Quantitative Evaluation 2

CS6501: Human-Computer Interaction

Seongkook Heo

Fall 2020, Department of Computer Science

University of Virginia | School of Engineering & Applied Science

Confounding Variables

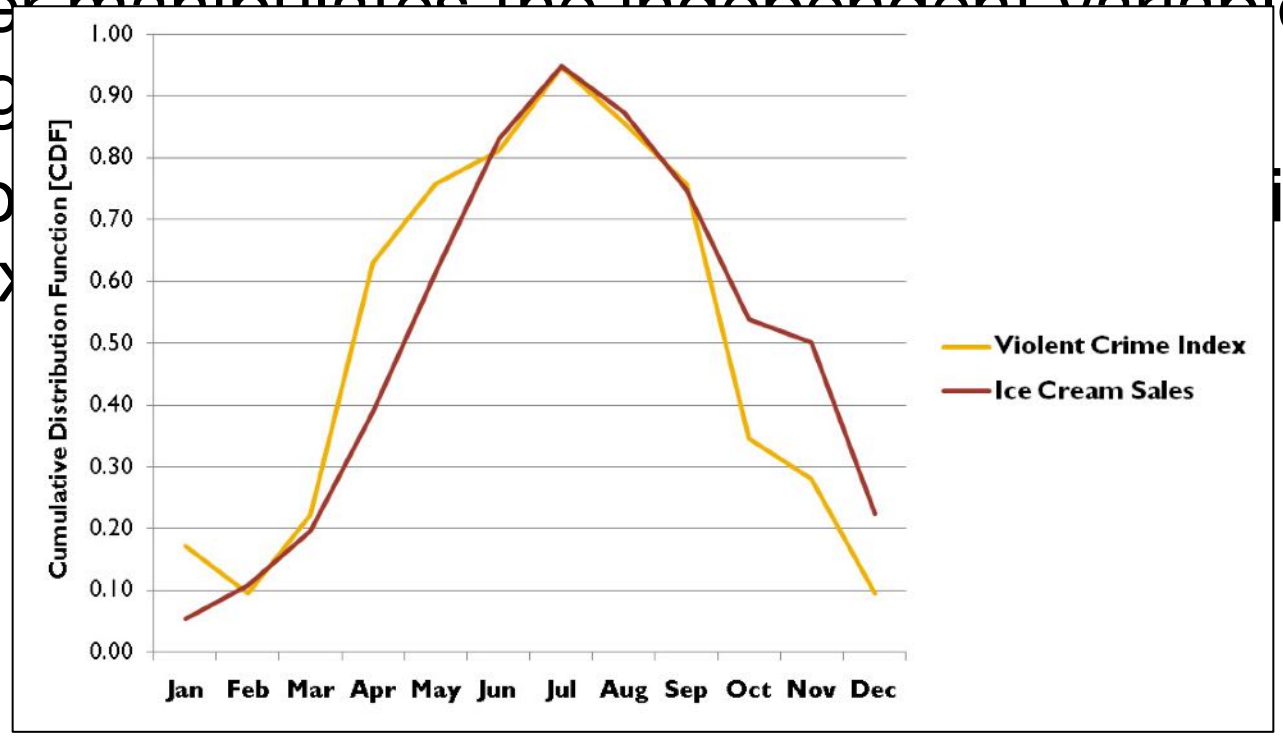
- Any circumstance that changes systematically as the experimenter manipulates the independent variable is a confounding variable.
- Identifying possible confounds is one of the most important jobs of an experimenter.

Confounding Variables

- Any circumstance that changes systematically as the experimenter manipulates the independent variable is a confounding variable.
- Identifying possible confounds is one of the most important jobs of an experimenter.

