HCI and Design

Admin

Assignment 5 is due today at 11:59pm!

No more assignments (you’re welcome)

Today:

Experimental Research

Types of HCI studies

Descriptive investigations focus on constructing an accurate description of what is happening.

Relational investigations enable the researcher to identify relations (correlations) between multiple factors. However, relational studies can rarely determine the causal relationship between multiple factors.

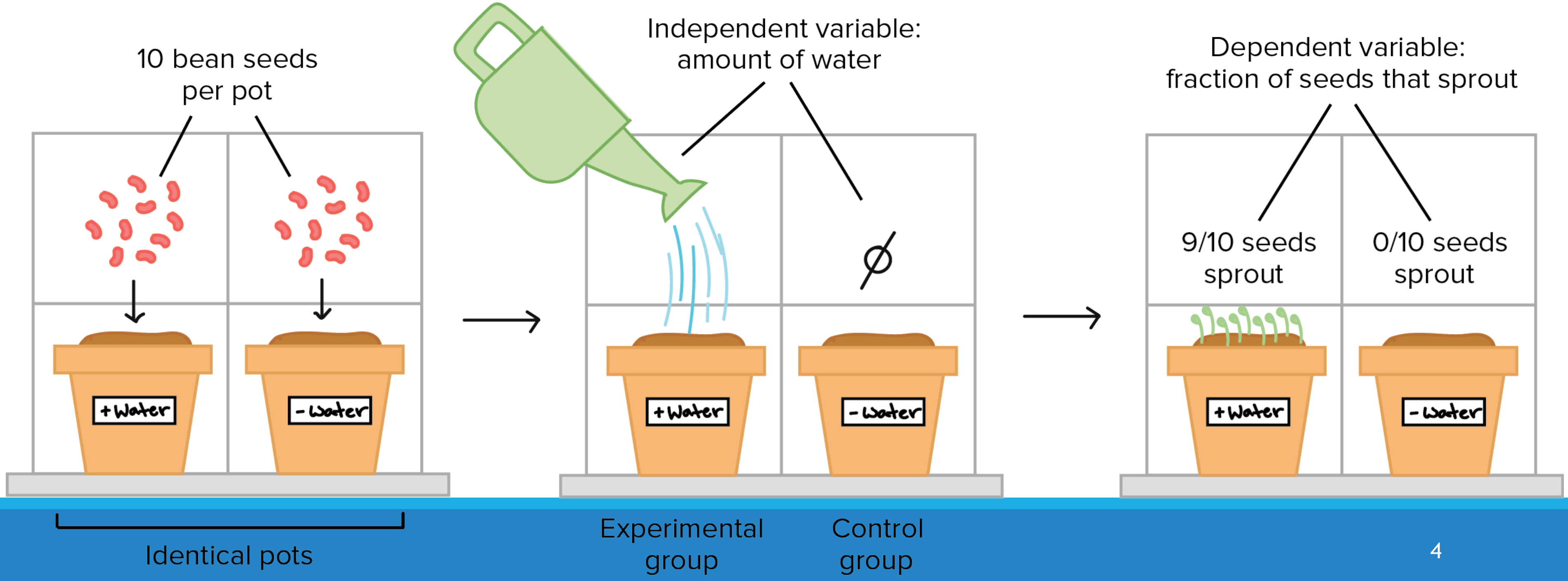
Experimental research allows the establishment of a **causal relationship**. Usually these are controlled experiments.

Why bother with experiment design?

To establish strong evidence linking manipulated **treatments** to changes in one or more **outcomes**.

To determine **causation.**

- Changes to x cause changes to y in this measurable way.



Hypotheses

An experiment normally starts with a hypothesis.

A hypothesis is a precise problem statement that can be directly tested through an empirical investigation.

Compared with a theory, a hypothesis is a smaller, more focused statement that can be examined by a single experiment.

Example: “The iOS virtual keyboard is faster and more accurate than the Android virtual keyboard.”

Null hypothesis

Null hypothesis: typically states that there is no difference between experimental treatments.

- *e.g., There are no detectable differences in the speed or accuracy of the iOS keyboard and the Android keyboard.*

The goal of an experiment is to find statistical evidence to confirm or reject null hypotheses in a reliable fashion.

A hypothesis should specify the independent variables and dependent variables.

