

CS 444: advanced methods for
human-computer interaction

controlled experiments - II

class 06

administrivia

project

- MSII was just due!
- design reviews on Friday
- briefly introduce MSIII

Updated: NSERC USRA

- http://www.nserc-crsng.gc.ca/Students-Etudiants/UG-PC/USRA-BRPC_eng.asp
- if interested contact me

2

MSIII: Workplan Recommendation

To allow yourself adequate time for all Parts A, B & C, we suggest the following workplan. You have approximately 3.5 weeks.

By the end of the 1st week: Complete Part A (Steps 1 & 2).

By the end of the 2nd week: Complete Step 3, and have Step 4 underway. Complete Step 5, and have Step 6 underway.

Decide which team members will be working on refining the experiment and which will be focused on prototype implementation.

By halfway through the 3rd week: Close to completing Step 4, and have significant progress on Step 6.

Use the lab to get feedback on the experiment design and the current version of the prototype. There should be a clear plan about which team members will be completing which steps in the deliverable. There is less than a week left in this stage.

3

today: part I

learning goals:

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 is the difference between statistical and practical significance?

other tests: what are correlation & regression?

4

analysis of variance (ANOVA)

a workhorse

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

terminology

- factor**
 - independent variable
 - i.e., Keyboard, Toothpaste, Age
- factor level**
 - specific value of independent variable
 - i.e., Qwerty, Crest, 5-10 years old

5

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

6

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

7

f statistic

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

= 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

8

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 sensitive to sample size

- Big N \rightarrow Big Power \rightarrow Easier to find significance
- Small N \rightarrow Small Power \rightarrow 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

9

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

10

single factor analysis of variance

compare means between two or more factor levels within a single factor

e.g.:

- dependent variable: typing speed
- independent variable (factor): keyboard
- between subject design

also called a one-way ANOVA

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

11

ANOVA terminology

- factorial design
 - cross combination of levels of one factor with levels of another
 - e.g., keyboard type (3) x expertise (2)

2-way factorial ANOVA

- cell
 - unique treatment combination
 - e.g., qwerty x non-typist

		Keyboard		
		Qwerty	Dvorak	Alphabetic
expertise	non-typist			
	typist			

12

ANOVA terminology

mixed factor

- contains both between and within subject combinations

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

13

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

14

ANOVA

in reality, we can rarely look at one variable at a time
example:

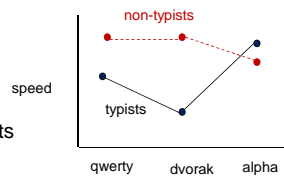
- t-test:

subjects faster on dvorak than qwerty



- anova: keyboard x expertise

alphabetic fastest for non-typists
dvorak fastest for typists



15

ANOVA case study

WIMP (GUI) vs. HYBRID (graphical command line)

motivation:

- WIMP interfaces are slow because of the mouse
- can we create a hybrid interface that is graphical but can be fully operated through the keyboard? (sort of like a command line)
- assume that one has been designed
- how should it be evaluated?

16

ANOVA case study

WIMP (GUI) vs. HYBRID (graphical command line)

independent variables:

- interface: WIMP, hybrid
- expertise: novice, expert
- command parameters: zero, one, two
 - E.g., bold (zero), font ariel (one), print -copies 2 -color greyscale (two)
 - **Note:** zero parameter commands can be done using shortcuts keys

dependent variables:

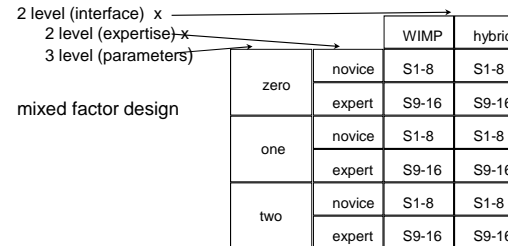
- performance: speed, error
- satisfaction

17

ANOVA case study

possible hypotheses:

- H1: experts will perform better than novices (not that interesting)
- H2: novices will perform better with WIMP than hybrid
- H3: experts will perform better with hybrid than WIMP, but only for commands with one or more parameters



