Controlled Experiments

September 12, 2018

Overview

Hypothesis testing

Experimental design

Intro to basic data analysis

Controlled Experiments

Test the effect of manipulating one or more *independent variables* on one or more *dependent variables*

Experimental Process

- Formulate a hypothesis
- Identify independent, dependent variables
- Design a controlled experiment
- Check for:
  - Confounds
  - Validity
  - Reliability
- Select representative participants
- Randomly assign to conditions
- Run experiment, collect data
- Analyze results

Hypothesis

A suggested explanation of a phenomenon
- "If I change A, then B will change in this manner…"

In experimentation, want hypothesis to be as specific as possible
- Makes it easier to test

Hypothesis

To test hypothesis, must identify what variables we think will lead to expected outcome
- "Users will complete tasks faster with keyboard shortcuts than without them"
- "Users will be able to select items faster with pie menus than with vertical context menus"

Clearly identify which variables will influence what outcomes, and how

Only manipulating independent variables increases our confidence that any observed changes in dependent variables due to changes in independent variables

Hypothesis Testing

In testing hypothesis, we are seeking to reject the *null hypothesis*

Null hypothesis
- There exists no relationship between manipulating the independent variables and the resultant changes in the dependent variables
- Example:
  - "There is no difference in selection speed between pie-menus and vertical context menus"

Experimental Design

Participant pool
- Are the study participants representative of the intended user population?
- E.g.,
  - College students vs. elder adults for a study on assistive technology

How will participants be assigned to conditions?
- Two options:
  - Between-subjects
  - Within-subjects

Between-Subjects

Each participant does one of the experimental conditions

Doesn't account for individual variability

Need more participants

No learning effects (good)

Also known as "randomized experiments"

Within-Subjects

Each participant completes all experimental conditions

Better able to account for individual differences

Requires fewer participants

Allows participants to make direct comparative statements

Learning effects are possible

- To account for these, order of conditions are usually *counterbalanced*

Designing Study Tasks

Tasks must:

- be externally valid
- exercise the key aspects of any new technology, theory, etc
- be feasible

Task Design

Often the toughest part of experiment design

Open-ended tasks:

- More natural, but harder to control

Restricted tasks:

- Less variability
- Greater internal validity

Examples?

Statistical Analysis

Calculations that tell us

- mathematical attributes about our data sets
  - mean, amount of variance, ...
- how data sets relate to each other
  - whether we are “sampling” from the same or different distributions
- the probability that our claims are correct
  - “statistical significance”

Statistical vs Practical Significance

When n is large, even a trivial difference may show up as a statistically significant result

- E.g. menu choice:
- mean selection time of menu a is 3.00 seconds
  menu b is 3.05 seconds

Statistical significance does not imply that the difference is important!

- Whether or not the difference matters is open to interpretation

T-test

A simple statistical test

- Allows one to say something about differences between means at a certain confidence level

Null hypothesis of the T-test:

- No difference exists between the means of two sets of collected data

Possible results:

- I am 95% sure that null hypothesis is rejected
  - (there is probably a true difference between the means)
- I cannot reject the null hypothesis
  - the means are likely the same

Different Types of T-tests

Comparing two sets of independent observations

- usually different subjects in each group (between-subjects)
- number per group may differ as well

Condition 1	Condition 2
S1–S20	S21–43

Paired observations

- usually a single group studied under both experimental conditions (within-subjects)
- data points of one subject are treated as a pair

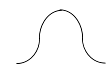
Condition 1	Condition 2
S1–S20	S1–S20

Different Types of T-tests

- Non-directional vs directional alternatives
  - Non-directional (two-tailed)
    - no expectation that the direction of difference matters
  - Directional (one-tailed)
    - Only interested if the mean of a given condition is greater than the other

T-test Assumptions

- Data points of each sample are normally distributed
  - but t-test very robust in practice



- Population variances are equal
  - t-test reasonably robust for differing variances
  - deserves consideration

- Individual observations of data points in sample are independent
  - must be adhered to

T-test…

- Significance level
  - Decide upon the level before you do the test!
  - In HCI, typically stated at the .05 (sometimes 0.1)
  - Some consider < 0.1 a “trend”
    - But this is controversial

Two-tailed unpaired T-test

- N: number of data points in the one sample
- $\sum X$: sum of all data points in one sample
- X: mean of data points in sample
- $\sum(X^2)$: sum of squares of data points in sample
- s^2: unbiased estimate of population variation
- t: t ratio
- df = degrees of freedom = N1 + N2 – 2

Formulas

$$s^2 = \frac{\sum(X_1^2) - \frac{(\sum X_1)^2}{N_1} + \sum(X_2^2) - \frac{(\sum X_2)^2}{N_2}}{N_1 + N_2 - 2} \qquad t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}}$$

How to maximize t?

Degrees of Freedom

Freedom of a set of values to vary independently of one another:

X = {21, 20, 24} N = 3

$\bar{X}$ = 65/3 = 21.6 <= $\bar{X}$ has N – 1 =2 df

Once you know the mean of N values, only N-1 can vary independently

Level of Significance for Two-Tailed Test

df	.05	.01
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947

df	.05	.01
16	2.120	2.921
18	2.101	2.878
20	2.086	2.845
22	2.074	2.819
24	2.064	2.797

Critical value: threshold that t statistic much reach to achieve significance.

How does the critical value change based on the degrees of freedom and the confidence level?

