CS 4249: Experimental Design

KAN Min-Yen

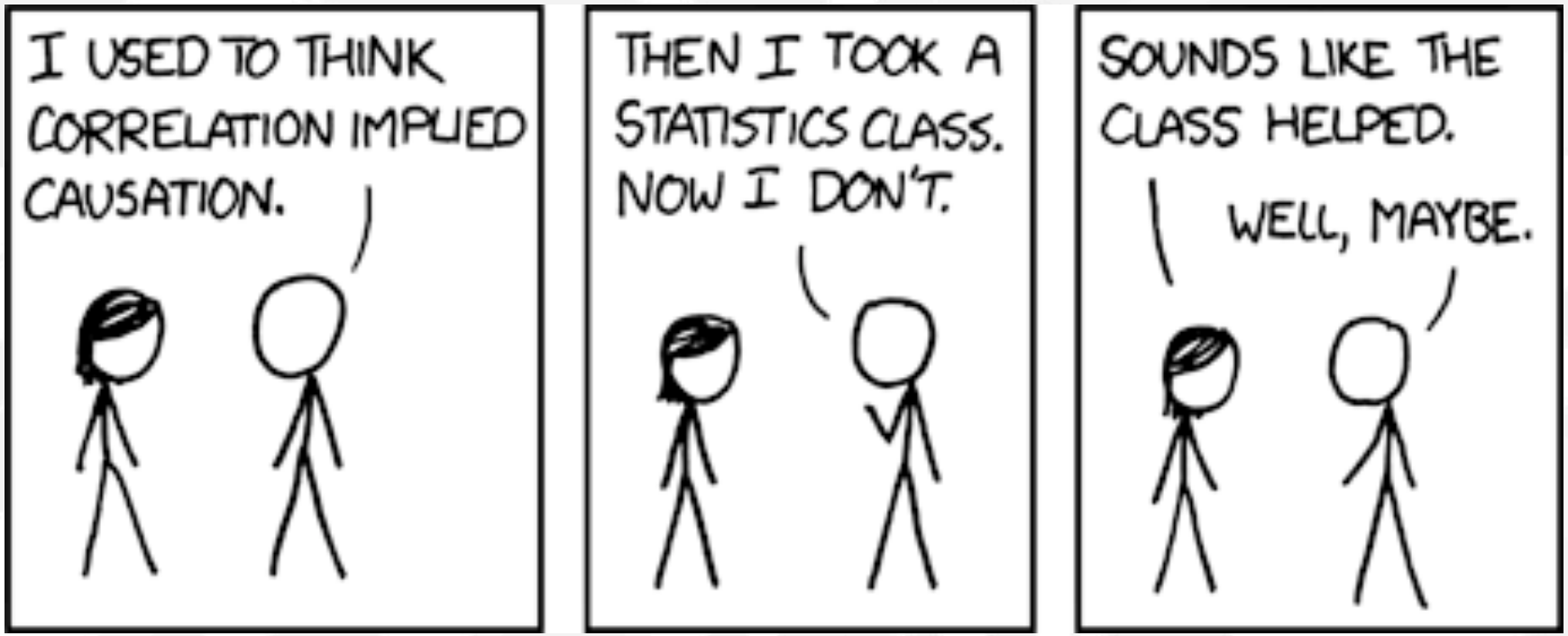
Week 5

(This week's lecture largely brought to you by Lazar, Feng and Hochheiser's *Research Methods in HCI*, Chapters 2, 3 and 4)

Behavioral Research

Research Type	Focus	Claims	Methods
Descriptive	Accurate description of the event	X is happening	Observation, Interviews, Focus Groups, Field Studies
Relational	Relation between factors, but not cause	X is related to Y	Observations, Surveys, Field Studies
Experimental	Uncovering cause	X causes Y	Controlled Experiments

Statistical Analysis



xkcd.com/552

Experimental Validation of Design

Design can be tested scientifically, when:

- A lucid and testable hypothesis exists
- Dependent variables can be controlled

- Our goal: create results that are replicable
  - Apply statistical significance testing

