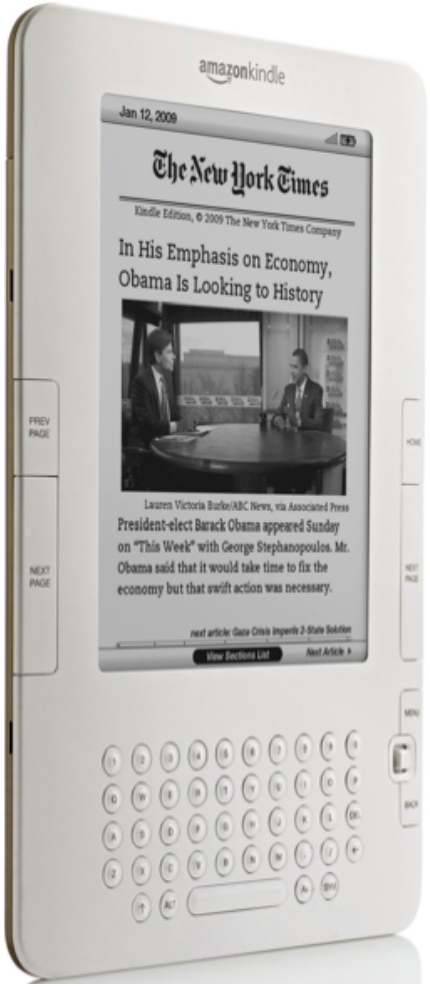
Lecture 18: Controlled experiments

April 14

Kindle vs. iPad



Typical steps to carry out controlled experiments in HCI

- Design
  - State a lucid, testable hypothesis
  - Identify independent and dependent variables
  - Design the experimental protocol
  - Choose the user population
- Run
  - Apply for human subjects protocol review (IRB)
  - Run some pilot participants
  - Modify and finalize the experimental protocol
  - Run the experiment
- Discover
  - Perform statistical analysis
  - Draw conclusion
  - Communicate results

Design

Designing experiments

1. State a lucid, testable **hypothesis**
2. Identify independent and dependent **variables**
3. Design the experimental **protocol**
4. Choose the user **population**

Usability testing vs. User testing

- Usability testing
  - Questions about the usability of a UI
- User testing
  - Questions about users

- We focus on usability testing

Typical research questions

- Is this design viable?
- Is it as good as or better than certain alternatives?
- Which of several alternatives is best?
- What are its performance limits and capabilities?
- What are its strengths and weaknesses?
- Does it work well for novices, for experts?
- How much practice is required to become proficient?

Hypothesis

"iPad is better than Kindles"

- Is it testable hypothesis?

Hypothesis

"iPad is better than Kindles"

- Is it testable hypothesis? **NO**

- Broad questions are not testable.
- Broad questions can be investigated by posing multiple narrow testable questions

Hypothesis

"iPad is better than Kindles"

- Is it testable hypothesis? **NO**

- What feature?
- What task?
- What measurement?
- What population?

Hypothesis

"iPad is better than Kindles"

- Is it testable hypothesis? **NO**

- What feature? **keyboard**
- What task? **typing**
- What measurement? **speed**
- What population? **College students**

Hypothesis

"College students (population) type (task) faster (measurement) using iPad's keyboard (feature) than using Kindle's keyboard"

Can still be even more narrow

e.g., in a classroom

Variables

- Independent
  - Things we want to compare
- Dependent
  - Things we want to measure
- Control
  - Things we don't want to interfere
- Nuisance
  - Thins we forgot to control

Variables

- Independent
  - Kindle or iPad keyboard
- Dependent
  - Typing speed
- Control
  - College students, age, experience
- Nuisance
  - Weather

Population

- Target population
  - E.g., College students
- Sample from this population
  - E.g., UMD students

- Inclusion criteria
- Exclusion criteria

- How many subjects do we need?
  - Depends on how diverse the population is
- How do we know we have enough subjects?
  - At the very least when there's statistical significance

Protocol

- What tasks should each subject perform?
  - E.g., what sentence to type
- Which conditions should each subject be tested on?
  - Within subject vs. between subject
- What to measure?
  - Task completion time

Protocol

- Within subject
  - Each subject is tested on all conditions
  - More efficient (need fewer subjects)
  - Each person is his or her own control
  - Need to deal with "order effects"
- Between subject
  - Each subject is tested on one condition
  - Simpler design and analysis
  - Easier to recruit participants (only one session)
  - Less efficient (need more subjects)

What will Simon say?

