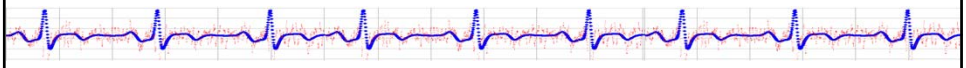
Empirical Research Methods in Information Science

IS 4800 / CS 6350

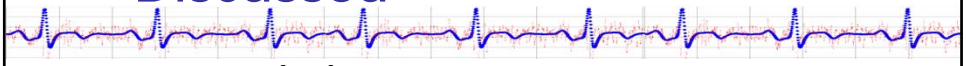


Lecture 21

Within-Subjects Designs

1

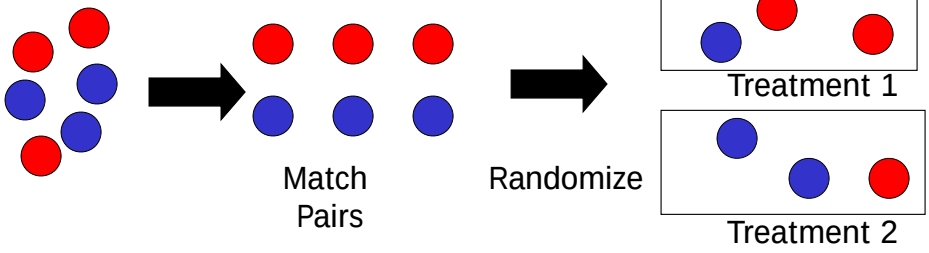
Types of Studies We've Discussed



- Descriptive
- Correlational
- Demonstration
- Experimental
  - One-factor, two-level, between-subjects
  - One-factor, two-level, within-subjects
    - aka "repeated measures" or "crossover"
  - Matched pairs

2

Review: Matched Group Design



- Use when you know some extraneous inter-subject variable has significant correlation with DV
- A between-subjects design

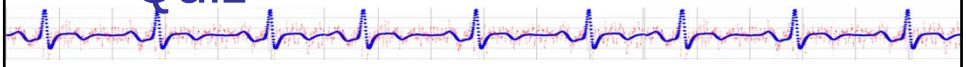
3

Review: Aaron & Aaron

- One-factor, two-level, within-subjects
  - Use “paired samples t-test” aka “t-test for dependent means”
- Matched pairs
  - Use “paired samples t-test” aka “t-test for dependent means”

4

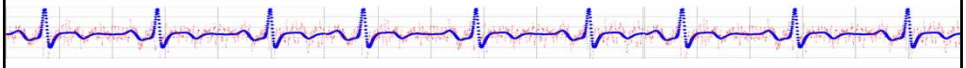
Quiz



- 15 minutes
- Closed book, closed notes

5

Types of Experimental Designs



- *Between-Subjects Design*
  - Different groups of subjects are randomly assigned to the levels of your independent variable
  - Data are averaged for analysis
  - Use t-test for independent means
  - We have discussed “single factor, two-level, between subjects” designs.

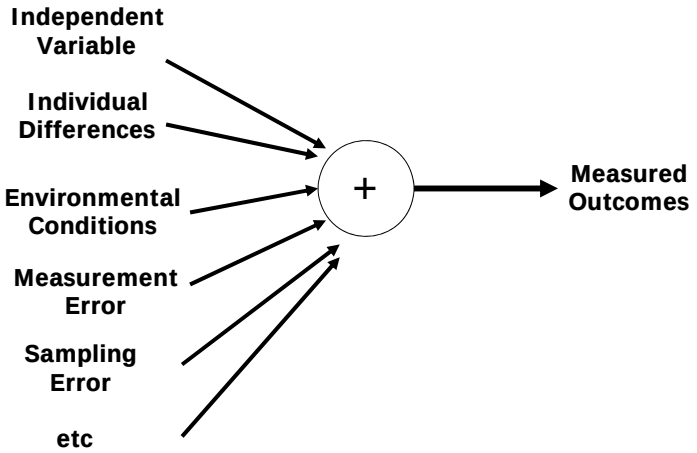
7

Types of Experimental Designs

- *Within-Subjects Design*
  - A single group of subjects is exposed to all levels of the independent variable
  - Data are averaged for analysis
  - aka “repeated measures design”, “crossover design”
  - Use t-test for dependent means aka “paired samples t-test”
  - We will discuss “single factor, two-level, within subjects” designs.

