HUMAN-COMPUTER INTERACTION THIRD EDITION

DIX FINLAY ABOWD BEALE

chapter 9

evaluation techniques (part 2)

Evaluating Implementations

Requires an artefact:

a simulation, a prototype, or a full implementation

Experimental evaluation

- controlled evaluation of specific aspects of interactive behaviour
- evaluator chooses hypothesis to be tested
- a number of experimental conditions are considered which differ only in the value of some controlled variable.
- changes in behavioural measure are attributed to different conditions

Experimental factors

- Subjects (i.e., the users, *aka* 'participants')
  - who – representative, sufficient sample
    - not the programmer's friend, boss, etc.
    - huge variability in performance of individuals
- Variables
  - things to modify and measure
- Hypothesis
  - what you'd like to show
- Experimental design
  - how you are going to show it
  - Includes 'Protocol' – what the subjects do

Variables

- independent variable (IV)
  - characteristic changed to produce different conditions
  - e.g. interface style, number of menu items
- dependent variable (DV)
  - characteristics measured in the experiment
  - e.g. time taken, number of errors.

Hypothesis

- prediction of outcome
  - framed in terms of IV and DV

    e.g. “error rate will increase as font size decreases”
- null hypothesis:
  - states no difference between conditions
  - aim is to disprove this

    e.g. null hyp. = “no change with font size”

Experimental design

- "within groups" design (also called "repeated measures")
  - each subject performs experiment under each condition
  - transfer of learning possible (practice makes performance better; or alternatively fatigue or boredom makes it worse)
  - less costly and less likely to suffer from user variation (each user is compared to themselves)
- between groups design
  - each subject performs under only one condition
  - no transfer of learning
  - more users required

Within v. Between

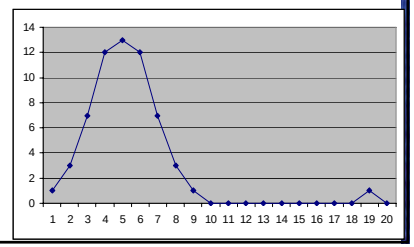
- Consider a test on the difference of beer v. vodka martinis on reaction time
  - Null hypothesis – no difference in increase in reaction time between the two beverages
- Design 1:
  - 30 people try beer; 30 other people try vodka – D.V. is change in reaction time pre- v. post drinking
    - Not bad – be sure to randomize who goes into beer group v. vodka group
    - But 'power' of the experiment will be reduced due to the great variability of individuals in reaction to alcohol

Within v. Between (contd.)

- Design 2:
  - All 60 people first try beer, then immediately try vodka
    - Problem of carryover effect
- Better Design:
  - All 60 try beer, then *a week later* try vodka
    - Now each individual is compared with themselves
    - Still possible problem of ordering effect (e.g., they might get a little better at the reaction time test)
- Best Design:
  - 30 try beer, then a week later vodka; 30 try vodka and then a week later beer

Analysis of data

- Before you start to do any statistics:
  - look at data (e.g. average=5.25 – but 4.9 without outlier)
  - save original data
- Choice of statistical technique depends on
  - type of data
  - information required
- Type of data
  - discrete
    - finite number of values
    - may be ordered, or unordered (e.g., colors)
  - continuous
    - any value



Analysis - types of test

- parametric
  - assume normal distribution
  - robust
  - powerful
- non-parametric
  - do not assume normal distribution
  - less powerful
  - more reliable
- contingency table
  - classify data by discrete attributes
  - count number of data items in each group

Analysis of data (cont.)

