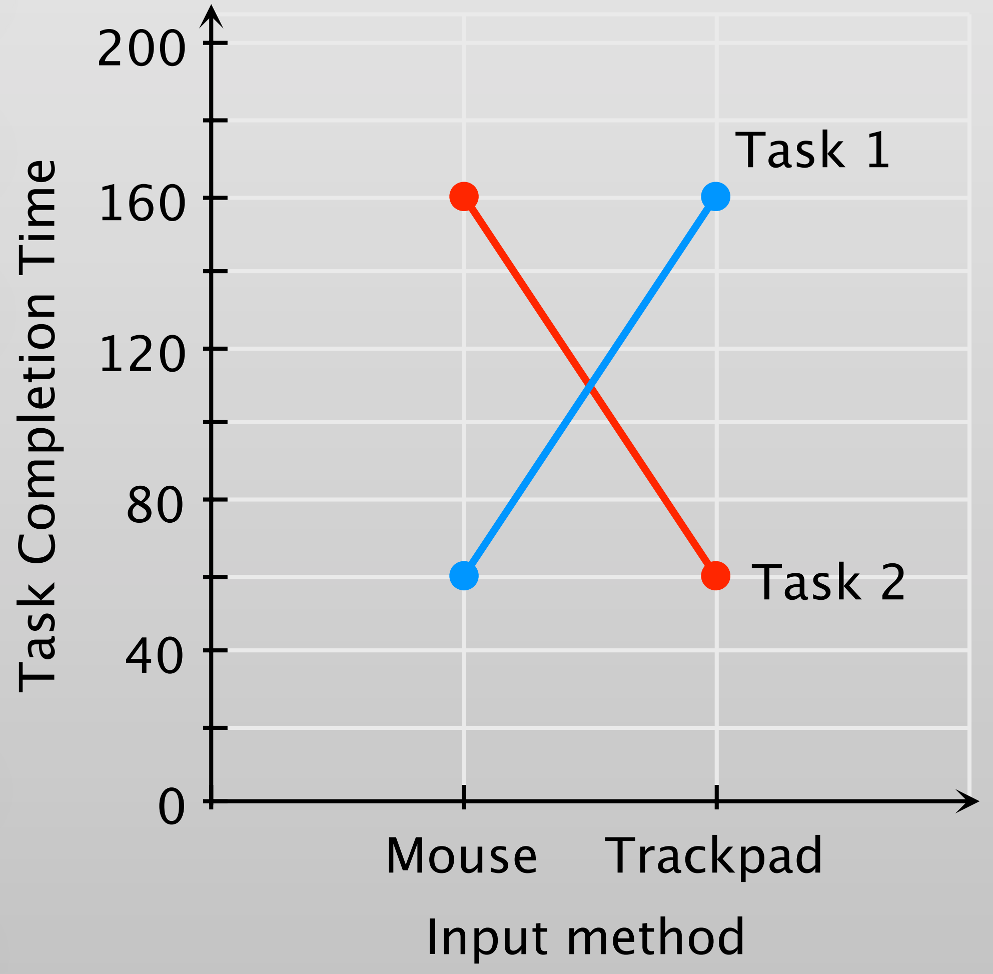
SLAPWidget.csv

Please follow along on your laptop!

In-class Exercise: Interaction Effects

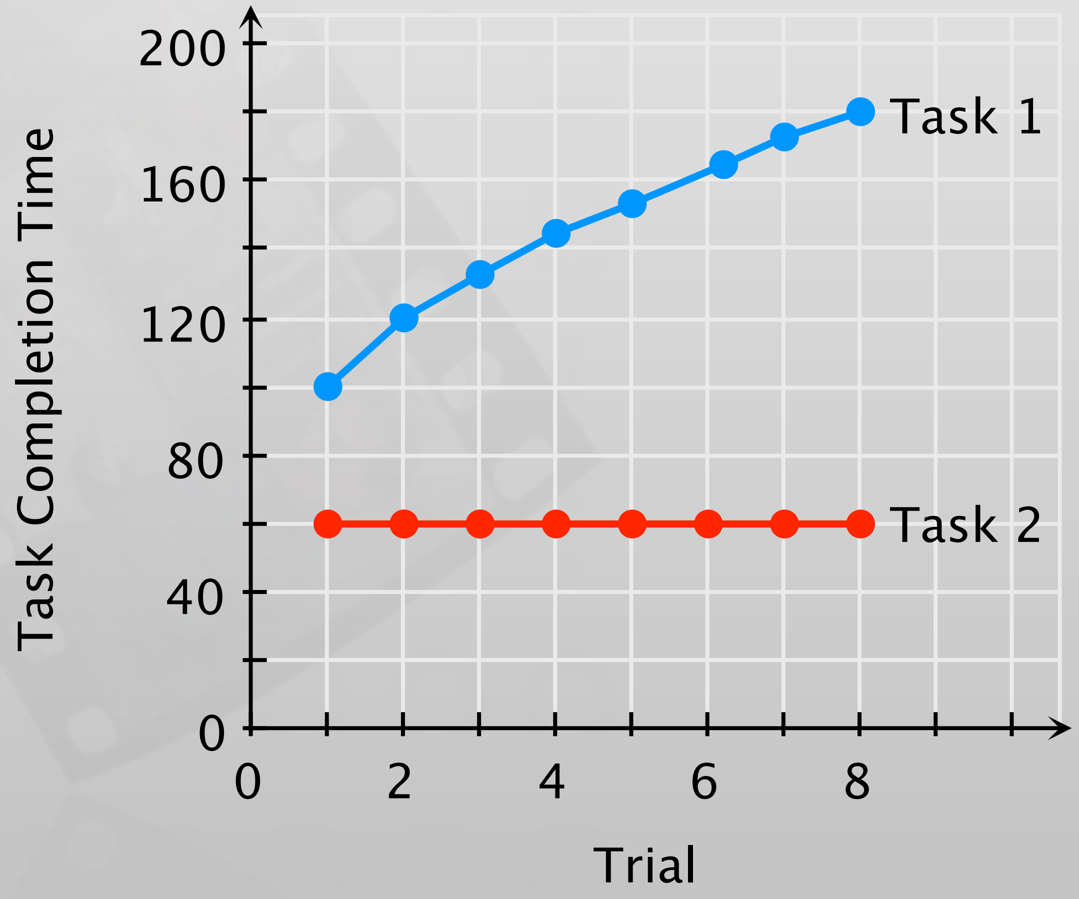
- Look at the following graphs. Make an educated guess whether there is a main effect, interaction, simple main effect, or nothing.