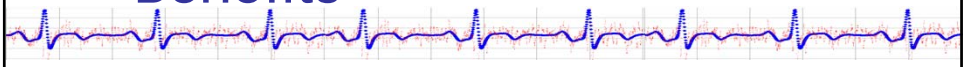
8

Error Variance



9

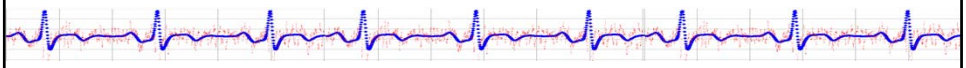
Within-Subjects Designs Benefits



- More Power! *Why?*
  - Controls for <u>all</u> inter-subject variability
  - Randomized between-subjects design just balances the effects between groups
  - (Matched-pair controls for identified and matched extraneous variables)
- Subjects can be asked to directly compare treatments

10

Within-Subjects Designs Disadvantages



- More demanding on subjects, especially in complex designs
- Subject attrition is a problem
- *Carryover effects:* Exposure to a previous treatment affects performance in a subsequent treatment

15

Carryover Example

- Embodied Conversational Agents to Promote Health Literacy for Older Adults

Brochure Computer

T0 → T1 → T2

Diabetes Knowledge Assessment	Diabetes Knowledge Assessment	Diabetes Knowledge Assessment

16

Sources of Carryover

- *Learning*
  - Learning a task in the first treatment may affect performance in the second
- *Fatigue*
  - Fatigue from earlier treatments may affect performance in later treatments
- *Habituation*
  - Repeated exposure to a stimulus may lead to unresponsiveness to that stimulus
- *Sensitization*
  - Exposure to a stimulus may make a subject respond more strongly to another
- *Contrast*
  - Subjects may compare treatments, which may affect behavior
- *Adaptation*
  - If a subject undergoes adaptation (e.g., dark adaptation), then earlier results may differ from later ones

17

Dealing With Carryover Effects

- *Counterbalancing*
  - The various treatments are presented in a different order for different subjects
  - May be complete or partial
  - Balances the effects of carryover on each treatment
  - Assumes carryover effect is independent of the order
- By randomizing treatment order you balance the influence of all time-related extraneous variables across conditions

19

Counterbalancing

- Full
  - N! treatment orderings
- Partial
  - Randomly select <N! treatment orderings
- Other partial counterbalancing strategies
  - Latin Square

20

Counterbalancing

- *The Latin Square Design*
  - Sample partial counterbalancing approach
  - Used when you make the number of treatment orders equal to the number of treatments (each treatment occurs once in every row and column)
  - Example: want to evaluate 4 different word processors, using 4 admins in 4 departments. A completely counterbalanced design would require 4x4x4=64 trials.
  - Latin square attempts to eliminate systematic bias in assignment of treatment to departments & subjects.

Subj	Department				
	1	2	3	4	
1	C	B	A	D	Treatments A-D
2	B	A	D	C	
3	D	C	B	A	
4	A	D	C	B	

21

Dealing With Carryover Effects

- Taking Steps to Minimize Carryover
  - Techniques such as pre-training, practice sessions, or rest periods between treatments can reduce some forms of carryover
- Make Treatment Order an Independent Variable
  - Allows you to measure the size of carryover effects, which can be taken into account in future experiments

22

Example of a Counterbalanced Single-Factor Design With Two Treatments

Order	Treatment Sequence
1	A B
2	B A

Subject	Order	Treatment A	Treatment B
1	2	23.5	14.2
2	1	14.6	11.5
...	...	...	...

How do you test for "order effects"?
2x2 ANOVA (later)

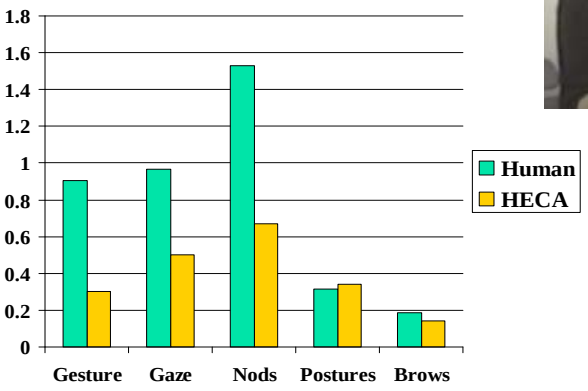
23

Design Criteria

- Between-subjects
  - Default
  - No carryover issues, but may require <u>many</u> subjects
- Within-subjects
  - More power, fewer subjects
  - Sensitive to carryover effects, requires more subject time
  - Allows direct comparison of treatments by subjects
- Matched-pairs
  - Suspect extraneous inter-subject variable highly correlated with DV
  - BUT, anticipate large carryover effect
- Other issues (e.g., recruiting) may be determining factor