- What information is required?
  - is there a difference?
  - how big is the difference?
  - how accurate is the estimate?
- Parametric and non-parametric tests mainly address first of these

ANOVA – analysis of variance

- Quite easy to test whether there's a significant difference between groups in Excel
  - Need to invoke Tools/Add-ins/Analysis Toolpack to enable
  - Then just apply Tools/Data Analysis/ANOVA: Single Factor to the data

ANOVA from Excel

Say we have three columns of numbers representing the time to complete a task for 5, 5 and 7 users using three variations of an interface

If P-value < 0.05 then we usually say the result is 'significant' (result is more than expected chance variation)

Anova: Single Factor

SUMMARY

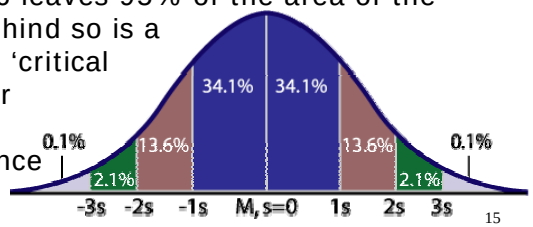
Groups	*Count*	*Sum*	*Average*	*Variance*
Group 1	5	82	16.4	9.3
Group 2	5	67	13.4	10.3
Group 3	7	79	11.28571	3.904762

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	76.28908	2	38.14454	5.244339	0.019959	3.738892
Within Groups	101.8286	14	7.273469			
Total	178.1176	16				

When is a difference a difference?

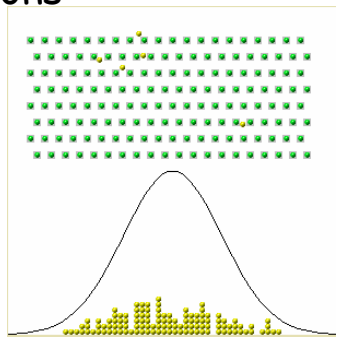
- In the world of parametric stats, we look for a statistic to be large enough to be 'significant'
  - On the Gaussian ('normal') curve a |Z|=1.96 leaves 95% of the area of the curve behind so is a common 'critical value' for claiming significance



Parametric assumptions

- Parametric statistics assume that some mathematically elegant assumptions hold true for the data
  - E.g., ANOVE (and standard 'regression') assume, among other things, normally distributed random error
  - Trivia: The mathematical form of the *probability density function* for the normal distribution is remarkably formidable
    - Centres on mean, μ, and is flattened by standard deviation, σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Galton machine simulates normal distribution (*aka* 'bell curve')

Exponential distribution models time between events happening with a constant average rate

Experimental studies on groups

More difficult than single-user experiments

Problems with:

- subject groups
- choice of task
- data gathering
- Analysis

- Unfortunately (in terms of experimental requirements) a lot of things that are interesting in the real world, involve computers mediating group behaviour

Subject groups

larger number of subjects
⇒ more expensive

longer time to `settle down'
… even more variation!

difficult to timetable

so … often only three or four groups

Groups (contd.): The task

must encourage cooperation

perhaps involve multiple channels

options:

- creative task — e.g. '*write a short report on …*'
- decision games — e.g. desert survival task
- control task — e.g. ARKola bottling plant

Groups (contd.): Data gathering

several video cameras
+ direct logging of application

problems:

- synchronisation
- sheer volume!

one solution:

- record from each perspective

Groups (contd.): Analysis

N.B. vast variation between groups

solutions:

- 'within groups' experiments (each group works under various conditions)
- micro-analysis (e.g., gaps in speech)
- anecdotal and qualitative analysis

look at interactions between group and media

controlled experiments may `waste' resources!

Groups (contd.): Field studies

Experiments dominated by group formation

Field studies more realistic:

distributed cognition ⇒ work studied in context
real action is *situated action*
physical and social environment both crucial

Contrast:

psychology – controlled experiment
sociology and anthropology – open study and rich data

About statistics

- It's an amazingly complex field
  - A lot of hidden complexities in running experiments and saying that the observed differences really make a difference
    - 'threats to validity' – are those things that make it possible that your experimental conclusion is in error
      - Threats to internal validity: like carryover effects, or lack of randomization
      - Threats to external validity: like that your whole population of subjects were unusual in some way, or the task was not representative of real use of the tool
  - When the outcomes are serious (e.g., medical trials) professional statisticians are always used in design of the experiment as well as analysis and reporting of the findings
  - Plenty of texts and courses on stats available (the Wikipedia is pretty good on this topics, too – e.g., for ANOVA)