Research Hypothesis and Variables

- A statement that can be tested through experimentation
  - Needs a null (H_0) and alternative (H_1) hypotheses
  - H_0 usually is "no difference" between treatments
- Independent variable – the different factor
  MANIPULATIONS
- Dependent variable – outcome being studied
  MEASURES

Your turn: Quick question

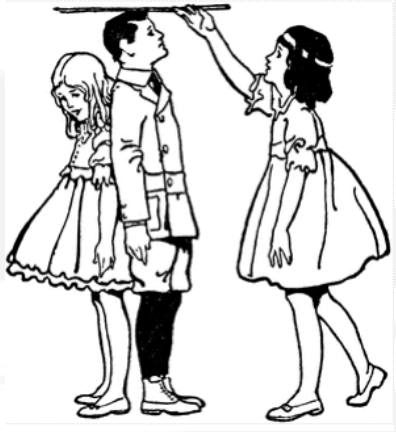
What are some of the typical independent variables studied in HCI research?

What about dependent variables?

Experimental Design

- Measuring dependent variable by varying the independent variable on the system
- Do this by random assignment of participants to values of the independent variable (treatments)
- Often will have multiple tasks, these often need randomization too (learning effects)

Why do we need significance tests?



Mike's height is 187cm. Mary's height is 172cm. So Mike is taller than Mary.

The average height of two males is 165cm. The average height of two females is 177cm. So females are taller than males.

We can't obtain data from all individuals from a population, so you need to create a (random, convenience [i.e.,] stratified[i.e.,]) sample. Significance testing allow us to determine our confidence that the results observed generalizes to the population.

Type I and II Errors

		Jury Decision	
		Not Guilty	Guilty
Reality	Not Guilty	(True Positive)	Type I Error (False Positive)
	Guilty	Type II Error (False Negative)	(True Negative)

Applies to US and SG courts: innocent (H_0) until proven guilty.

Quick Q: why might Type I errors be considered worse than Type II?

Significance tests measures the likelihood of falsely concluding Type I errors.

Experiment Design

- Factorial Design
  - When more than one independent variable is being observed.

A) An experiment to compare typing speed when using two types of keyboards and the tasks of composition and transcription.

B) An experiment to compare the satisfaction of subjects using an experimental UI, LINC and a database frontend to find books in the Central Library.

Your Turn: Which is which?

Within-Subject (Repeated Measures)