Analysis of Variance (ANOVA)

A statistical workhorse

- Supports moderately complex experiment designs (relative to t-test)
- Lets you examine multiple independent variables at the same time

Analysis of Variance (ANOVA)

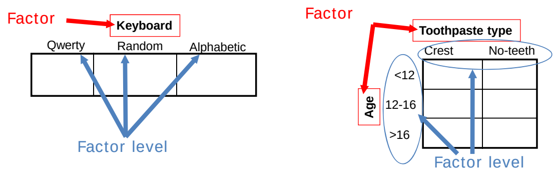
Examples

- There is no difference between people's mouse typing ability on the Random, Alphabetic and Qwerty keyboard
- There is no difference in the number of cavities of people aged under 12, between 12-16, and older than 16 when using Crest vs No-teeth toothpaste

Analysis of Variance (ANOVA)

Terminology

- Factor = independent variable
- Factor level = specific value of independent variable



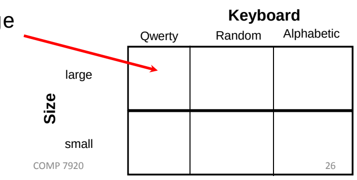
ANOVA Terminology

Factorial design

- cross combination of levels of one factor with levels of another
- Eg. keyboard type (3) x size (2)

Cell

- unique treatment combination
- E.g. qwerty x large



ANOVA Terminology

Mixed factor

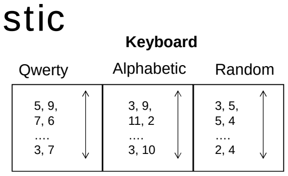
- contains both between and within subject combinations
- within subjects: keyboard type
- between subjects: size

Size	Keyboard: Qwerty	Random	Alphabetic
Large	*S1-20*	*S1-20*	*S1-20*
Small	*S21-40*	*S21-40*	*S21-40*

f Statistic

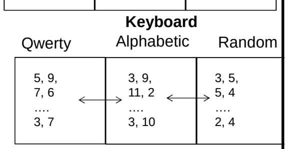
Within group variability (WG)

- Individual differences
- Measurement error



Between group variability (BG)

- treatment effects
- individual differences
- measurement error



these two variabilities combine to give total variability

we are mostly interested in between group variability because we are trying to understand the effect of the treatment

f Statistic

f = $\frac{BG}{WG} = \frac{treatment + ID + m.error}{ID + m.error}$ = ?

= 1, if there are no treatment effects

> 1, if there are treatment effects

Within-subjects design:

the ID component in numerator and denominator "cancels" out, therefore a more powerful design

f Statistic

Similar to the t-test, we look up the f value in a table, for a given alpha and degrees of freedom to determine significance

Thus, f statistic sensitive to sample size
- Large sample => Easier to find significance
- Small sample => Difficult to find significance

What we (should) want to know is the effect size
- does the treatment make a big difference (i.e., large effect)?
- or does it only make a small difference (i.e., small effect)?
- depending on what we are doing, small effects may be important findings

Data Analysis: Terminology

Main effect
- There is some difference between levels of a factor
  - But, doesn't tell you where the difference lies (if you have > 2 levels)

Post-hoc analysis
- Where does the difference lie?
  - E.g., pairwise comparisons
- Corrections (e.g., Bonferroni) used to protect against Type I error

ANOVA

Compares relationships between many factors

- Considers the interactions between factors

ANOVA Interactions

Example interaction

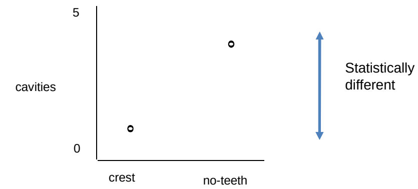
- typists are faster on Qwerty than the other keyboards
- non-typists perform the same across all keyboards
- cannot simply say that one keyboard is best without talking about typing ability

	Qwerty	Random	Alpha
non-typist	S1-S10	S11-S20	S21-S30
typist	S31-S40	S41-S50	S51-S60

ANOVA - Interactions

Example:

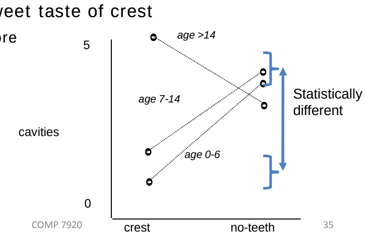
- t-test: crest vs no-teeth
  - subjects who use crest have fewer cavities
- interpretation: recommend crest



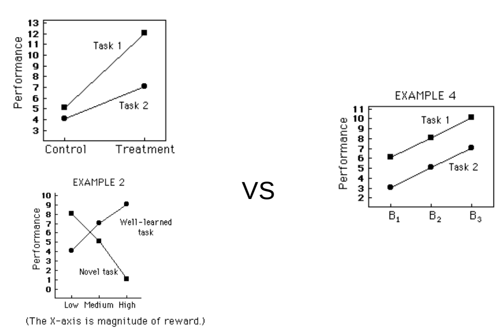
ANOVA - Interactions

Example:

- anova: toothpaste x age
  - subjects 14 or less have fewer cavities with crest.
  - subjects older than 14 have fewer cavities with no-teeth.
- interpretation: the sweet taste of crest
  - makes kids use it more
  - repels older folks



Example Interactions



Types of Validity

Construct validity
- Are you measuring what you say you are measuring

Internal validity
- The changes in the dependent variables are caused by the independent variables

External validity
- Results can be generalized to other settings, populations, tasks, etc.

Ecological validity
- To what extent do the study conditions mimic those in the real world
- Related to external validity, but not the same

Learning Outcomes

Now you…
- Are familiar with basic experimental design
- Can explain the difference between-subjects and within-subject designs
- Know that there are a number of different statistical methods that can be applied to different experimental designs
- Are familiar with two statistical tests (T-tests and ANOVA)
- Are familiar with ANOVA terminology (e.g., factors, levels, cell, factorial design)
- Can explain the difference between statistical and practical significance