In-class Exercise: Interaction Effects

- Look at the following graphs. Make an educated guess whether there is a main effect, interaction, simple main effect, or nothing.



Within-groups:One-Way Repeated Measures ANOVA

- Used for within-groups design because it reduces differences caused by each participant from between-group differences

- More powerful in the same data set

- But: Sphericity assumptions
  - Variance between any two pairs of conditions do not differ significantly
  - Determined using Mauchly's sphericity test: cannot assume sphericity if p < .05
  - Assumption violated: Use corrected p values, e.g., Greenhouse-Geiser method

Demo:
Repeated Measures ANOVA

feedback.csv

Please follow along on your laptop!

Summary

- NHST supports alternative hypothesis by indicating that if null hypothesis is true, the measured data is unlikely
  - *p*-value: Asssuming that the null hypothesis was true, this is the probability that the data would occur as measured
- One-Way ANOVA partitions variance from between-groups factors
  - Tukey's Test: comparing all conditions pairwise to determine differences (post-hoc)
- Non-parametric tests: use only when parametric test assumptions are violated, e.g., non-interval data (Kruskal-Wallis something instead of ANOVA)
- Repeated-measure ANOVA does not assume independent samples. Use for within-groups design.
- Main effect, interaction, and simple main effect need to be identified when we have more than one IV

Beyond the Basics: What We Didn't Cover

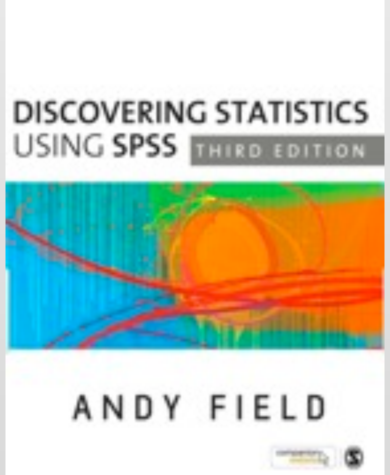
- Assumptions for statistical tests
  - We know: if the data is not interval, you cannot use ANOVA
  - There are more assumptions, e.g., normality of the data or equal variances.
  - There are statistical tests (Shapiro-Wilk, Bartlett) and visualizations (Q-Q plot) to check these assumptions
  - Use transformation to change data to a form suitable for analysis (with some trade-offs)
  - Bootstrap procedures allow you to analyze the data by re-sampling
- What to do if your results are *not* statistically significant
  - Try increasing the number of samples
  - Use power analysis to determine the number of samples needed

Beyond the Basics: What We Didn't Cover

- Counting and proportional data
  - Distribution differs from interval data
  - There are special tests for that, e.g., Chi-square
- Data from non-experiments (surveys,...)
  - Correlational statistics allow you to draw some conclusions
- Modeling and prediction
  - Linear or logistic regression allows you to create a model to predict output
  - E.g., Fitts' law assignment

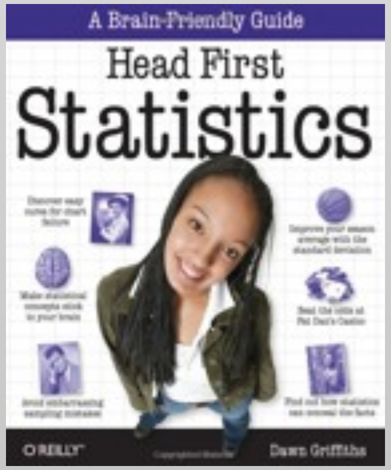
Want More?

- 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/
- Discovering Statistics Using SPSS by Andy Field
  - Easy to read, lots of examples, detailed explanations
  - SPSS is not required to understand the concepts
- Head First Statistics by Dawn Griffiths
  - Mostly basic statistics and probability theory
  - Helps getting the basics right for advanced understanding



DISCOVERING STATISTICS USING SPSS THIRD EDITION

ANDY FIELD



A Brain-Friendly Guide

Head First Statistics

O'REILLY

Dawn Griffiths

Theory

- ✓ Models of interaction
  - ✓ Affordances, mappings, constraints, types of knowledge, errors
- ✓ Design principles
- ✓ Human cognition and performance
- ✓ Interaction design notation
- ✓ History and vision of HCI

Practice

- ✓ Sketching
- ✓ User observation
- ✓ Iterative design
- ✓ Prototyping
- ✓ Ideation
- ✓ User studies and evaluation