Confounding Variables

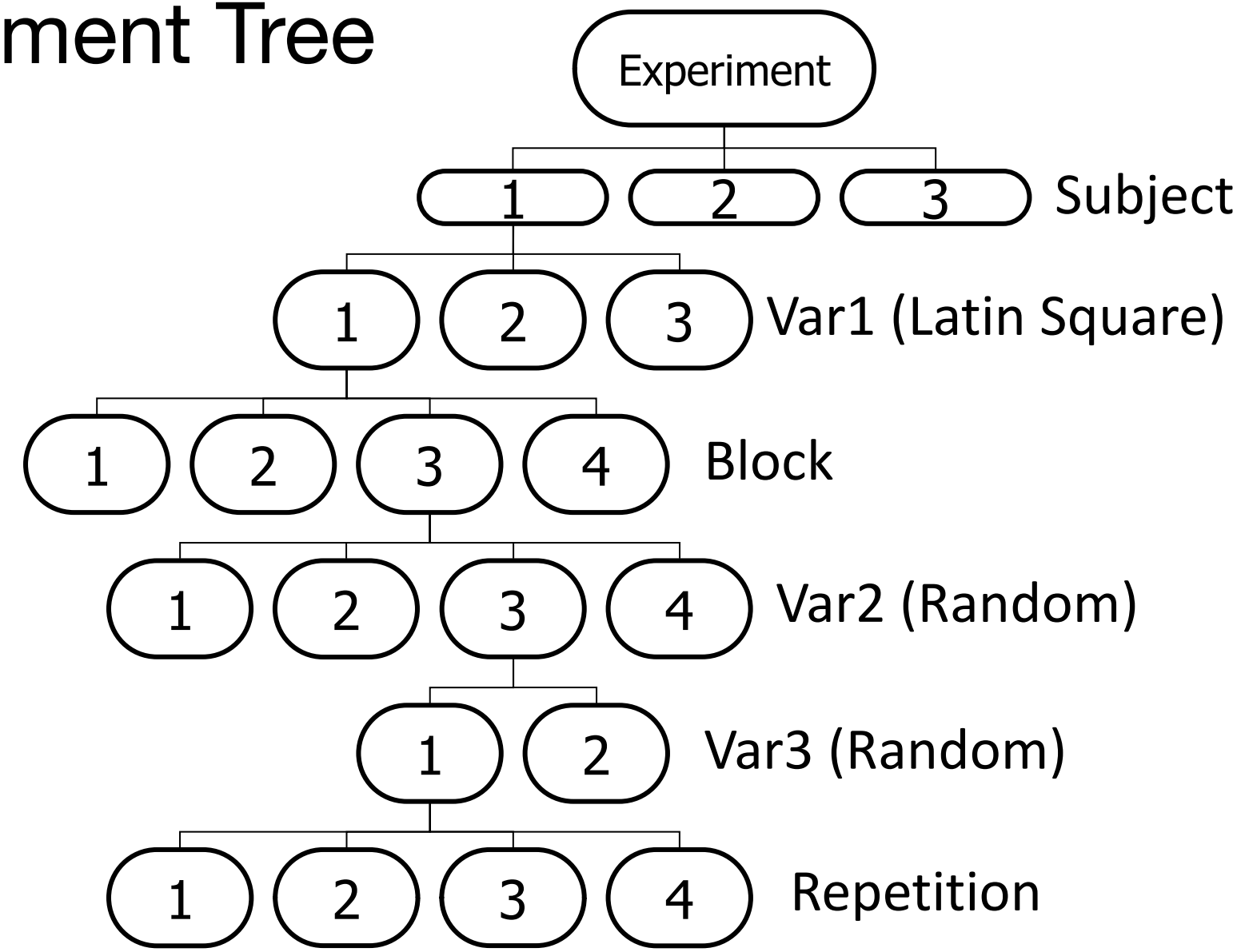
- Any circumstance that changes systematically as the experimente[illegible]e is a confounding
- Identifying p[illegible]important jobs of an ex



Other Definitions

- Block
  - A significant section of the experiment
  - Repeated to analyze learning
- Trial
  - An individual measurement for a single condition/cell
- Repetition
  - A trial which is repeated within a block
  - Increase number of data points, reliability
- Determining number of blocks/repetitions
  - Reasonable experiment duration
  - Enough data points for significant effects

