HUMAN-COMPUTER INTERACTION THIRD EDITION

DIX FINLAY ABOWD BEALE

chapter 9

evaluation techniques

Analysis of data

- Before you start to do any statistics:
  - look at data
  - save original data
- Choice of statistical technique depends on
  - type of data
  - information required
- Type of data
  - discrete - finite number of values
  - continuous - any value

Measurement

- **Levels of measurement** = a particular way of assigning numbers to measure something.
  - amount of measurable information
  - different relations
  - permissible maths operations
  - isomorphism
- **Types**
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Levels of measurement

Nominal

- Numbers = labels only
- Sorting things into categories
- Math operations: count only
- Statistical descriptor: mode

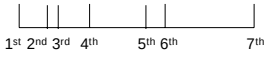
E.g. gender: Male/Female
Religious affiliation: atheist=1, Buddhist=2, Christian=3
Political Views: liberal, moderate, conservative
Birthplace, Hair colour, Blood type, Telephone numbers, Social security numbers, Occupation, Marital status

Levels of measurement

Ordinal

- All properties of nominal scale
- Number show a relative ranking: "*greater than*"
- How much greater ?
- Math operations: rank
- Statistical descriptor: median

E.g. Econ. and Social Class: Upper, Middle, Lower
Military rank, School grade



Levels of measurement

Ordinal more examples...

- Measuring satisfaction: very satisfied, somewhat satisfied, not satisfied at all.
- Measuring importance: not very important, fairly important, very important, most important thing.
- Measuring frequency: always, frequently, sometimes, occasionally, never.
- Measuring condition: excellent, good, fair, poor.

Levels of measurement

Interval

- All properties of ordinal scales
- Distance between any two data points can be determined
- Math operation: f(x)=ax+b
- Statistical descriptor: any

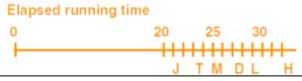
E.g. Fahrenheit and centigrade temperature
F = 9/5 C +32
constructed measures – IQ scores

Levels of measurement

Ratio

- The most powerful
- All properties of interval scales
- **true zero point**
- Math operation: any
- Statistical descriptor: any

E.g. Height, Weight, Distance
Income in Dollars; Temperature in Kelvin
Age; Length of residence in a given place



Levels of measurement

- **Nominal level**
  Do you like ice cream?
  ___ 1. Yes ___ 2. No
- **Ordinal level**
  Rank the next ice cream flavours in order of preferences from best to worst
  ___ Chocolate
  ___ Vanilla
  ___ Strawberry
  ___ Mint

Levels of measurement

- **Interval level**

  Semantic differential scale
  This brand of ice cream is

	1	2	3	4	5	
Excellent						Poor

- **Ratio level**

  How many boxes of ice cream do you usually consume on a monthly base?
  ___ boxes of ice cream

Levels of measurement

Comparing two Nominal values

Categories of two cases
- are the same
- are different

Comparing two Ordinal values

- one case comes before or after another case
- not *how far* before or after
- no arithmetic

Levels of measurement

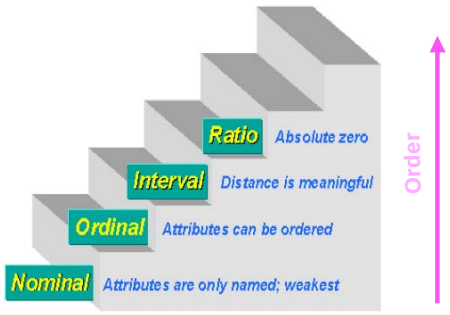
Comparing two Interval values

- both order and distance between points
- not *how big* one value is as a fraction of another
- Addition, subtraction

Comparing two Ratio values

- order, distance between points
- Relative size – as ratio
- Addition, subtraction, multiplication, division

Levels of measurement



Levels of measurement

Nominal				
Ordinal				
Interval				
Ratio				

Levels of measurement

	Same or different	Bigger or smaller	How much bigger?	How far of 0?
Nominal				
Ordinal				
Interval				
Ratio				

Levels of measurement

	Same or different	Bigger or smaller	How much bigger?	How far of 0?
Nominal	X			
Ordinal	X	X		
Interval	X	X	X	
Ratio	X	X	X	X

Levels of measurement

- Important: Meaning of data
- When in question, seek the highest level of measurement possible.
- Can always convert higher level to lower level but NOT the reverse.