Between-Subject

Definitions

A) All participants try all treatments

B) Each participant tries one treatment

Disadvantages

I) Counterbalancing for ordering effects needed

II) Participant Fatigue

III) Requires large sample size

IV) Effect of individual differences hard to separate

Split-Plot

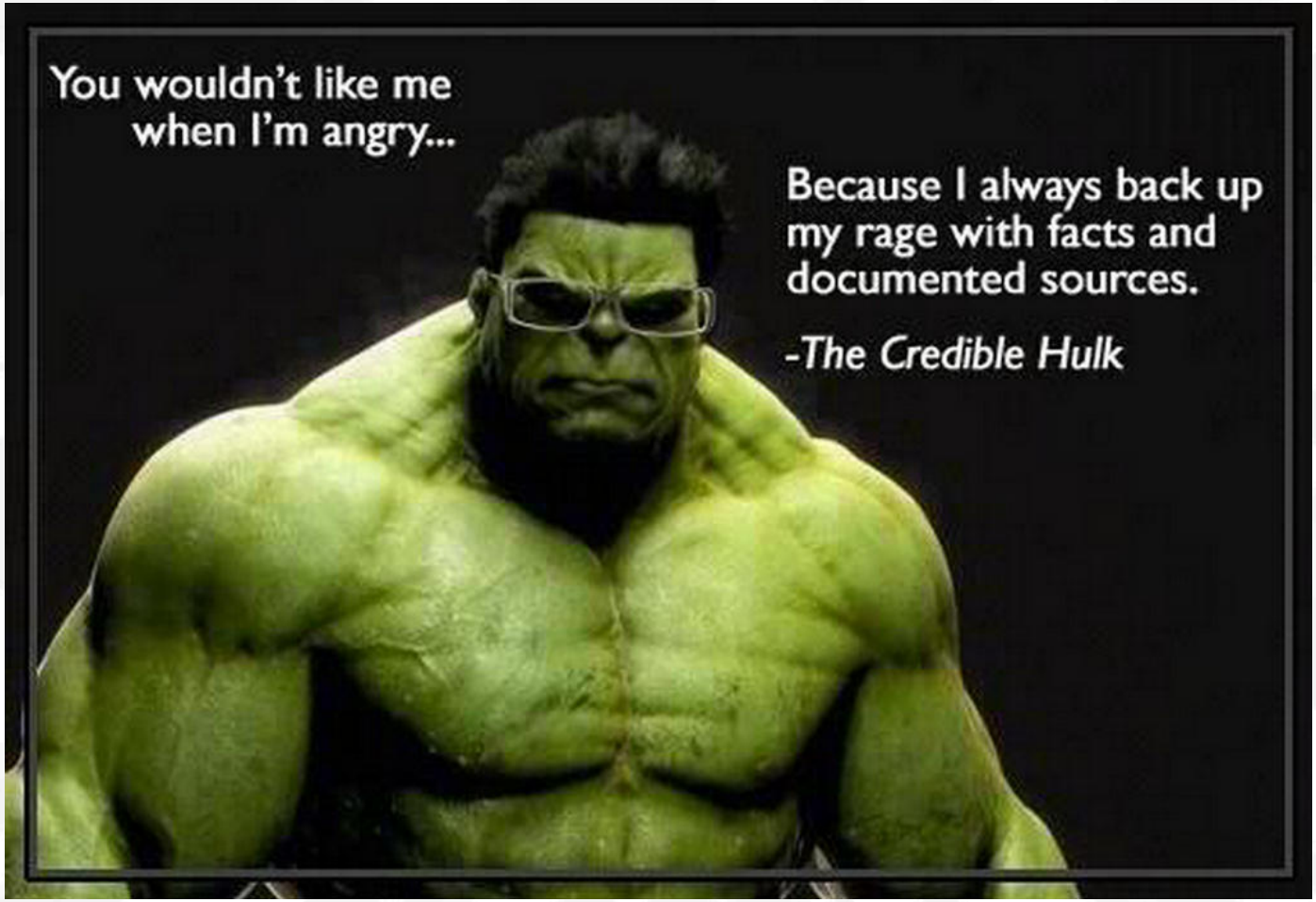
- Split-plot designs are also a possibility with factorial designs
  - A portion of variables are investigated through between-subject design
  - The rest, within-subject design.

		QWERTY	DVORAK
Composition		A	C
Transcription		B	D

Data Coding

- When coding information, compile a guideline to follow.
- For consistency, documentation and to help multiple annotators. E.g.:
  - Age: _____ (9, 9.5, nine, nine yrs. old)
  - Code mouse clicks that open to a new screen, but not any pop-up modal dialogs.
- Good to show annotator consistency : Cronbach's α

