CPSC 544: Experiments II

Prof. Narges Mahyar

LEARNING GOALS

- why are statistics used?
- What is a T-test?
- what is an analysis of variance (ANOVA)?
- what is the important terminology in ANOVA?
- what are the different types of ANOVA?
- when would one choose to use an ANOVA?
- what other statistics are relevant to HCI?

Acknowledgement: Some of the material in this lecture is based on material prepared for similar courses by Saul Greenberg (University of Calgary)

STATISTICAL ANALYSIS

- what is a statistic?
  - a number that describes a sample
  - sample is a subset (hopefully representative) of the population we are interested in understanding
- statistics are calculations that tell us
  - mathematical attributes about our data sets (sample)
    - mean, amount of variance, ...
  - how data sets relate to each other
    - whether we are "sampling" from the same or different populations
  - the probability that our claims are correct
    - "statistical significance"

T-TEST

allows one to say something about differences between two means at a certain confidence level

null hypothesis of the t-test:

- no difference exists between the means

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 (number may differ as well)

- Condition 1 Condition 2
- S1–S20 S21–S43

paired observations
usually single group studied under separate experimental conditions

data points of one subject are treated as a pair

- Condition 1 Condition 2
- S1–S20 S1–S20

Which one is within-subject? Between-subject?

DIFFERENT TYPES OF T-TESTS

comparing two sets of independent observations **(between subjects)**

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

- Condition 1 Condition 2
- S1–S20 S21–S43

paired observations **(within subjects)**
usually single group studied under separate experimental conditions

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

TWO-TAILED UNPAIRED T-TEST

n: number of data points in the one sample ($N = n_1 + n_2$)

ΣX: sum of all data points in one sample

$\overline{X}$: mean of data points in sample

$\Sigma(X^2)$: sum of squares of data points in sample

s^2: combined sample variance

t: t ratio

df = degrees of freedom = $n_1 + n_2 - 2$

How to maximize t?

Formulas

$$s^2 = \frac{\Sigma(X_1^2) - \frac{(\Sigma X_1)^2}{n_1} + \Sigma(X_2^2) - \frac{(\Sigma X_2)^2}{n_2}}{n_1 + n_2 - 2}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

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 must reach to achieve significance.

How does critical value change based on *df* and confidence level?

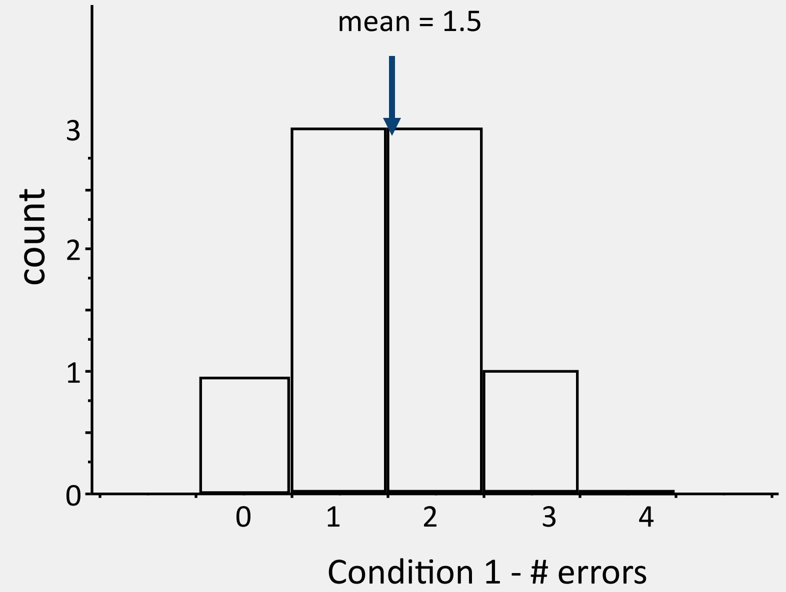
BACK TO EXAMPLE:

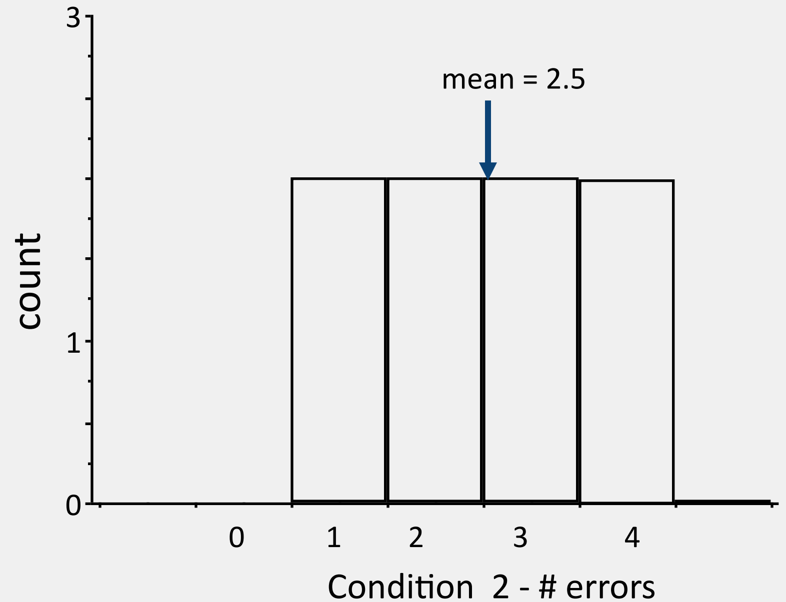
scenario 2: assume we ran a between-subjects experiment, where we counted the # of errors under each condition

condition 1 (pop-up) : 0, 1, 1, 1, 2, 2, 2, 3

condition 2(pull down) : 1, 1, 2, 2, 3, 3, 4, 4

Is there ***a significant*** difference between the means?





TWO-TAILED UNPAIRED T-TEST

Condition one (pop up): 0, 1, 1, 1, 2, 2, 2, 3

Condition two (pull down): 1, 1, 2, 2, 3, 3, 4, 4

What the results would look like in R.

data: my_data$Condition.1 and my_data$Condition.2

t = -1.8708, df = 13.176, p-value = 0.08374

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
 -2.1531955 0.1531955

sample estimates:
mean of x mean of y
 1.5 2.5

is the difference **significant**?

TWO-TAILED UNPAIRED T-TEST

Condition one (pop up): 0, 1, 1, 1, 2, 2, 2, 3

Condition two (pull down): 1, 1, 2, 2, 3, 3, 4, 4

data: my_data$Condition.1 and my_data$Condition.2

t = -1.8708, df = 13.176, p-value = 0.08374

hint

probability that means are from the same underlying population

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-2.1531955 0.1531955

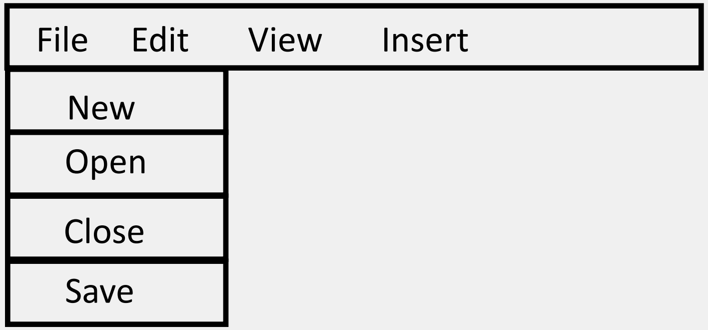
sample estimates:
mean of x mean of y
1.5 2.5

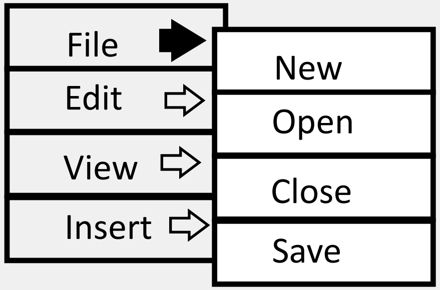
How does the outcome change for a confidence level of 0.10?