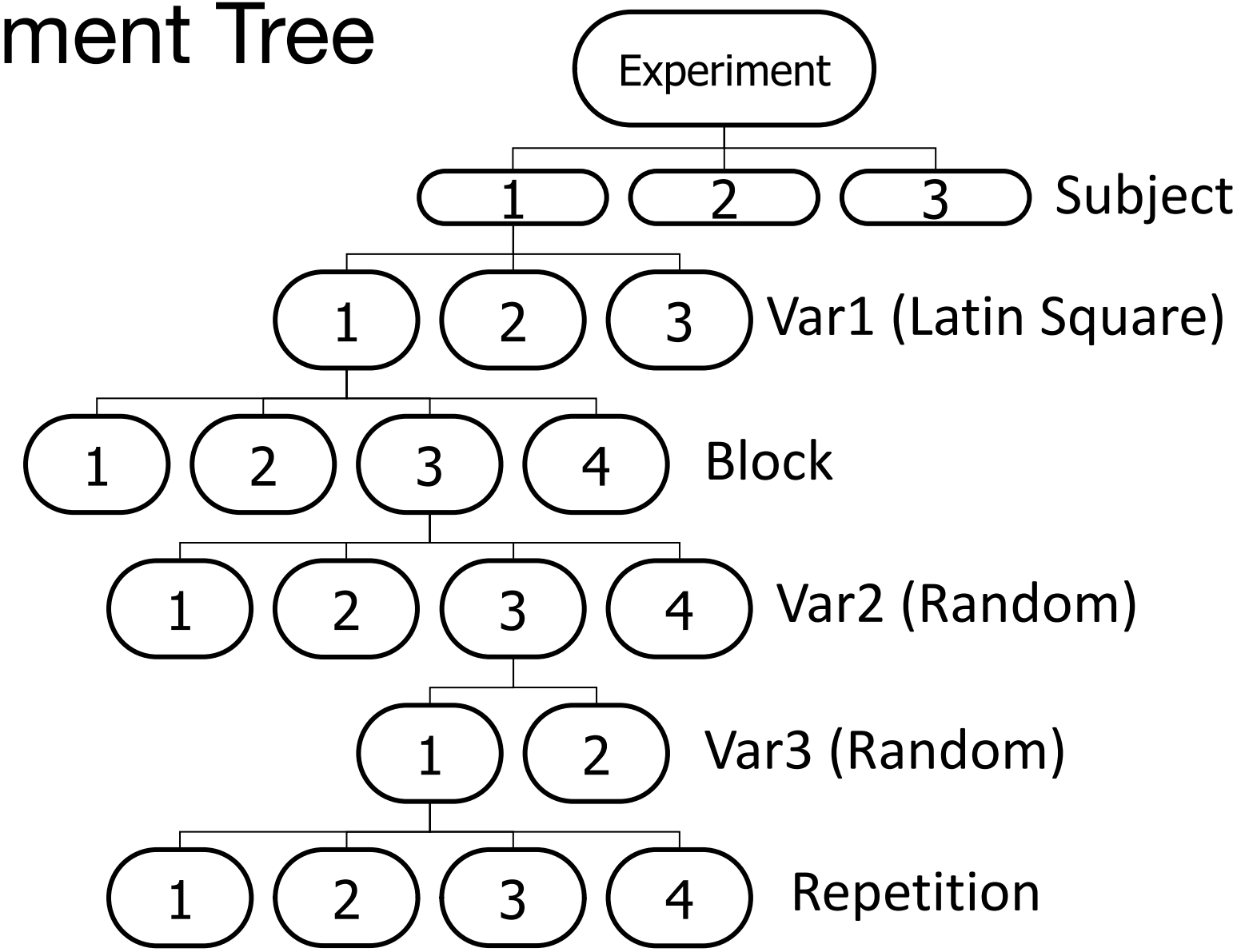
Example Experiment Tree



Example Experiment Tree

3 Subjects
x 3 Var1
x 4 Blocks
x 4 Var2
x 2 Var3
x 4 Repetitions

= 1152 Total Trials



Example Experiment Tree

3 Subjects
x 3 Var1
x 4 Blocks
x 4 Var2
x 2 Var3
x 4 Repetitions

= 1152 Total Trials

384 trials per subject
x 10s per trial

= 64 minutes per subject

64 minutes
+ setup
+ breaks
+ warmup trials
+ questionnaires

= 90 Minute Study

Reliability/Repeatability

- Would the same results be achieved if repeated?
  - Perfectly reliable if you get same results each time experiment is repeated
- Problems
  - Individual differences:
    - Best user 10x faster than slowest
    - Best 25% of users ~2x faster than slowest 25%
  - Unreliable instruments
    - Stretchable rubber ruler vs. steel ruler
- Partial Solution
  - Reasonable number and range of users tested
  - Correlate data from repeated measurements

Determining Reliability

- Test-retest
  - Repeat experiment on same group at a later time
- Alternative-form
  - Similar experiment given to same group at a later time
- Split-half
  - Split experiment results in half and analyze separately

Validity

- Are you measuring what you think you're measuring?
  - Errors in equipment
  - Errors in procedure
  - Incorrect pool of subjects
  - Errors in questions asked
  - Errors in variables measured

Validity

- Internal Validity
  - The degree to which the results are attributable to the independent variable and not some other rival explanation

- External Validity
  - The extent to which the results of a study can be generalized

Conducting an Experiment

Experiment Design

Conducting an Experiment

Conducting an Experiment

Experiments involve humans need IRB (Institutional review board) approval

Reviews research protocols and materials, such as

- Research methodology
- The risks or benefits
- The rights of the participants
- Anonymity and confidentiality

Conducting an Experiment

Participants can be recruited in various ways

- Flyers
- Online Forums
- Crowdworkers

But carefully consider how you can get the right participants: specify the conditions in detail in the recruitment ad.

Conducting an Experiment

Always run a pilot study

- Greet the participant