How will others criticize our finding using this design?

- Is it **internally** valid?
  - Could the typing speed differences be explained by other factors, such as technical experiences?
- Is it **externally** valid?
  - How about other college students?
  - How about high school students?
  - How about outside of classrooms?
  - How about other sentences?

Internal validity

- Effects are due to the test conditions
- Differences in the means are due to the test conditions
- Variances are due to participant differences
- Other potential sources of variance are controlled

External validity

- Findings are generalizable to other people and other situations?
- Participants represent broader intended population of users
- Test environment and experimental procedures represent real world situations where the UI will be used

Designing experiments

1. State a lucid, testable **hypothesis**
2. Identify independent and dependent **variables**
3. Design the experimental **protocol**
4. Choose the user **population**

Run

Running experiments

1. Apply for human subjects protocol review (**IRB**)
2. Run some **pilot** participants
3. Modify and **finalize** the experimental protocol
4. Run the **actual** experiment

- *In a web site usability study, a 12 year old girl who wanted information on the White House was supposed to type in www.whitehouse.gov but instead typed www.whitehouse.com and was suddenly transported to a porn site.*

IRB

- Institutional Review Board (IRB)
- Goals:
  - Review all research involving human (or animal) participants
  - Safeguard the participants
- Required for proposed research that involves
  1. intervention or interaction with human subjects
  2. the collection of identifiable private data on living individuals
  3. data analysis of identifiable private information on living individuals

Informed Consent: Why?

- People can be sensitive about the research process and issues
- Errors will likely be made, participants may feel inadequate
- Studies may be mentally or physically strenuous

Informed Consent: Content

- Purpose of the research Study
- What you will be asked to do in the study
- Time Required
- Risks and Benefits
- Compensation
- Confidentiality
- Voluntary participation
- Right to withdraw from study
- Whom to contact for questions about the study
- Whom to contact about your rights as a research participant in this study
- Agreement signatures

Pilot test

- Why pilot test?
  - There are always unexpected problem.
  - It's not too late to make chances to the experimental protocol.
  - Get a feel whether the hypothesis can be verified
  - It's not too late to stop the experiment if results don't seem promising.

Modify and finalize the protocol

- Improve instruction
- Change the number of tasks
- Adjust the allotted time for each task
- Fix UI bugs

- Once the protocol is finalized and the actual experiment is running, you can not change it any more

Running the actual study: Before the study

- Prepare well
- Don't let participants wait
- Maintain privacy
- Explain procedures without compromising results
- Let participants know they can quit anytime
- Ask participant to sign a consent form

During each session

- Make sure participant is comfortable
- Session should not be too long

After the session

- Debrief the participant
  - State how findings will help improve UI
  - Explain errors and failures are not their problems
  - Show how to perform failed tasks
- Do not compromise privacy
  - Store data anonymously, securely, or destroyed

Analyze

Analyzing results

1. Perform **statistical** analysis
2. Draw **conclusion**
3. Communicate results

Scenario: Kindle vs. iPod

- Population
  - All college students
- Sample
  - Some UMD students
- Independent variable: iPad or Kindle
- Dependent variable: typing time

Subj	Kindle Time (s)	iPad Time (s)
1	43	34
2		33
3	43	36
4	35	31
5	36	41
6	39	39
7	42	5
8	43	29
9	41	30
10	39	41