RECALL MENU HYPOTHESES

This time lets just hypothesize about error rate:

- H0: **there is no difference in error rate** when selecting a single item from a pop-up or a pull down menu *- cannot reject at 0.5 level*
- H1: **selecting from a pop-up menu will be less error prone** than selecting from a pull down menu





SUMMARY OF THE T-TEST

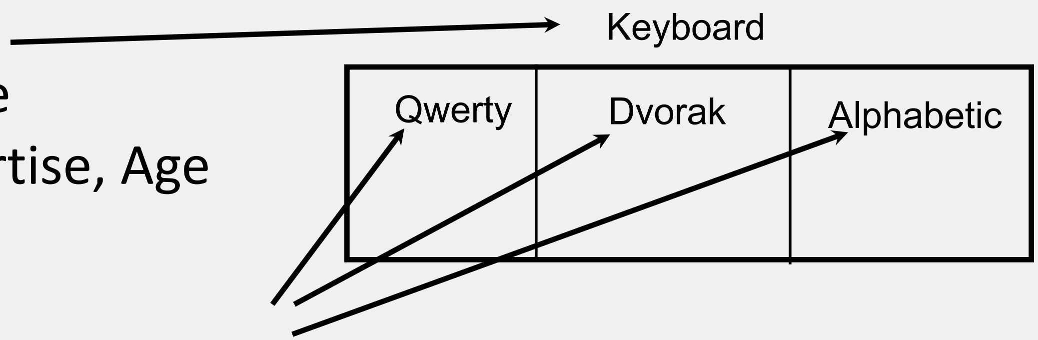
- the point: establish a confidence level in the difference we've found between 2 sample means.
- the process (what your stats software does under the hood):
  - compute df
  - choose desired significance, p (aka α)
  - calculate value of the t statistic
  - compare it to the critical value of t given p, df: t(p,df)

  - if t > t(p,df), can reject null hypothesis at p

ANALYSIS OF VARIANCE (ANOVA)

- a workhorse
  - allows moderately complex experimental designs (relative to t-test)

- terminology
  - factor
    - independent variable
    - e.g., Keyboard, Expertise, Age
  - factor level
    - specific value of independent variable
    - e.g., Qwerty, novice, 10-12 year olds



ANOVA TERMINOLOGY

between subjects

- a subject is assigned to only one factor level of treatment
- problem: greater variability, requires more subjects

Keyboard

Qwerty	Dvorak	Alphabetic
S1-20	*S21-40*	*S41-60*

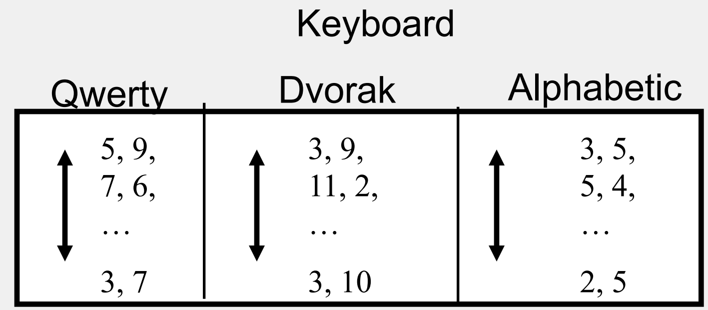
within subjects

- subjects assigned to all factor levels of a treatment
- requires fewer subjects
- less variability as subject measures are paired
- problem: order effects (e.g., learning)
- partially solved by counter-balanced ordering