18

task

assume that the task is to enter a whole series of commands, one after the other

there is an equal number of 0, 1, and 2 parameter commands used

the identical commands are used in both interface conditions

19

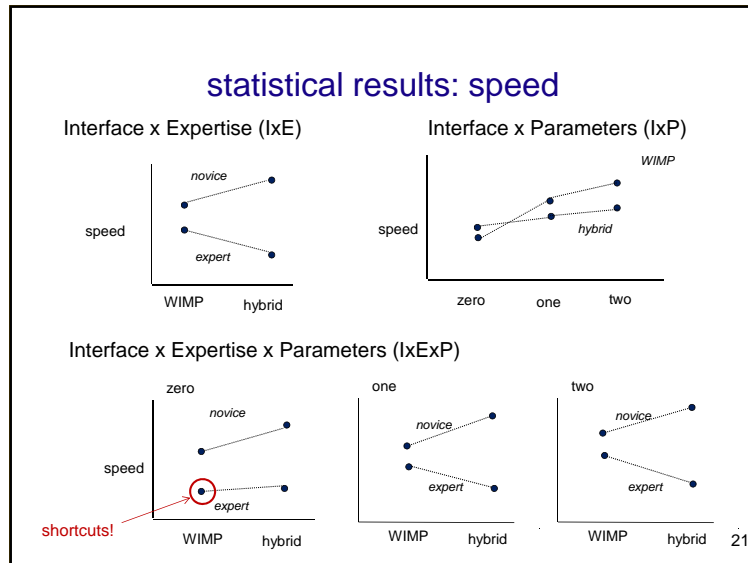
statistical results: speed

	<i>F-ratio.</i>	<i>p</i>	
Interface (I)	0.4		} main effects
Expertise (E)	5.5*	<0.05	
Parameters (P)	31.0**	<0.01	
IxE	15.2*	<0.05	} interactions
IxP	8.0*	<0.05	
ExP	5.0		
IxE _x P	14.1*	<0.05	

main effect: the effect of the variable **collapsing across** all levels of other variables in the experiment

interaction effect: the effect of one variable differs depending on the level of another (other) variable(s)

20



summary of results

Assuming same results for errors as speed...

H1: experts will perform better than novices (not that interesting)
Supported: main effect of expertise, showing experts better

H2: novices will perform better with WIMP than hybrid
Supported: 2-way interaction effect of interface and expertise, showing novices overall better with WIMP

H3: experts will perform better with hybrid than WIMP, but only for commands with one or more parameters
Supported: 3-way interaction effect of interface, expertise, and number of parameters, showing experts better with hybrid, but only with one and two parameters

22

case study conclusions

- expertise makes a big difference
- WIMP interaction should be kept for novices
- hybrid interaction should be available for experts

23

other tests: correlation

measures the extent to which two concepts are related

- e.g., years of university training vs computer ownership per capita

how?

- obtain the two sets of measurements
- calculate correlation coefficient
 - +1: positively correlated
 - 0: no correlation (no relation)
 - -1: negatively correlated

dangers

- attributing causality
 - a correlation does not imply cause and effect
 - cause may be due to a third "hidden" variable related to both other variables
 - e.g., (above example) age, affluence
- drawing strong conclusion from small numbers
 - unreliable with small groups
 - be wary of accepting anything more than the direction of correlation unless you have at least 40 subjects

24

non-HCI sample study: cigarette consumption

crude Male death rate for lung cancer in 1950 per capita consumption of cigarettes in 1930 in various countries