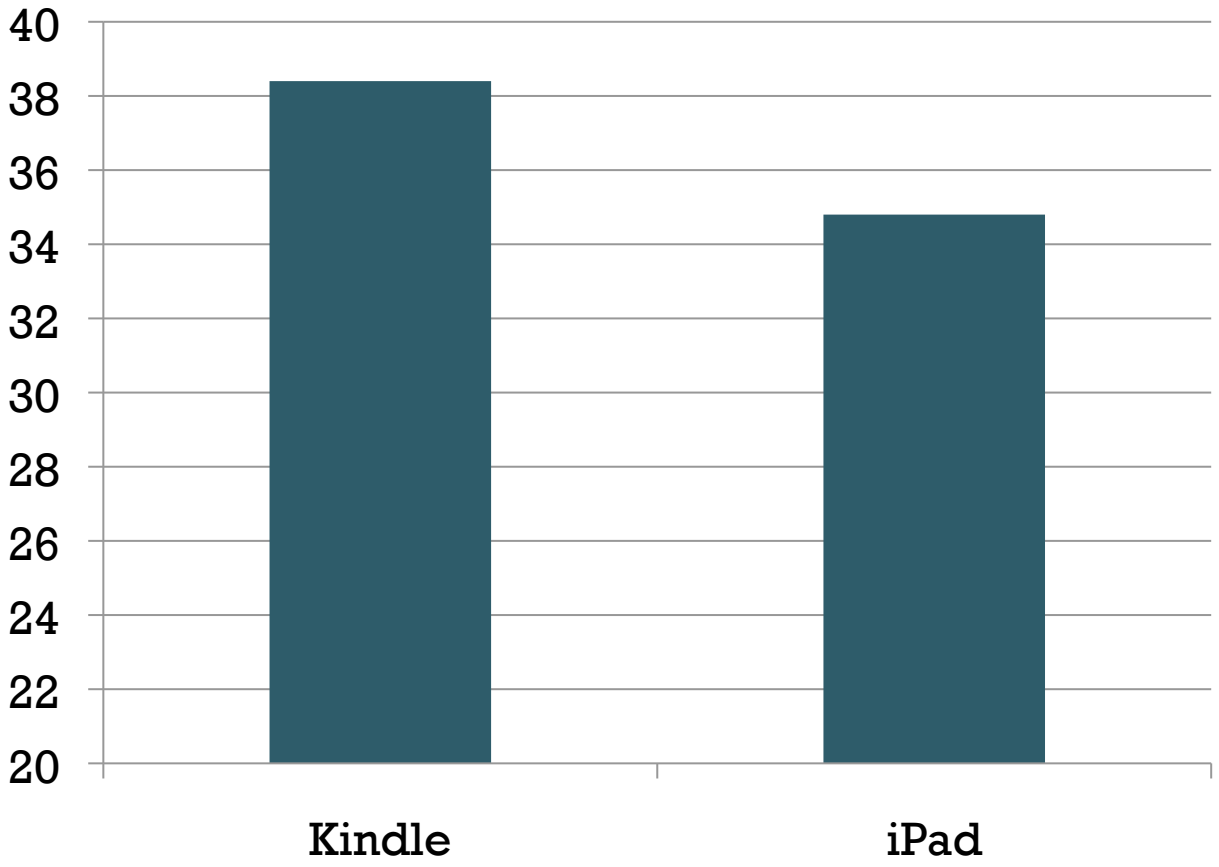
Pop quiz:

Between subject or within subject design?

Cleanup the data

- Are there outliers?
- Are there junk or missing data?
  - Some participants may have fallen asleep
  - Some participants may do random things just to earn the money
  - Recoding devices may have failed a couple times

Subj	Kindle Time (s)	iPad Time (s)
1	43	34
2		33
3	43	36
4	35	31
5	36	41
6	39	39
7	42	5
8	43	29
9	41	30
10	39	41