Keyboard

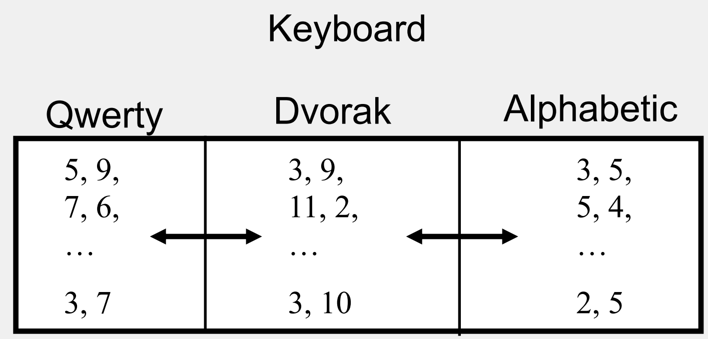
Qwerty	Dvorak	Alphabetic
S1-20	*S1-20*	*S1-20*

F STATISTIC



within group variability (WG)

- individual differences
- error (random + systematic)



between group variability (BG)

- treatment effects
- individual differences
- error (random + systematic)

these two variability's combine to give total variability

- we are mostly interested in ______________ variability because we are trying to understand the effect of the treatment

F STATISTIC

ANOVA is what we call an omnibus test

- tells us if ($\bar{x}_1 = \bar{x}_2 = \bar{x}_3$) IS NOT true
- doesn't tell us HOW the means differ (i.e. $\bar{x}_1 > \bar{x}_2$)

Intuition…

$$f = \frac{BG}{WG} = \frac{\text{treatment} + \text{id} + \text{error}}{\text{id} + \text{error}} = ?$$

= 1, if there are no treatment effects

> 1, if there are treatment effects

within-subjects design: the id component in numerator and denominator factored 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 α and degrees of freedom to determine significance

- thus, f statistic is sensitive to sample size
  - Big N → Big Power → Easier to find significance
  - Small N → Small Power → 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

STATISTICAL SIGNIFICANCE VS. PRACTICAL SIGNIFICANCE

- when N is large, even a trivial difference (small effect) may be large enough to produce a statistically significant result
  - e.g., menu choice:
    mean selection time of menu A is 3 seconds;
    menu B is 3.05 seconds

- statistical significance does not imply that the difference is important!
  - a matter of interpretation, i.e., subjective opinion
  - should always report means to help others make their opinion

- there are measures for effect size
  - regrettably they are not widely used in HCI research

SINGLE FACTOR ANALYSIS OF VARIANCE

- compare means between two or more factor levels within a single factor
- e.g.:
  - dependent variable: typing speed (time)
  - independent variable (factor): keyboard
  - between subject design

also called a one-way ANOVA

Qwerty	Alphabetic	Dvorak
S1: 25 secs S2: 29 … S20: 33	S21: 40 secs S22: 55 … S40: 33	S51: 17 secs S52: 45 … S60: 23

ANOVA TERMINOLOGY

- factorial design
  - cross combination of levels of one factor with levels of another
  - e.g., keyboard type (3) x expertise (2)
- Cell [or condition]
  - unique treatment combination
  - e.g., qwerty x non-typist

2-way factorial ANOVA

		Keyboard		
		Qwerty	Dvorak	Alphabetic
expertise	non-typist			
	typist			

ANOVA TERMINOLOGY

- mixed factor [split-plot]
  - contains both between and within subject combinations

		Keyboard	
	Qwerty	Dvorak	Alphabetic
expertise non-typist	*S1-20*	*S1-20*	*S1-20*
expertise typist	*S21-40*	*S21-40*	*S21-40*

ANOVA

- compares the relationships between many factors
- provides more informed results
  - considers the interactions between factors
  - e.g.,
    - typists type faster on Dvorak, than on alphabetic and Qwerty
    - non-typists are fastest on alphabetic

		Keyboard		
		Qwerty	Dvorak	Alphabetic
expertise	non-typist	*S1-20*	*S1-20*	*S1-20*
	typist	*S21-40*	*S21-40*	*S21-40*

Other statistical tests commonly used in HCI

- Your reading does a very good job of covering these, and we won't cover them further
  - Correlation
  - Regression
  - Non-parametric tests
    - Chi-squared
    - Mann-Whitney
    - Wilcoxon signed-rank
    - Kruskal-Wallis
    - Friedman's