p < 0.01

Understanding Statistics for HCI and Related Disciplines

Part 3 – Gaining Power

the dreaded 'too few participants'

Alan Dix

http://alandix.com/statistics/

don't you hate it when …

=== Reviewer 2

Although the premise seems sound there are too few participants to be able to draw meaningful conclusions.

=== Reviewer 3

P= 0.673 not significant

statistical power – what is it?

if there is a real effect
how likely are you to be able to detect it?

avoiding false negatives

increasing power

standard approach …
add more participants

but not the only way!

can get more power …
but often sacrifice a little generality
need to understand and explain

with great power comes great responsibility ;-)

noise–effect–number triangle

three routes to gain power

recall Stats 101 (for simple data)

σ – standard deviation of data

often 'noise' – things you can't control or measure
e.g. individual variability

s.e. – standard error of mean (s.e.)

the accuracy of your estimate (error bars)

s.e. = $\sigma / \sqrt{n}$ (if σ is an estimate $\sqrt{n-1}$)

to half standard error you must quadruple number.

effect size

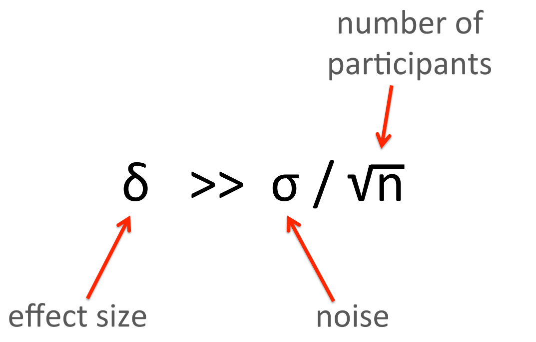
how big a difference do you want to detect?

call it δ

the accuracy (s.e.) needs to be better than δ

$$\delta >> \sigma / \sqrt{n}$$

Understanding Statistics for HCI and Related Disciplines – Alan Dix

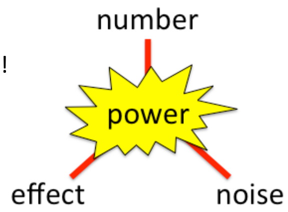


Understanding Statistics for HCI and Related Disciplines – Alan Dix

noise–effect–number triangle

to gain power
address any of these

not just more subjects!



Understanding Statistics for HCI and Related Disciplines – Alan Dix

general strategies

increase number
- the standard approach … but …
- square root often means very large increases

reduce noise
- control conditions (physics approach)
- measure other factors and fit (e.g. age, experience)

increase effect size
- manipulate sensitivity (e.g. photo back of crowd!)

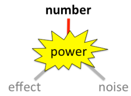
Understanding Statistics for HCI and Related Disciplines – Alan Dix

Understanding Statistics for HCI and Related Disciplines – Alan Dix

subjects

control or manipulate who

more subjects or trials

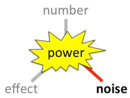


more subjects
- average out *between* subject differences

more trials
- average out *within* subject variation
  e.g. Fitts' Law experiments

... but both both may need lots
e.g. to reduce noise by 10, need 100 times more

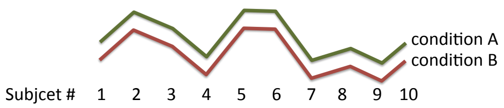
within subject/group



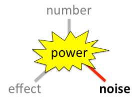
cancels out between subject variation

helpful if effect reasonably consistent
but between subject variability high

may cause problems with order effects, learning



narrow/matched users

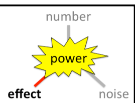


aims to reduce between subject variation

choose subjects who are very similar to each other or in some way have matched
e.g. balance gender, skills

allows between subject experiments
how do you know what is important to match?

targeted user group



aims to increase the effect size

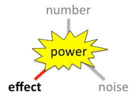
choose group of users who are likely to be especially affected by the intervention
e.g. novices or older users

but ... generalisation to other users will be by theoretical argument not empirical data

tasks

control or manipulate what

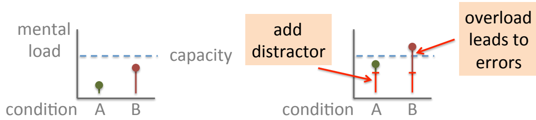
distractor tasks



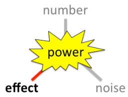
aim to saturate user's cognitive resources
so make them more sensitive to intervention

e.g. count backwards while performing task

helpful when coping mechanisms mask effects



targeted tasks

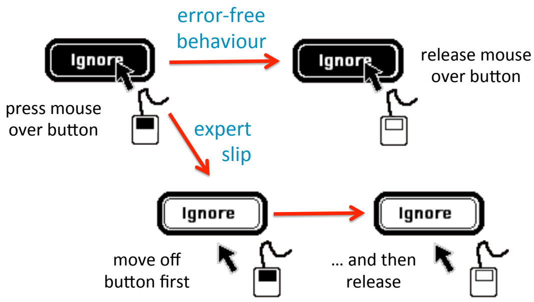


design a task that will expose effect of intervention

e.g. trouble with buttons paper (expert slip)

… but … care again with generalisation!

example: trouble with buttons



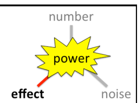
trouble with buttons (2)

novices:
- work more slowly – less likely to make slip
- notice lack of semantic feedback – so they recover

experts:
- act quickly – so make more slips
- focused on next action, so miss feedback

problem:
- experts slips don't happen often … never in experiments
- needed to craft task to engineer expert slips

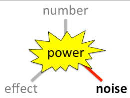
demonic interventions!



extreme version create deliberately nasty task!
- e.g. 'natural inverse' steering task
  - Fitts' Law-ish experiment
  - added artificial errors to cause overshoots

… but … again generalisation
… and … subjects may hate you!

reduced vs. wild

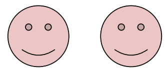


in the wild has lots of extraneous effects
= noise!

control environment => lab or semi-wild

reduced task
- e.g. scripted use in wild environment

reduced system
- e.g. mobile tourist app with less options



Understanding Statistics for HCI and Related Disciplines – Alan Dix