Statistical Analyses

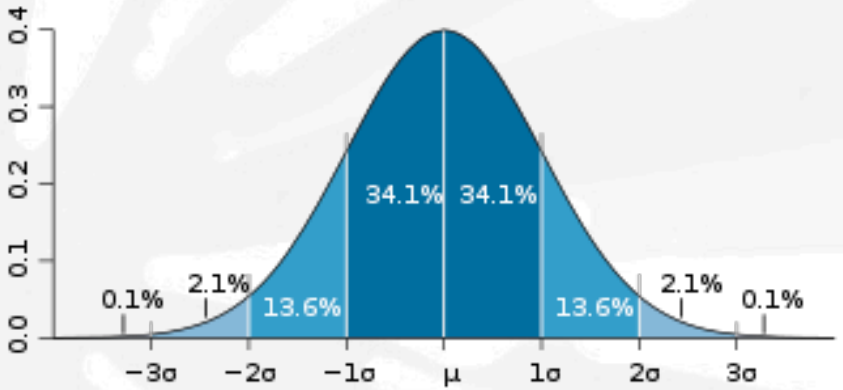


Hypothesis testing for simple designs

Experiment Design	Indep. Vars. (IV)	Conditions for each IV	Test Type	Degrees of Freedom
Between-subject	1	2	Independent-sample t test	$n_1 + n_2 - 2$ (2 groups of 10 subjects, df = 18)
Within-subject	1	2	Paired-samples t test	n-1 (10 subjects, df = 9)

Normal Distribution

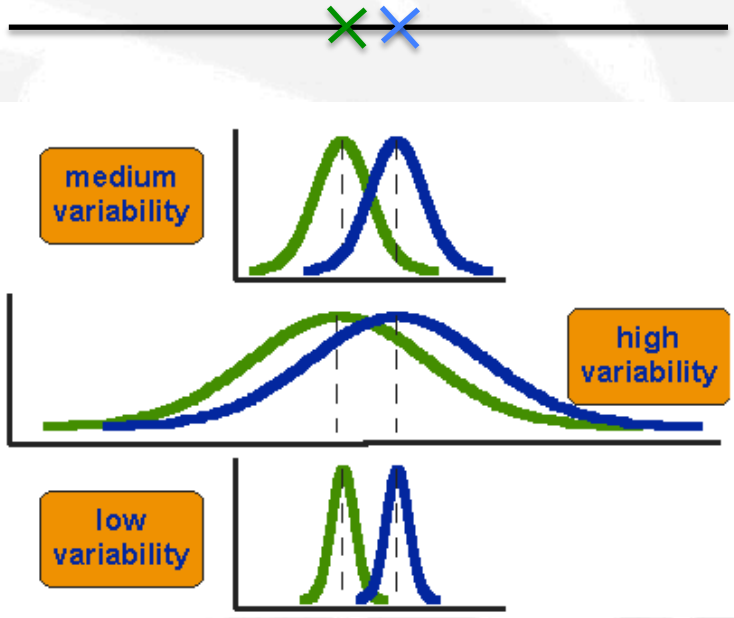
- A.K.A. Gaussian distribution
- Very convenient; closed form with μ and σ^2.
  MEAN VARIANCE
- The Central Limit Theorem roughly states that the variances of many variables will also be a normal distribution



Credits: Wikipedia

Q Question: Does "Six Sigma" have anything to do with this sigma (σ)?

Do the means overlap?



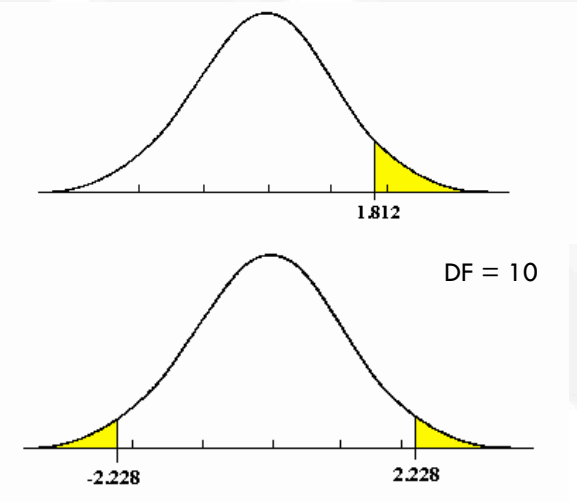
- We need to know not just the mean, but the variance. Is that sufficient?
- No, it's still not enough. We need to also know the sample size.

William S. Gosset

Student's *t*-Test

- Two groups are unrelated:
  - Independent Sample *t* Test (e.g., Between-subject)
- Two groups contribute two data sets:
  - Paired Sample *t* Test (e.g., Within-subject)

- One tailed
  - If a direction is specified
    e.g., "improvement"
- Two tailed
  - If only difference is needed



Interpreting the *t*-Test

- T-tests return a *t* value, where larger *t* suggests a higher probability that H_0 is false.
- Given a *t* and the degrees of freedom, we can look up the probability that the means are not the same.
  "SAMPLE SIZE"
- If that probability is above 95% (99%) we can conclude significance.