- Introduce the experiment, get a consent form signed
- Get demographic information and experience
- Give instructions to completing tasks
  – should be consistent across all participants

Be polite, professional, and neutral.

Conducting an Experiment

- **Check if data are valid**
- Analyze data using proper analysis methods,
  as you initially defined in the experiment design
- Do not only report the numbers and test results,
  discuss findings

- you are the most knowledgeable person for that experiment

Analyzing Results

- Observation:
  - How did the independent variables (IV) affect the dependent variables (DV)?
  - What type of trends occurred?
- Analysis:
  - What conclusions can be made?
  - How can future results be predicted?

Conveying Results

- What are the most important findings?
  - Based on fundamental questions
- How can the results be illustrated?
  - Graphs, charts, etc.

Research Hypotheses

- An experiment normally starts with a research hypothesis.
- A hypothesis is a precise problem statement that can be directly tested through an empirical investigation.

Card Play

- If I choose 10 cards, how many will be red?

When were you confident?

Consecutive Blacks	Probability	
1	1/2	0.5
2	1/4	0.25
3	1/8	0.125
4	1/16	0.063
5	1/32	0.031
6	1/64	0.016
7	1/128	0.008
8	1/256	0.004
9	1/512	0.002
10	1/1024	0.001

p < .05

Types of Hypotheses

- **Null hypothesis**

  Typically states that there is no difference between experimental treatments

- **Alternative hypothesis**

  A statement that is mutually exclusive with the null hypothesis

- **Goal of experiment**

  Typically to find statistical evidence to reject the null hypothesis in order to support the alternative hypothesis

Types of Hypotheses

- **Null hypothesis**

  The chance of drawing a red card and a black card is equal

Types of Hypotheses

- **Null hypothesis**

  The chance of drawing a red card and a black card is equal

- **Alternative hypothesis**

  Something fishy is going on…

Types of Hypotheses

- **Null hypothesis**

  The chance of drawing a red card and a black card is equal
- **Alternative hypothesis**

  Something fishy is going on…
- **Statistical evidence and conclusion**

  The probability of obtaining the result that we did (10 blk in a row) was 0.001.