Independent Variables

Independent variables are things the experimenter manipulates.

Independent variables (IV): the factors that the researchers are interested in studying or the possible “cause” of the change.

- IV is independent of a participant’s behavior.
- IV is usually the treatments or conditions that the researchers can control.

Typical independent variables in HCI

Those that relate to technology

- Types of technology or device (e.g. keyboard type)
- Types of design (e.g. design A vs. B)

Those that relate to users: age, gender, computer experience, professional domain, education, culture, motivation, mood, and disabilities

Those that relate to context of use:

- Physical status
- User status
- Social status

Dependent Variables

Dependent variables are things the experimenter measures.

Dependent variables (DV) refer to the outcome or effect that the researchers are interested in.

- DV is dependent on a participant's behavior or the changes in the IVs
- DV is usually the outcomes that the researchers need to measure.

Typical dependent variables in HCI

Efficiency:

- e.g., task completion time, speed

Accuracy:

- e.g., error rate

Subjective satisfaction:

- e.g., Likert scale ratings

Ease of learning and retention rate

Physical or cognitive demand

- e.g., NASA task load index

Factors

Same as independent variables.

An experiment with a control group and a treatment group is a single-factor (or one-way) experiment.

Example:

- Two groups:
  - *treatment* gets broccoli every morning,
  - *control* does not.
- The factor or independent variable might be called *food*.
- The measure or dependent variable is HCI test score.

Levels

Levels are values a factor can assume (i.e. groups).

Examples:

- Factor food has two levels: broccoli, no-broccoli
- Factor keyboard has two levels: iOS, Android
- Factor posture has three levels: sitting, standing, walking

Finding differences among levels is what an experiment is all about.

Between-subjects design

Each participant (subject) experiences only one level of a factor

- requires more participants
- but avoids possible confounds
- easier to analyze statistically

- Example:
  - Participants type using either iOS keyboard OR Android keyboard, but not both.
  - Most AB tests are between-subjects

Within-subjects design

Each participant (subject) experiences all levels of a factor

- much more powerful statistically
- but can introduce confounds
- Example:
  - Participants complete typing tasks using BOTH an iOS keyboard AND Android keyboard.

Carryover effects

The effect of one condition “carries over” into the next condition

Common in within-subject designs

e.g., learning from one condition to the next

Neutralize carryover effects with **counterbalancing**

Two conditions:
p1: iOS, Android
p2: Android, iOS
p3: iOS, Android
p4: Android, iOS
…

Three conditions:
p1: A, B, C
p2: A, C, B
p3: B, A, C
p4: B, C, A
p5: C, A, B
p6: C, B, A

fully counterbalanced

Mixed factorial design

Contains at least one between-subjects factor and one within-subjects factor.

Also called split-plot designs.

e.g. Do men and women perform differently when using each of two different mobile keyboards?

- Between subjects factor sex with two levels: man, woman
- Within subjects factor keyboard with two levels: iOS and Android

Confounds

Any unaccounted for factors that could explain your results. Serious confounds ruin experiments.

Examples:

- unequal treatments or procedures (e.g., participants typed 5 phrases with iOS and 20 with Android)
- sources of non-random variation (e.g., all participants who used iOS were teenage boys)
- systematic measurement error (e.g., task start time was different for Android than for iOS)
- various other biases (discussed in previous classes)

Avoiding confounds

Remove or exclude

- simply make the confound not exist

Spread equally

- randomize such that the confound is ‘noise’

Manipulate as a factor

- systematically control a confound’s influence

Record as a covariate

- we can then test whether it had an effect

Randomization

Randomization: the random assignment of treatments to the experimental units or participants

In a totally randomized experiment, no one, including the investigators themselves, is able to predict the condition to which a participant is going to be assigned

Practice?

When comparing a ‘new thing’ to an ‘old thing,’ how can we make a fair comparison?

How do we handle practice?

Example:

- Typing on a familiar QWERTY keyboard versus a new, unfamiliar experimental keyboard.

Handling practice

Recruit participants with equal (non-)familiarity with treatments

- could we find people who have never used a QWERTY keyboard?

