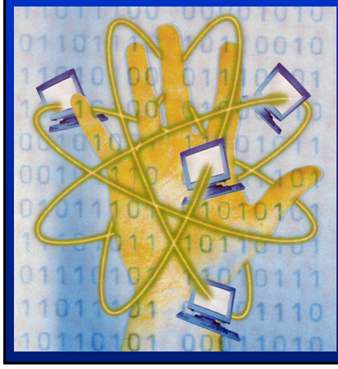
HUMAN-COMPUTER INTERFACE DESIGN

University of Essex



Course EE212

Part 1, Section 6

Evaluation - experiment design and statistics

Computing & Electronic Systems

Autumn 2008

John Foster (module supervisor) and Edward Tsang

DESIGNING EXPERIMENTS - Overview

- An experiment
  - is a sequence of tests
- Choosing variables and conditions
  - to cover the area of interest
  - trying to exclude distracting or unnecessary variables
- Selecting subjects
  - to represent the intended users
- Designing the procedure
  - explaining the experiment to trial users, with consistency
  - allowing for learning effects by altering the test sequence
  - constructing a 'statistically-useful' sequence of tests
  - capturing user responses or reactions - in words, numbers or actions

EVALUATION EXPERIMENTS - aims and techniques

- Aims
  - construct experiment that is simple and unbiased
  - these aims are often *conflicting*
- Techniques
  - replication
  - balancing
  - randomisation
  - blocking structure

SIMPLE EVALUATION - example (1)

- Experiment - for a text-editing application
  - suppose we want to compare the speed of positioning the edit cursor, using either cursor keys or mouse 'move and click'
  - one variable, that can take two values (treatments): keys or mouse
- Method
  - measure the cumulative time taken to position the cursor correctly
  - use same task (text and original position of cursor) for each test, otherwise variation in task will affect the measured time

Treatment	Time(secs)
A	45
B	28

one trial user
uses treatment A - keys
then treatment B - mouse

what is wrong …

learning effect

EVALUATION EXPERIMENTS - example (2)

- Add another trial user - *replication*
  - each user experiences one test - no learning effect is possible
  - first user tries treatment A, second user tries treatment B

Subject	Treatment	Time(secs)
1	A	40
2	B	30

mouse (treatment B) still seems faster

- Fixed one problem, but introduced another problem …
  - measured difference might be due to differences between people
    - maybe user 2 has faster reactions or movements than user 1
    - maybe user 1 has little experience of using a keyboard

EVALUATION EXPERIMENTS - example (3)

- Add another pair (or many pairs) of trial users
  - reduce (average-out) the variation due to the different users
  - first user tries treatment A, second user tries treatment B

Subject	Treatment	Time(secs)
1	A	40
2	B	30
3	A	38
4	B	32

average of treatment A = 39
average of treatment B = 31
difference A_{AV} and B_{AV} = 8
difference between users = 2
(for same treatment)

mouse seems *truly* faster

- Measured difference is much larger than user variation
  - so results form a reasonable conclusion, without more analysis
  - not always like that …

6

EVALUATION EXPERIMENTS
- observations (about observations)

- A single measurement (observation)
  - of each treatment is not sufficient …
  - it does not allow *independent* assessment of treatment variability
- What if the difference between treatments is small ?
  - increase the number of trial users
    - variation due to N users typically changes as $N^{-0.5}$
  - use statistical methods to decide if measured difference is significant
- Experiments with large numbers of subjects
  - suppose there are 64 trial users, and each test takes 5 minutes
    - total time for experiment is at least 5 hours 20 minutes
    - that's enough time for the room to get hot or stuffy,
    - enough time for the amount of daylight to change, etc.
- *Balance* the sequence of treatments - not 'all A' then 'all B'
  - *randomise* the test sequence - use tables of random numbers

7

SENSITIVE EXPERIMENTS
- to measure small differences

- Is there a better way …
  - than adding more and more trial users ?
    - yes - use each trial user more than once
- Earlier examples
  - all used the 'between subjects' method of experimental design
    - each subject tested one treatment
  - measured differences were between treatments, but also between users
- 'Within subjects' design - more efficient use of trial users
  - each user tests all of the treatments
  - advantage is reduced variability due to different users - more precision
    - each user acts as their own 'control' or reference case
  - disadvantage is *much* greater opportunity for learning effects to occur
  - total time for the experiment does not change

8

SENSITIVE EXPERIMENTS
- 'within subjects' design method

- *Blocking* structure
  - where each user contributes a block (here, a row) of results
  - randomly allocate trial users to blocks

Subject	Period 1 Treatment A	Period 2 Treatment B	Subject average
1	40	28	34
2	45	30	37.5
3	38	27	32.5
4	36	32	34
Treatment average	39.75	29.25	34.5 Overall average

possible to calculate:
averages for each treatment
averages for each user
these are independent …
greater precision in measuring
both treatment and user effects

learning effect is a
serious problem

- One further improvement to the design is needed …
  - balance the treatment sequence