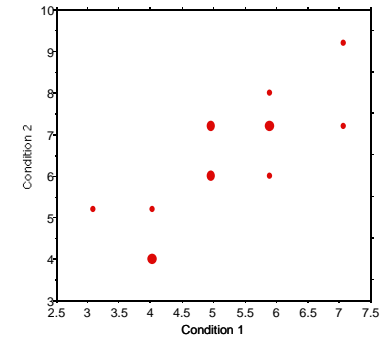


25

correlation

$r^2 = .668$

condition 1	condition 2
5	6
4	5
6	7
4	4
5	6
3	5
5	7
4	4
5	7
6	7
6	6
7	7
6	8
7	9



26

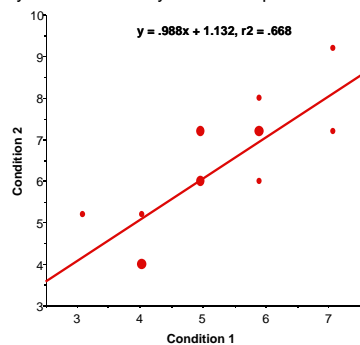
regression

calculate a line of "best fit"

use the value of one variable to predict the value of the other

- e.g., 60% of people with 3 years of university own a computer

condition 1	condition 2
5	6
4	5
6	7
4	4
5	6
3	5
5	7
4	4
5	7
6	7
6	6
7	7
6	8
7	9



27

part I – you now know

there are many statistical methods that can be applied to different experimental designs

- t-tests
- single factor ANOVA
- factorial ANOVA (case study)
- correlation and regression

ANOVA terminology

- factors, levels, cells
- factorial design
 - between, within, mixed designs

difference between statistical and practical significance

28

part II

learning goals:

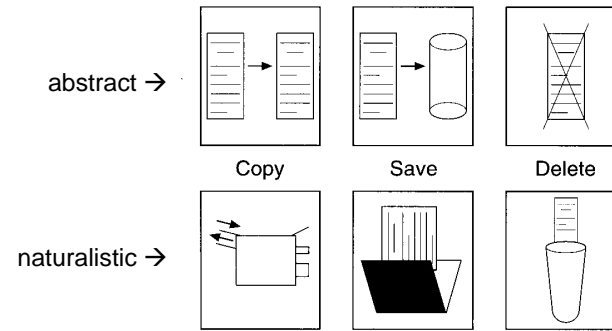
t-test example

- how do I construct an experiment to compare 2 things?
- how can I use Excel to test the alternate hypothesis?

29

t-test example: evaluating icon designs

two styles: which will be easiest for users to remember?



30

hypothesis?

H_1 : natural icons are easier to recall than abstract icons

H_0 : there is no difference in recall between the natural icons and the abstract icons

31

variables

independent variable(s)?

dependent variable(s)?

for this example, let's assume:

1. **number of mistakes** in selection (error rate)
2. **time taken to select icon**

32

experiment task

2 interfaces which are identical in every way except icon design

selection task which can be repeated for both conditions (natural, abstract) - ???

1. produce a document
a natural task: results may be more transferable
2. select appropriate icon in response to a prompt
an artificial task: may be easier to control

first, we may want to give subject time to **learn** icons to avoid learning effects in data

33

experiment task, cont. for this experiment, let's choose:

- **prompt user:** e.g. 'save a document', then require user to select the proper icon
- **one task** = randomly generate a series of X prompts using all of the 3 different types of icon (copy, save, delete) for a given type of icons (either natural or abstract)
- **for each subject:** one task with **each** type of icons

how many times? (what is X?)

→ 10-30 reasonable for this (relatively easy-to-learn) task
consider desired session duration, expected learning curves, effect of boredom on performance.

34

data and analysis

[iconData.xls](#)

sample tables: (many available on web)

http://en.wikipedia.org/wiki/Student's_t-distribution

35

conclusions

reject the null hypotheses and conclude that naturalistic icons are faster to use than abstract ones

would not have been able to draw this conclusion without pairing the observations, i.e., conducting a paired samples t-test

36

re-thinking the analysis

what might be the experiment design if you used an ANOVA instead of a t-test?

37

part II – now you know...

t-test example

- how to construct a simple experiment to compare 2 things
- how to use Excel to test the alternate hypothesis
 - how to “manually” do a t-test and paired t-test
 - how to use Excel’s built-in functions for the same

38

on deck

experiments III

- example of ANOVA from research paper
- value of lecture will be *substantially* greater if you have done the reading

- types of validity
- two types of errors

39