Give fixed amount of practice

- can be fixed amount of time, or fixed # of trials

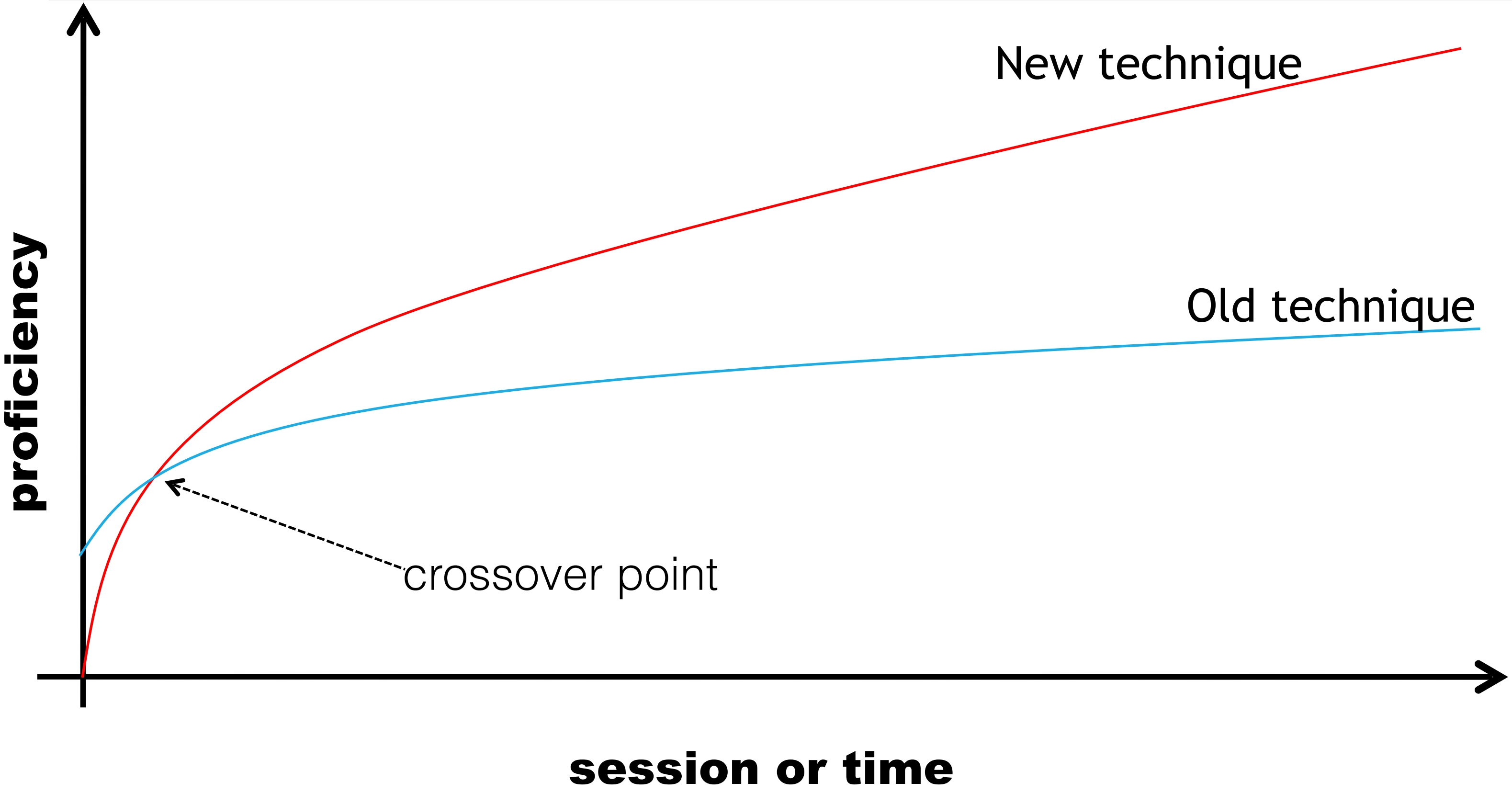
Practice until a certain proficiency is reached

- requires real-time feedback, go until performance is equal, report time taken to that point, then study further

Run a longitudinal study

- test over multiple sessions and construct learning curves

Learning curves



Example 1

HCI researchers wanted to determine if the size of a device’s screen affects how quickly people are able to read news articles. They created an experiment in which they asked 40 participants to read a news article on either a smartwatch, smartphone, tablet, or desktop. They measured how long it took each participant to read the article.