Types of Hypotheses

- **Null hypothesis**

  The chance of drawing a red card and a black card is equal
- **Alternative hypothesis**

  Something fishy is going on…
- **Statistical evidence and conclusion**

  The probability of obtaining the result that we did (10 blk in a row) was 0.001.

  ➔ Therefore, reject the null hypothesis

Types of Hypotheses

- **Null hypothesis**

  The chance of drawing a red card and a black card is equal

- **Alternative hypothesis**

  Something fishy is going on…

- **Statistical evidence and conclusion**

  The probability of obtaining the result that we did (10 blk in a row) was 0.001.

  ➔ Therefore, reject the null hypothesis

  ➔ **Professor is a trickster!**

What is Hypothesis Testing?

- The use of statistical procedures to answer research questions
- Typical research question (generic):

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

- For hypothesis testing, we instead use a statement:

> There is no difference in the mean time to complete a task using Method A vs. Method B.

- This is the null hypothesis (assumption of “no difference”)
- Statistical procedures can be used to reject the null hypothesis

Type I and Type II Errors

- All significance tests are subject to the risk of Type I and Type II errors
- Type I error (also called a “false positive”):
  - Rejecting the null hypothesis when it is true
- Type II error (also called a “false negative”):
  - Not rejecting the null hypothesis when it is false
- It is generally believed that Type I errors are worse than Type II errors
  - A Type I error may result in a condition worse than the current state
  - A Type II error can cost the opportunity to improve the current state

Type I and Type II Errors

		Study conclusion	
		No difference	Touchscreen ATM is easier to use
Reality	No difference	√	Type I error
	Touchscreen ATM is easier to use	Type II error	√

Traditional ATM or Touchscreen ATM easier to use?

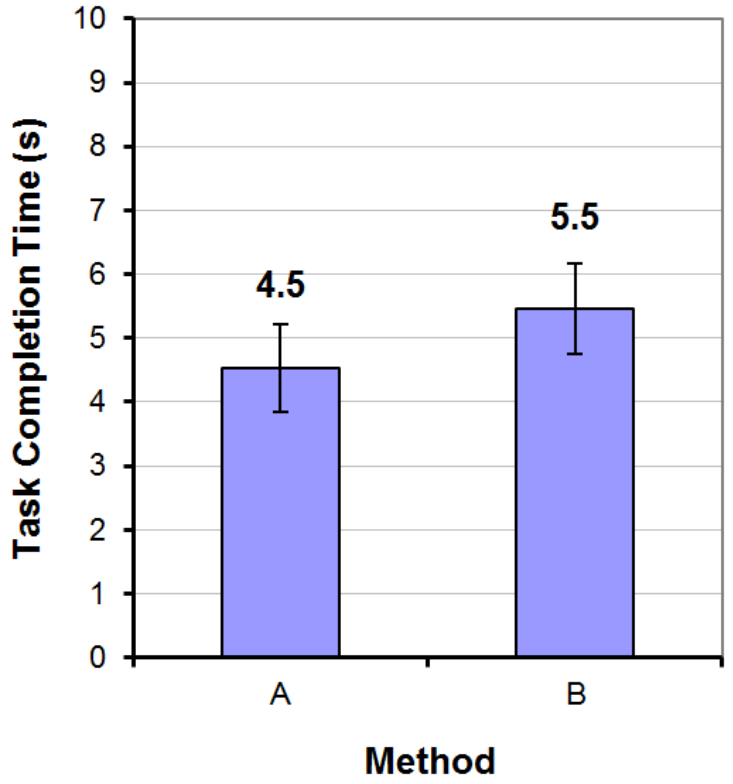
- It is generally believed that Type I errors are worse than Type II errors
  - A Type I error may result in a condition worse than the current state
  - A Type II error can cost the opportunity to improve the current state

