design and test

- designing the right experiment
- choosing the right test

design

what to vary

what to measure

what can go wrong ...

what to look for in the results

what kind of data

- continuous:
  e.g. time to complete task (12.73 secs)
- discrete ...
  - arithmetic:
    e.g. number of errors (average makes some sense)
  - ordered / ordinal:
    e.g. satisfaction rating (?average rating?)
  - nominal / categorical:
    e.g. menu item chosen ((File+Font) / 2 = FIml ?)

what kind of variable

independent*:

- what you choose to vary

dependent:

- what you want to measure

extraneous:

- what you haven’t thought of!

* N.B. different meaning of independent

several independent variables

- fix all but one
  - ✘ doesn't tell you about interactions
    (e.g. change menu and icon metaphor)
  - ✘ lots of little experiments
  - ✔ simple!
- vary several
  - ✘ one enormous experiment
  - ✘ confusing effects and difficult sums
  - ✔ let the computer do them!

several dependent variables

- common in field studies
- not 'independent' of each other
  (e.g. speed and accuracy)

statistical connection $\neq$ causality

(may be due to common cause)

extraneous variables

- try to think of them
- control them:
  - fix them
    - level playing field
  - balance them
    - don't put all the experts in the same group!
- at least measure them
    - become like more dependent variables
- <u>very</u> difficult for interface design ideas

what can go wrong

- too much variability
  especially with people!
- confusing effects (aliasing)
  e.g. all experts in one group
- wrong tests
  false results (+ve or –ve)

too much variability

either:

- increase number

  double sensitivity $\Rightarrow$ quadruple size

- cancel out variability

  – paired tests

basis for pairing

- people are very variable
  (also other things like farm fields!)
- different personal traits:
  expertise, dexterity, intelligence
- often similar effects on results:
  faster / slower, more / less accurate

paired experiment

- try several things on each person
- basis of analysis
  - differences within individual
- cancels out
  - differences between individuals

example

(from Dix, Finlay, Abowd and Beale, 1993)

Subject number	Presentation order	(1) Natural (secs.)	(2) Abstract (secs.)	(3) Subject mean	(4) Natural (1)–(3)	(5) Abstract (2)–(3)
1.	AN	656	702	679	–23	23
2.	AN	259	339	299	–40	40
3.	AN	612	658	635	–23	23
4.	AN	609	645	627	–18	18
5.	AN	1049	1129	1089	–40	40
6.	NA	1135	1179	1157	–22	22
7.	NA	542	604	573	–31	31
8.	NA	495	551	523	–28	28
9.	NA	905	893	899	6	–6
10.	NA	714	803	759	–44	44
	mean (μ)	698	750	724	–26	26
	s.d (σ)	265	259	262	14	14
		s.e.d. 117			s.e. 4.55	
	Student's t	0.32 (n.s.)			5.78 (p<1% 2 tailed)	

beware

- transfer effects:
  positive – training
  negative – confusion
- randomised order helps
  – but look at data
- use the right test!

other types of design

- factorial:
  - try everything with everything
- Latin square:
  - assume no interactions

	A	B	C	D
a	α	β	γ	δ
b	γ	δ	α	β
c	β	α	δ	γ
d	δ	γ	β	α

- as it comes – just measure:
  - often only option for fieldwork
  - don't worry let stats package sort it out!

basic principles

- reduce variability
  - control extraneous variables
  - use same subjects
- avoid aliasing
  - try to balance out independent variables
  - if uneven stats more difficult but possible
- replication
  - to improve averages
  - to estimate error
- always keep your raw data!

results – what to look for

- main effects
  - changing A affects B

5	10	7.5

s.e. =1.3

- trends
  - increasing A increases B

5	10	17

- interactions
  - when both A&C ...

5	4
3	9

- the unexpected
  - +ve or –ve results

use the right test

- tests make assumptions
  - pairing, normality, independence

- if not true of <u>your</u> data:

  false positive
  - think there's an effect when there isn't

  false negative
  - miss a real effect (see example)

kinds of test

- parametric
  - 'well behaved' data
- non-parametric
  - use ordering only
- contingency tables
  - for 'occurrence' data
- Baysian statistics
  - use prior knowledge

parametric tests

- assume a distribution
  - often Normal, but not always
- many are robust
  - OK if data is nearly normal!
- data distribution ≠ test assumption
  - choose different test
  - modify data e.g. log transformation

non-parametric tests

- no distribution assumed
- simply use relative size of data
- do assume independence
- little ‘power’
  - effects need to be large

  ⇒ use parametric when possible

contingency tables

- if dependent variable(s) are <u>nominal</u>
  (that is no intrinsic order… e.g. red / green / blue)
- use occurrence in each category
- still assume independence
- no assumed distribution for data
  - not normally classed as non-parametric
  - use χ^2 in testing (actually an approximation)

non-independent data

recall:

- positive correlation
  - decreases <u>measured</u> variability
  - false positives
- negative correlation
  - increases <u>measured</u> variability
  - false negatives

✔ can modify tests ... ask an expert!

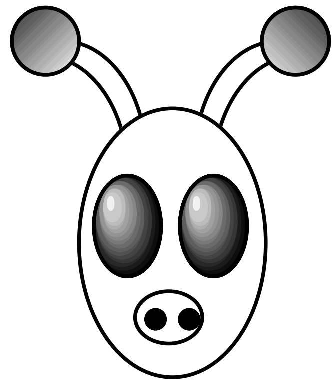
Baysian statistics

- philosophical stance

? what do you know about the world

- traditional statistics
  - nothing!
  - reason from unknowledge
- Baysian statistics
  - 'prior' probabilities
  - reason from guess-timates

Baysian thinking



(what you think before any evidence)

prior probability of meeting:

a Martian	=	0.000 001
a human	=	0.999 999

probability of having antennae:

if Martian	=	1
if human	=	0.001

! you meet someone with antennae

posterior probability of being:

a Martian	≈	0.001
a human	≈	0.999

Baysian issues

- how do you get the prior?
  - actually often doesn't matter too much!
  - traditional stats rather like uniform prior
- handling multiple evidence
  - can re-apply iteratively
  - problems with interactions
- internecine warfare
  - traditionalists and Baysians often fight