An independent-samples t-test suggests that there is a significant difference in the task completion time between the group who used the standard word-processing software and the group who used the prediction software ($t(15) = 2.169$, $p < 0.05$).

Analysis of Variance

- When there are more than two means, we use ANOVA, which return a value f (instead of t)
  - Repeated t-tests would give the wrong conclusion
- F values can then be checked to see whether they indicate a probability (e.g., $p > .95$) to reject the null hypothesis; again, given specific degrees of freedom.

ANOVA Designs

Experiment Design	Indep. Vars. (IV)	Conditions for each IV	Test Type
Between-subject	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-subject	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Split-Plot	2 or more	2 or more	Split-Plot ANOVA

The idea behind ANOVA is to compare the ratio of **between group** variance to **within group** variance. If the variance caused by the interaction between the samples is much larger when compared to the variance that appears within each group, then it is because the means are not the same.

Sample ANOVA statement

The analysis result suggests that there is no significant difference between between subjects who completed the transcription tasks and those who complete the composition tasks ($F(1,42) = 1.41$, n. s.). There is significant difference among participants who used different input methods ($F(2,42) = 4.51$, $p < 0.05$)

	Standard	Prediction	Speech
Transcription	Group A: 8 subjects	Group B: 8 subjects	Group C: 8 subjects
Composition	Group D: 8 subjects	Group E: 8 subjects	Group F: 8 subjects

DF = 48 subjects – 6 groups = 42

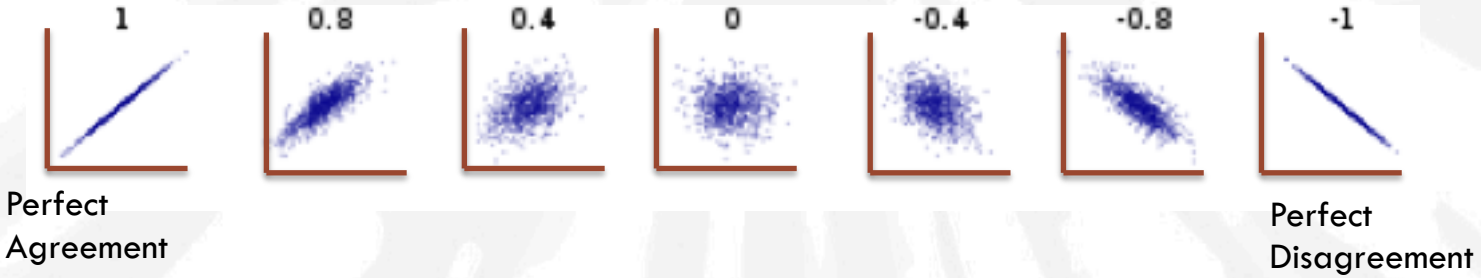
Reliability of Experimental Results

- Random Errors ("Noise")
- Systematic Errors ("Biases")
  1. Measurement instruments: Low battery mechanical stopwatch
  2. Experimental procedure: Learning and Fatigue
  3. Participants: Recruiting
  4. Experimenter behavior: Leading, Impatience
  5. Environment: Workplace vs. Laboratory

Identifying correlations

- Test whether variables x and y are related.
  - E.g., want to know whether computer experience correlates with typing speed.

- Pearson's R returns a value between {-1,+1}



Intervening variables

- In correlation, watch out for relations that have an intervening variable that may explain the factors.

- Observed: Subjects with high income can find things slower than those with low income on a web site. What should we conclude?

Summary

- Experimental Design
  - Within, Between Subjects & Split-Plot
- Significance Testing
  - T-tests and ANOVA (F-tests*)
- Correlation
  - Pearson's product moment (R)