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

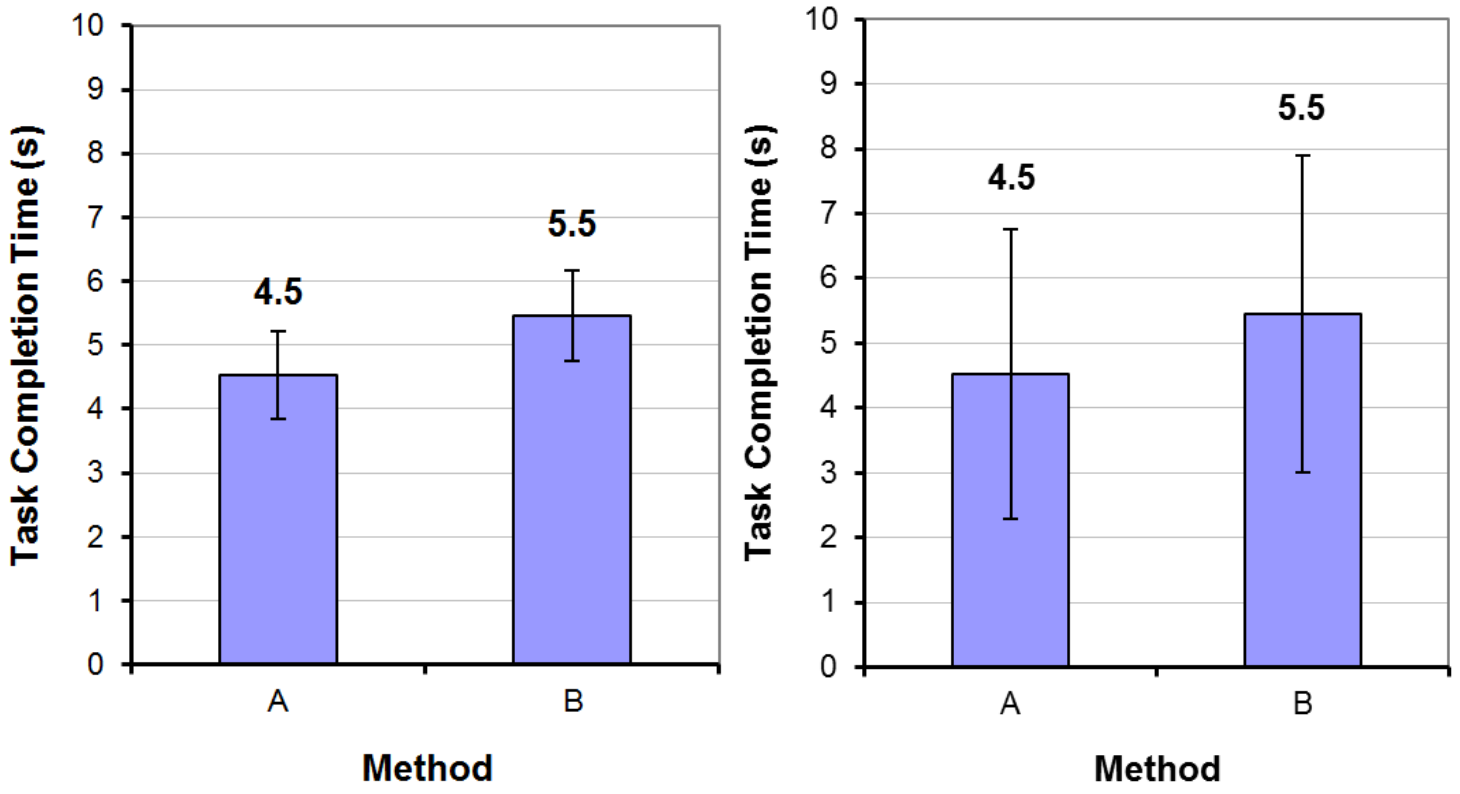
Descriptive Statistics

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



Descriptive Statistics

- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of spread
  - 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 does have an effect on DV y

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?