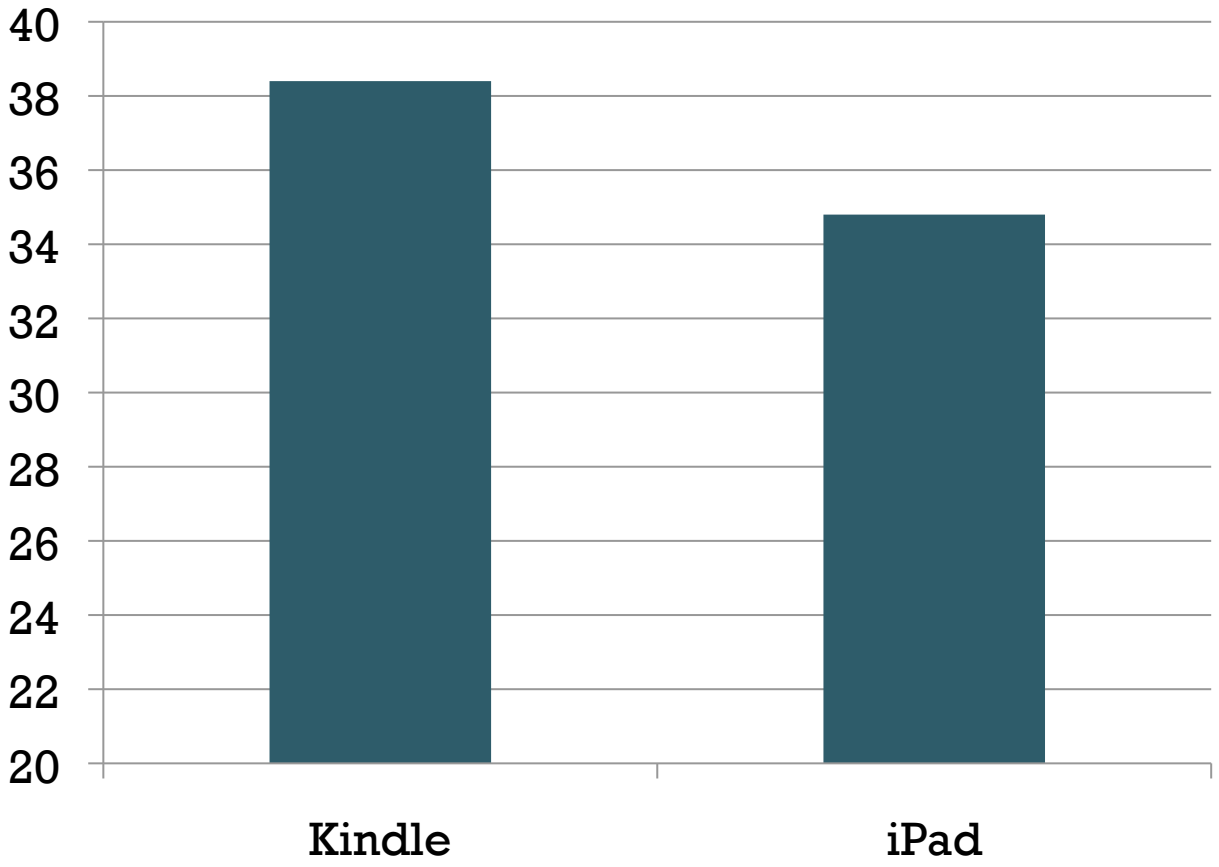

40
38
36
34
32
30
28
26
24
22
20
Kindle
iPad

Statistical questions

- Descriptive questions (What)
  - What is the typical performance?
  - How large are the differences between individuals?
- Analytical questions (yes or no)
  - Is there a difference?
  - Is the difference large or small?
  - Is the difference significant or due to chance?

Statistical tools

- Descriptive statistics (what)
  - Mean
  - Median
  - Standard deviation
  - Correlation
  - Regression
- Analytical statistics (yes or no)
  - T-test
  - ANOVA
  - Post-hoc test


40
38
36
34
32
30
28
26
24
22
20
Kindle
iPad

- Is this difference significant?
- What if there's no inherent difference?
- If two are the same, can we get this result by chance?
  - We may happen to select those with higher values from one group and those with lower values from another group.