Factor(s) / Independent variable(s)?

- Within or between subjects?
- Levels?

Dependent variable(s)?

Possible issues/confounds to think about?

Example 2

Icons in user interfaces can be used for many purposes. But are icons always better than text, or a mix of icons and text, or just the text? Nicki tried to answer this question by performing a controlled experiment with three different interfaces 1. Icons only 2. Icons with command name and 3. Command names only. The experiment measured users preference for each of the three interfaces.

Factor(s) / Independent variable(s)?

- Within or between subjects?
- Levels?

Dependent variable(s)?

Possible issues/confounds to think about?

Example 3

Researchers wanted to study how the temperature of the room affected male and female students performance on their final exam. They split the class into two groups, with each group having approximately equal numbers of males and females. One group completed the test in a room at 60 degrees and the other at 90 degrees. Researchers measured their overall test score along with the time that it took to complete the test.

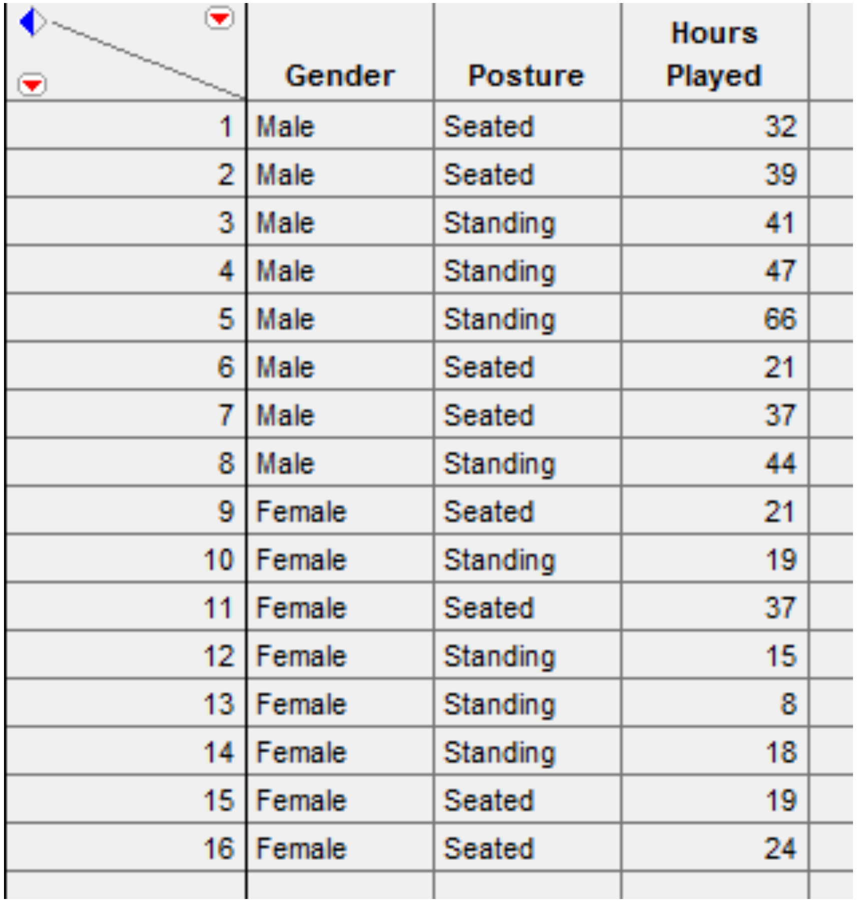
Factor(s) / Independent variable(s)?

- Within or between subjects?
- Levels?

Dependent variable(s)?

Possible issues/confounds to think about?

Significance tests



	Gender	Posture	Hours Played
1	Male	Seated	32
2	Male	Seated	39
3	Male	Standing	41
4	Male	Standing	47
5	Male	Standing	66
6	Male	Seated	21
7	Male	Seated	37
8	Male	Standing	44
9	Female	Seated	21
10	Female	Standing	19
11	Female	Seated	37
12	Female	Standing	15
13	Female	Standing	8
14	Female	Standing	18
15	Female	Seated	19
16	Female	Seated	24

Is there a significant difference in this data?

Significance tests

Why do we need significance tests?