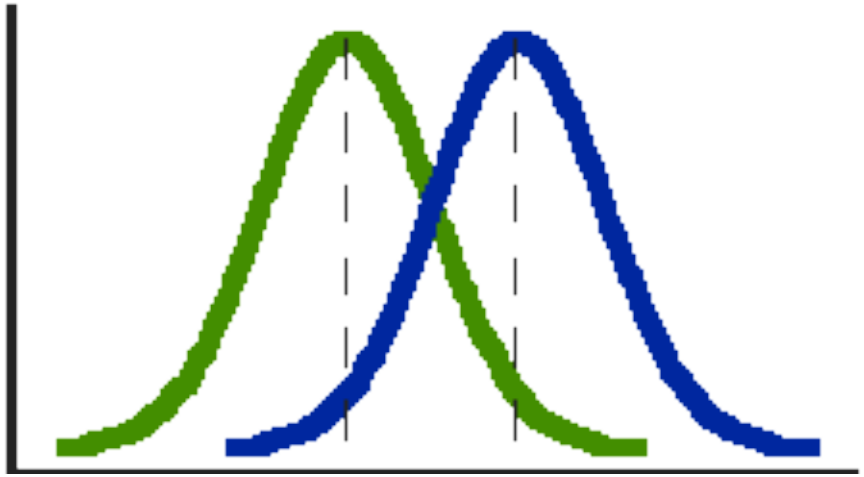
- Seems odd that we analyze the variance, when the research question is concerned with the overall means:

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

- Let's explain through the t-test…

Comparing Two Means: t-test

- Test if means are statistically different
- Equation produces t value
- t value maps to a probability



$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Comparing Two Means: t-test

- Independent-samples t test: between-group design

Group	Participants	Task completion time	Coding
No prediction	Participant 1_a	245	0
No prediction	Participant 2_a	236	0
No prediction	Participant 3_a	321	0
No prediction	Participant 4_a	212	0
No prediction	Participant 5_a	267	0
No prediction	Participant 6_a	334	0
No prediction	Participant 7_a	287	0
No prediction	Participant 8_a	259	0
With prediction	Participant 1_b	246	1
With prediction	Participant 2_b	213	1
With prediction	Participant 3_b	265	1
With prediction	Participant 4_b	189	1
With prediction	Participant 5_b	201	1
With prediction	Participant 6_b	197	1
With prediction	Participant 7_b	289	1
With prediction	Participant 8_b	224	1

Comparing Two Means: t-test

- Paired-sample t test: within-group design

Participants	No prediction	With prediction
Participant 1	245	246
Participant 2	236	213
Participant 3	321	265
Participant 4	212	189
Participant 5	267	201
Participant 6	334	197
Participant 7	287	289
Participant 8	259	224

Comparing Two Means: t-test

- Test if means are statistically different
- Equation produces t value
- t value maps to a probability
  - Lower variance -> Higher t value -> Lower probability
- Only compares two groups

Project Proposal (Presentation + Document)

- You will need
  - HCI Problem
    - The problem you want to solve
  - Related work
    - What others have done
  - Research Question
    - What you want to know by conducting this research
  - Method
    - What you suggest or design

Project Proposal (Presentation + Document)

- **Project Proposal (Due Sep 25)**
  - Similar to Introduction of a CHI paper
  - Three sections
    - Motivation & Background
    - **Related Work**
    - Research Question and Method

- 3+ papers that are relevant to the problem
- 2+ papers that are relevant and aligned with the solution you are suggesting
- (optional) suggesting a similar solution but used for other problems
- (optional) study papers that help understanding the problem

Project Proposal (Presentation + Document)

- **Project Presentation (Sep 28)**
  - Each team will have up to 10 minutes
  - Should include all four components
  - We will have a shared document for feedback and questions (and this is where participation counts)
  - If you can’t join live, you can record and send the video

Team Discussion

Thank you!