T-test

- Goal:
  - Test if the difference in sample means can be explained by the difference in true means
- Two sided (or two tails)
  - There is a difference
- One sided (or one tail)
  - There is a difference, and the difference goes in a particular direction

Scenario: Kindle vs. iPod

- Population
  - All college students
- Sample
  - Some UMD students
- Independent variable: iPad or Kindle
- Dependent variable: typing time

T-test: Example

- Assume that the populations using Kindle and iPad ***are not*** different in terms of the task completion time
- Compute the mean task completion time of the two samples:
  → 35 vs 38
- Compute the difference between the two means
  → 3
- Compute the chance of observing this much difference
  → 2%
- Since 2% is so low, there may be a contradiction
- Thus, the assumption is unlikely to be true.
- Thus, the population means ***are*** different.

T-test

- Analogous to proof by contradiction
  - Assume that the true means of the two populations ***are not*** different
  - Compute the means of the two samples
  - Compute the difference between the two sample means
  - Compute the chance of observing this much difference
  - If the chance is low, this seems contradictory.
  - Thus, the assumption is unlikely to be true.
  - Thus, the true means ***are*** different.

T-test

- A H0: null hypothesis ontradiction
  - Assume that the true means of the two populations ***are not*** different
  - Compute the means of the two samples
  - Compute the difference between the two sample means
  - Compute the chance of observing this much difference
  - If the chance is low, this seems contradictory.
  - Thus, the assumption is unlikely to be true.
  - Thus, the true means ***are*** different.

T-test

- Analogous to proof by contradiction
  - Assume that the true means of the two populations ***are not*** different
  - Compute the means of the two samples
  - Compute the difference between the two sample means
  - Compute the chance of observing this much difference
  - If the chance is low, this seems contradictory.
  - Thus, the assumption is unlikely to be true.
  - Thus, the true means ***are*** different.

p value

T-test

- Analogous to proof by contradiction
  - Assume that the true means of the two populations ***are not*** different
  - Compute the means of the two samples
  - Compute the difference between the two sample means
  - Compute the chance of observing this much difference
  - adictory.
  - e true.
  - Thus, the true means ***are*** different.

H1: alternative hypothesis

T-test

- Analogous to proof by contradiction
  - Assume that the true means of the two populations ***are not*** different
  - Compute the means of the two samples
  - Compute the difference between the two sample means
  - Compute the chance of observing this much difference
  - If the chance is low, this seems contradictory.
  - Thus, the assumption is unlikely to be true.
  - Thus, the true means ***are*** different.

$p < 0.05$

P = 0.05

- If they ***are not*** different, the chance of seeing at least this much difference between the two samples is 5%.

- If the efficiency of iPad and Kindle ***are not*** different, the chance of seeing at least 3 sec. difference between two samples is 5%.

P = 0.25

- If they ***are not*** different, the chance of seeing at least this much difference between the two samples is 25%.

- If the efficiency of iPad and Kindle ***are not*** different, the chance of seeing at least 3 sec. difference between two samples is 25%.

p = 0.05: wrong interpretations

- This result is 95% right.
- This result is invalid only in 5 out of 100 people.
- This result is invalid only in 5 out of 100 trials.
- The actual difference is 5%.

Draw conclusions

- Typically $p < 0.05$ is considered statistically significant and can be scientifically published.

- In some cases, $p < 0.01$ is desired
  - when we may not know which hypothesis to test and just choose one randomly.
  - E.g., one of many possible causes of cancer

Advanced analysis techniques

- ANOVA
  - More than one independent variables
    - E.g., iPad, Kindle, Nook
- Post-hoc test
  - Find promising pairs using ANOVA
  - Run T-Test on that pair
  - Requires a stricter p value

Communicate

- Translate the stats into words so regular people can understand

- "the numbers suggest computer experience may influence performance especially speed"

Feeding back into design

- Found 2 sec. saving per operation.
- Quantitative terms
  - Translated into monetary saving
- Qualitative terms
  - Workers are happy, less stressed, less tired.