9

SENSITIVE EXPERIMENTS
- 'within subjects' design method

- Alternate the order of treatments in different blocks
  - known as 'Latin Squares' design - fully balanced, randomised block
  - there are as many users as treatments as periods

Subject	Period 1	Period 2	Subject average
1	A 41	B 27	34
2	B 31	A 44	37.5
3	A 39	B 26	32.5
4	B 33	A 35	34

Treatment averages: A - 39.75, B - 29.25 | 34.5 Overall average
Period averages: 1 - 36, 2 - 33

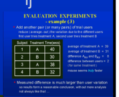
users are the 'blocking factor'
on the rows
periods are 'blocking factor'
on the columns
contribution of treatments to
the total variability is
independently measurable

- Learning effects cancel out
  - *if* they are symmetrical ie. learning using A = learning using B

10

STATISTICAL METHODS
- modelling the variability of the results

- For example (3) slide 6
  - mathematical model of sources of variability in the test score x_{ij}
  - assumes there is some overall or average performance
  - some effect due to the different treatments
  - and some residual or random error



$$x_{ij} = \mu + t_i + e_{ij} \qquad (i = 1,2;\ j = 1,\dots,4)$$

where x_{ij} = observation on subject j and treatment i
μ = overall average
t_i = effect of ith treatment
e_{ij} = random error

the grand mean x of all observations x_{ij} is an estimate of μ
mean of all observations on ith treatment, less x, estimates t_i

11

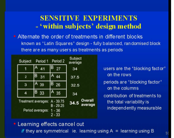
STATISTICAL METHODS
- modelling the variability of the results

- For the final (Latin Squares) example slide 10
  - model of sources of variability in test score x_{ijk}
  - assumes there is some overall or average performance
  - separate effects due to different treatments, users and periods
  - and some residual or random error - should be smaller than Example (3)

$$x_{ijk} = \mu + t_i + s_j + p_k + e_{ijk} \qquad (i = 1,2;\ j = 1,\dots,4;\ k = 1,2)$$

where x_{ijk} = observation on subject j with treatment i in period k
μ = overall average
t_i = effect of ith treatment
s_j = effect of jth subject (block)
p_k = effect of kth period
e_{ijk} = random error

the grand mean x of all observations x_{ijk} is an estimate of μ
mean of all observations on ith treatment, less x, estimates t_i
mean of all observations on subject j, less x, estimates s_j
mean of all observations on period k, less x, estimates p_k



12

STATISTICAL METHODS - formal terminology

- In comparative experiments
  - to estimate the probability that the differences measured in the experiment might have happened by chance
  - if probability is low enough, result is said to be *significant* at that level
  - eg. if result is significant at 5% level, then only 1 in 20 chance that this result happened by chance
- Use of hypotheses
  - null hypothesis assumes there is no difference between treatments
  - alternative hypothesis assumes some difference between treatments
  - assume null hypothesis is true, then look for significant differences - if found reject null hypothesis and accept alternative hypothesis

13

STATISTICAL METHODS - data distributions and significance tests

- Method used depends on :
  - number of treatments
  - assumptions about the nature, and internal dependencies, of the data
- Distribution of experimental data
  - often assume random processes, so data follows normal distribution (Gaussian)
- Tests for statistical significance
  - if there are two treatments - ' t-test'
  - more than two treatments - 'ANOVA' (analysis of variance), 'F-test' these only test the null hypothesis, if any significance is found then further tests are used to establish which variable is significant - ' multiple t-test' and ' Tukey's test' (Wetherill, 1981)

14

Scales of Measurement

- Nominal
  - Categories, such as colours (red, blue), gender, marital status
- Ordinal
  - Rank order, e.g. 1st, 2nd and 3rd as in horse racing
- Interval
  - Like ordinal, but difference between 1st and 2nd is the same as distance between 2nd and 3rd
- Ratio
  - Such as mass, length, age
- Reference: S. S. Stevens, *On the theory of scales of measurement*, Science 103, 1946, pp667-680

STATISTICAL METHODS - other distributions and issues

- Normal distribution of experimental data
  - is not always true
- Frequency measurements
  - eg. number of users in a category - follow Poisson distribution
- Ranking measurements
  - eg. ratings in questionnaire answers - use Mann-Whitney / Wilcoxon tests (Gibbons, 1971)
- Statistical significance
  - is a mathematical idea …
  - a non-significant difference is *not* the same as proof of no difference just means that the experimental conditions found no difference
  - significant difference is not *necessarily* interesting or useful difference

15

EVALUATION - summary

- Effective evaluation is not easy
  - classic mistakes are easy to make
  - but good evaluations save time / money - avoid rework after production
- Evaluate early and often
  - use simple methods at first, more complex methods later in design
  - variety of opinion, from experts and several kinds of users, is valuable
- Trial users
  - think about potential bias (intended or not) when you select them
- The design of evaluations is important
  - advance planning, for all kinds of evaluation, is essential
  - writing good questionnaires, and holding useful interviews, needs care
  - sequence of tests, and organisation of arithmetic, makes big difference
  - statistical analysis is powerful, but significant results not always useful

16