Tuesday, October 17, 2006
10:14 AM

Exp Design (2nd ed.)

From last time:

- formulate a hypothesis
- specify indep. variables (what gets manipulated?)
  - (how many IVs ≃ experiment complexity ↓ keep it tractable)
  - note the levels of these
- specify factorial design, if that's appropriate
  - e.g. 2×3 experiment — factorial design
    - 2 IVs, one at 2 levels
    - " " 3 levels
    - ⟹ 6 trials needed

- Keep track (stipulate) DVs — what gets measured

objective { Performance measures : efficiency (speed) eg time to complete
effectiveness (accuracy) eg. # errors

subjective { satisfaction eg. questionnaire
"SUS" metrics (usability)

objective { process measures : eye movements eg. fixations, fix. duration, transitions
dialog — think-aloud

Tuesday, October 17, 2006
10:14 AM

Rule of thumb: vary only one condition (IV) at a time,
keep all else constant
(hard to do practically — various conditions
(cause & effect) — initiate — did participants eat lunch
before conducting trials ...
goofy stuff like this helps
control)

Tuesday, October 17, 2006
10:14 AM

Methodology:

- Hypothesis: what outcomes are expected? Based on what assumptions / prior knowledge?
- Design: type of design: observational study? laboratory or in-field?
  within-subjects / between-subjects
  factorial, e.g. 2x3, 4x3x2, ...
- Participants: demographics (M, F, average age, corrected vision)

→ counterbalanced order of trials, e.g. 2x3 within-subjects ⇒ 6 conditions

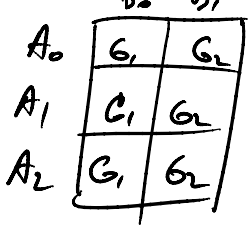
S1	A, B, C, D, E, F	Latin Square for counterbalancing repeated measures order — counteracts learning/fatigue effects
S2	B, C, D, E, F, A	
S3	C, D, E, F, A, B	
S4	D, E, F, A, B, C	
S5	E, F, A, B, C, D	
S6	F, A, B, C, D, E	

(or randomize)

Tuesday, October 17, 2006
10:14 AM

- Apparatus: Tobii eye tracker, etc
  - sample rate
  - accuracy
- Procedures:
  - what subjects were told, instructions given?
    any training? calibration (how many points? when was it done?)
- Tasks:
  - what did subjects do
  - how much did they know — how much was divulged)
  - does it deceive participants, but does it give anything away either
    ↳ info may influence performance
    (also Hawthorne effect)

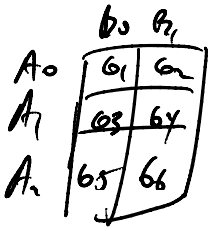
- factorial design:



2 x 3
↑ 2 levels of B
↖ 3 levels of A

A: varies with subj
B: varies between subjects

"mixed" design "nested"



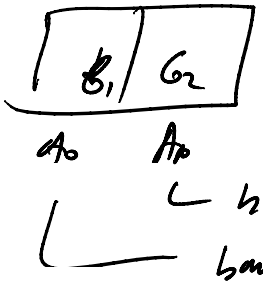
2x3 fully independent

↓

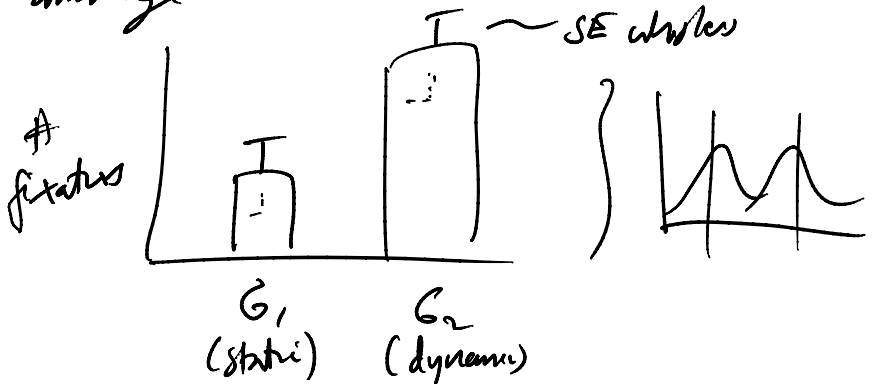
important for analysis

- Statistical Analysis : math 'tool' (eg. ANOVA, t-test, etc)
  depends on type of data
  & notion of independence

- basic approach:
  - collect data for groups of participants
  e.g. 2 groups: banner ad on web page



- run 10 participants per group, count fixations, take average

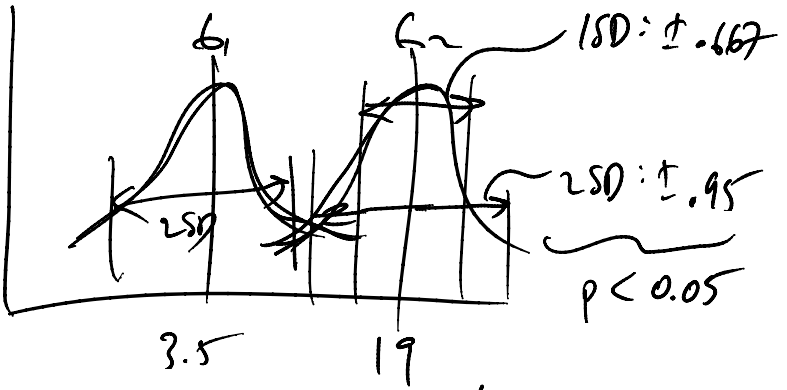


- In Excel:

	G_1	G_2
S_1	3	10
S_2	4	20
S_3	2	17
S_4	1	16
S_5	3	12
.		
S	1	1

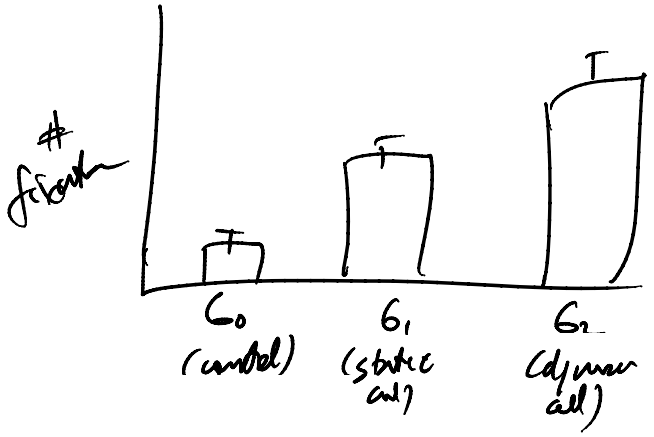
mean: 3.5 19

- is there an effect? independent t-test (between-subjects)



if no overlap, effect is significant

- t-test: comparing 2 means (2 columns of data)
- Let's add in control condition - no ads
- 5 people per group, $n = 15$

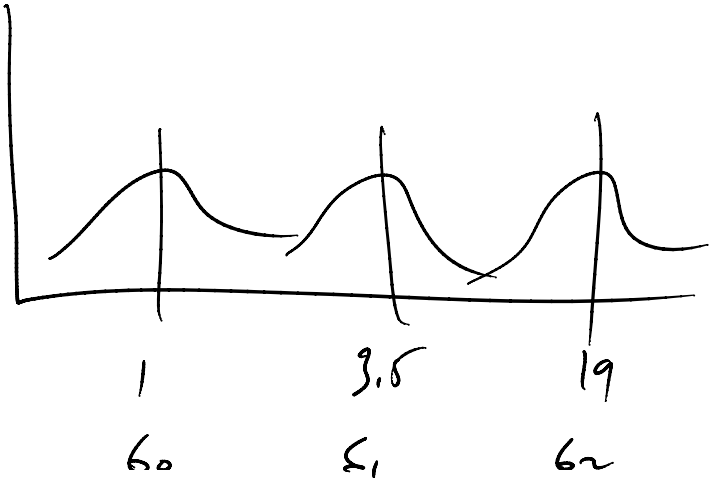


- can't do a 3-way t-test — use ANOVA (Analysis of variance) — like a multi-column extension of t-test
- ANOVA: plug in 3 columns of data, get back $F = 4.761$ $p = 0.03$

here, we see an significant effect $p < 0.05$

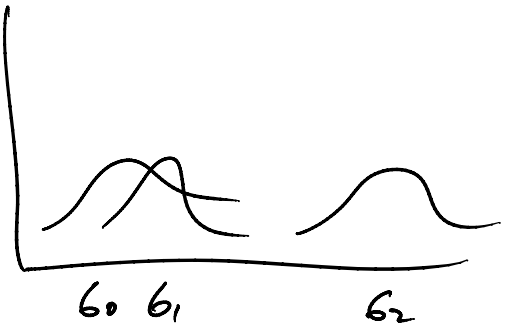
to report, $F(2,12) = 4.761$, $p < 0.05$

df: degrees of freedom
should add up to $n-1$ (14)
first number is conditions -1 (columns)
$3-1 = 2$, $14-2 = 12$



- ANOVA says 2 or more "Gaussians" are farther apart than 2 SDs (at $p < 0.05$)
  $\Rightarrow$ 2 don't overlap — but which two?

Could be:



ANOVA still reports $p < 0.05$
but not where the effect is.

You need to do 'post-hoc' analysis (?)

$\Rightarrow$ pairwise mean comparisons — to find out whether effect is between
μ_0, μ_1
μ_0, μ_2
μ_1, μ_2

- can do t-test here on pairs (if you can assume data fits normal dist.)

or Kruskal-Wallis test if can't make this assumption.

For pair of data:

	(between) independent	(within) non-independent
Nominal	Chi-square	Sign test
Ordinal	Mann-Whitney U	Wilcoxon signed pairs
Parametric	t-test for indep. means	t-test

Nominal: usually most meaningful — categorical data, e.g. # blue-eyed people

Ordinal: only denotes order (e.g. top 5 scores on midterm)

Parametric: uniform, equal interval scale

(basically what we have to use
are fixations on an equal interval scale?
do 4 fixations over an item (RoI)
denote "twice" something as 2 fixations?
(as informative,
demanding, interesting?)

Maybe: cognitive load is represented by
fixations (operational assumption)

Tuesday, October 17, 2006
10:14 AM

For df > 2

	indep	non-indep
non-parametric	chi-sq.	Kruskal-Wallis
parametric	ANOVA	ANOVA

Most people use ANOVA because:

- multiple conditions (multiple columns of data)
- data is assumed to fit normal dist
- data is on equal distance scale

these should hold, if not don't use ANOVA

Tuesday, October 17, 2006
10:14 AM

Statistical Tools / packages:

- Excel has stats package – need to enable
  - not the most powerful
  - ok for descriptive (mean, SD ...)
  - ok for bar charts
- SPSS
- Minitab
- SAS
- R — free version of S+
  - ↳ windows version
  - ↳ Linux version (on our machines in CS)
- Statistix