Levels of measurement

Why Cares?

- they use different procedures
- they produce different results
- they determine partly the ***type of statistics*** that can be applied to the data.

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

Experimental studies on groups

More difficult than single-user experiments

Problems with:
- subject groups
- choice of task
- data gathering
- analysis

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

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

Data gathering

several video cameras
+ direct logging of application

problems:
- synchronisation
- sheer volume!

one solution:
- record from each perspective

Analysis

N.B. vast variation between groups

solutions:
- within groups experiments
- micro-analysis (e.g., gaps in speech)
- anecdotal and qualitative analysis

look at interactions between group and media

controlled experiments may `waste' resources!

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

Observational Methods

Think Aloud
Cooperative evaluation
Protocol analysis
Automated analysis
Post-task walkthroughs

Think Aloud

- user observed performing task
- user asked to describe what he is doing and why, what he thinks is happening etc.

- Advantages
  - simplicity - requires little expertise
  - can provide useful insight
  - can show how system is actually use
- Disadvantages
  - subjective
  - selective
  - act of describing may alter task performance

Cooperative evaluation

- variation on think aloud
- user collaborates in evaluation
- both user and evaluator can ask each other questions throughout

- Additional advantages
  - less constrained and easier to use
  - user is encouraged to criticize system
  - clarification possible

Protocol analysis

- paper and pencil – cheap, limited to writing speed
- audio – good for think aloud, difficult to match with other protocols
- video – accurate and realistic, needs special equipment, obtrusive
- computer logging – automatic and unobtrusive, large amounts of data difficult to analyze
- user notebooks – coarse and subjective, useful insights, good for longitudinal studies

- Mixed use in practice.
- audio/video transcription difficult and requires skill.
- Some automatic support tools available

automated analysis – EVA

- Workplace project
- Post task walkthrough
  - user reacts on action after the event
  - used to fill in intention
- Advantages
  - analyst has time to focus on relevant incidents
  - avoid excessive interruption of task
- Disadvantages
  - lack of freshness
  - may be post-hoc interpretation of events

post-task walkthroughs

- transcript played back to participant for comment
  - immediately → fresh in mind
  - delayed → evaluator has time to identify questions
- useful to identify reasons for actions and alternatives considered
- necessary in cases where think aloud is not possible

Query Techniques

Interviews
Questionnaires

Interviews

- analyst questions user on one-to -one basis usually based on prepared questions
- informal, subjective and relatively cheap
- Advantages
  - can be varied to suit context
  - issues can be explored more fully
  - can elicit user views and identify unanticipated problems
- Disadvantages
  - very subjective
  - time consuming

Questionnaires

- Set of fixed questions given to users
- Advantages
  - quick and reaches large user group
  - can be analyzed more rigorously
- Disadvantages
  - less flexible
  - less probing

Questionnaires (ctd)

- Need careful design
  - what information is required?
  - how are answers to be analyzed?
- Styles of question
  - general
  - open-ended
  - scalar
  - multi-choice
  - ranked

Physiological methods

Eye tracking
Physiological measurement

eye tracking

- head or desk mounted equipment tracks the position of the eye
- eye movement reflects the amount of cognitive processing a display requires
- measurements include
  - fixations: eye maintains stable position. Number and duration indicate level of difficulty with display
  - saccades: rapid eye movement from one point of interest to another
  - scan paths: moving straight to a target with a short fixation at the target is optimal

physiological measurements

- emotional response linked to physical changes
- these may help determine a user's reaction to an interface
- measurements include:
  - heart activity, including blood pressure, volume and pulse.
  - activity of sweat glands: Galvanic Skin Response (GSR)
  - electrical activity in muscle: electromyogram (EMG)
  - electrical activity in brain: electroencephalogram (EEG)
- some difficulty in interpreting these physiological responses - more research needed

Choosing an Evaluation Method

when in process:	design vs. implementation
style of evaluation:	laboratory vs. field
how objective:	subjective vs. objective
type of measures:	qualitative vs. quantitative
level of information:	high level vs. low level
level of interference:	obtrusive vs. unobtrusive
resources available:	time, subjects, equipment, expertise