- When the values of the members of the comparison groups are all known, you can directly compare them and draw a conclusion. No significance test is needed since there is no uncertainty involved.
- When the population is large, we can only sample a sub-group of people from the entire population.
- Significance tests allow us to determine how confident we are that the results observed from the sampling population can be generalized to the entire population.

Significance tests: p-values

How do we determine if something is statistically significant?

We perform an analysis of variance and get a p-value.

The p-value comes from the distribution of the sample mean.

The p-value is the probability of randomly getting a test statistic as (or more) extreme than what you observed if the null hypothesis was true.

i.e. the probability that your results occurred by chance

p = 0.45 means there is a 45% chance the data occurred by chance.

p = 0.05 means there is a 5% chance the data occurred by chance.

Significance tests

We now need to use the p-value to choose a course of action . . .

Either reject the null hypothesis, or fail to reject the null hypothesis

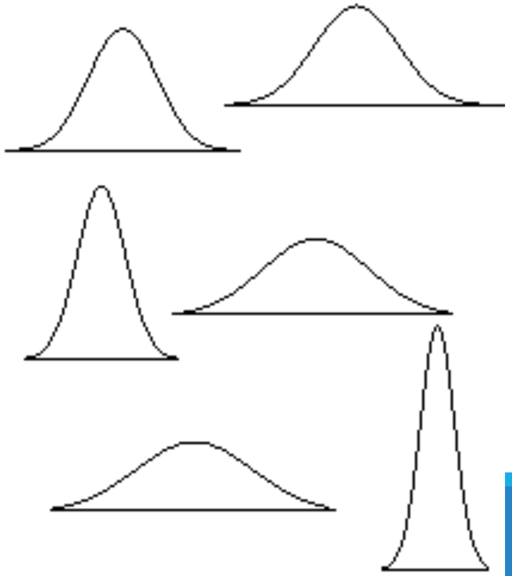
We need to decide if our sample result is unlikely enough to have occurred by chance.

Standard cutoff (in HCI) is $p < .05$ .i.e. we're at least 95% confident that our results did not occur by chance.

ANalysis Of VAriance (ANOVA)

- Statistical Workhorse
  - Supports moderately complex experimental designs and statistical analysis
  - Lets you examine differences between multiple independent variables at the same time
  - Assumes a normal distribution (Bell curve)

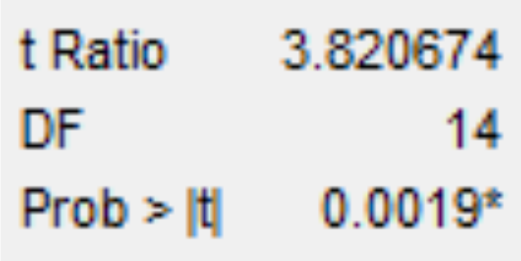
- ANOVA - analysis of variance
  - t-test (two levels)
  - F-test (more than two levels)



Reporting statistical significance

In HCI, p<0.05 is usually considered "significant"

e.g., t-test



t Ratio	3.820674
DF	14
Prob > |t|	0.0019*

"Gender had a significant effect on hours of game-play (*t*(14)=3.82, p<.01)."

Tests where p>.05 are "nonsignificant"

- Or "not detectably different"
  - (t(14)=1.23, n.s.)

They are not "insignificant."

Does *not* show equality!

Usually report p-values for only…

- p<.05
- p<.01
- p<.001
- p<.0001

Summary of Experimental Research

Experimental research allows the establishment of a **causal relationship**. Usually these are controlled experiments.

Experimental research requires well-defined, testable hypotheses that consist of a limited number of dependent and independent variables.

Use statistical tests to establish “significance” of results.

Experimental research requires strict control of factors that may influence the dependent variables.

- Strict control may not be a good representation of users’ typical interaction behavior.
- Experiments done “in the wild” can be difficult to control.

Activity (in pairs)

In your project, what controlled experiment could you run?

- What is your hypothesis? What is the null hypothesis?
- What is your “control” group, “treatment” group(s)?
- What would you have participants in each group do?
- What data would you collect?
  - *What is your independent variable(s) - Factor(s) and levels?*
  - *What is your dependent variable(s) - outcome/measure(s)?*
  - *Would you use a between-subjects or within-subjects design? Why?*
- What are possible confounds or issues to keep in mind?
- **Write your NetIDs and names on it and turn it in!**