24

Example Study: Handheld ECAs



Modality Study

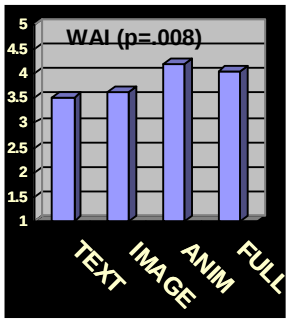
- Compared 4 modalities:
  - Text only
  - Text + Static agent image
  - Animated agent
  - Animated agent + nonverbal sounds
    - Backchannels, Discourse markers, etc.
- DVs: WAI, Credibility, Comfort

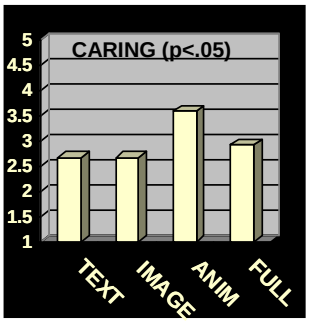
Modality Study

- 4 treatments
- How to design?

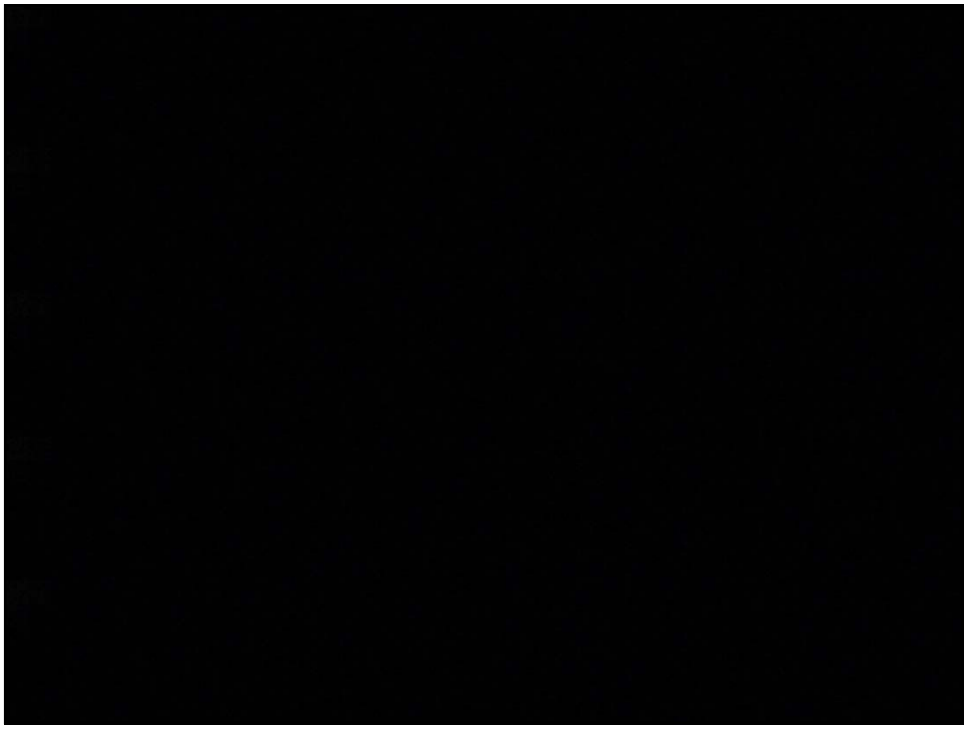

- 4 characters
- 4 topics
- Counterbalance treatment order AND randomize pairing of character/topic/treatment


Modality Study



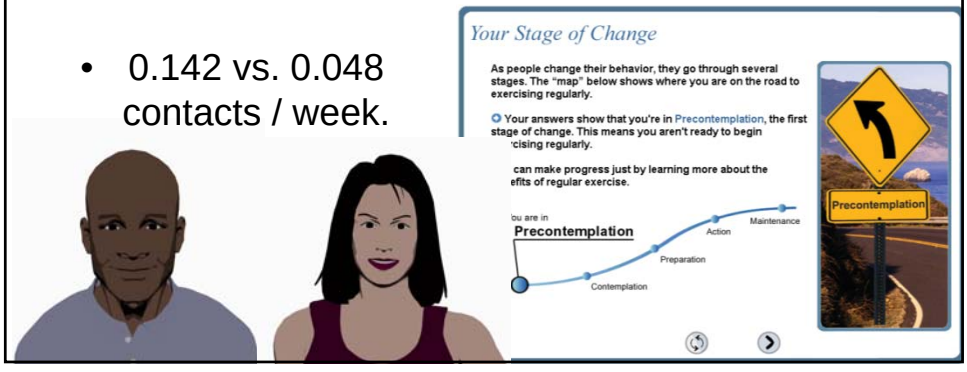


- Animated agent also scored higher (approaching significance) on *credibility of health information* and *comfort using in the workplace.*



RAISE: Web-based Intervention with and without Agent

- Compare: 1) existing web-based intervention to 2) intervention + agent to 3) control (N=1,200).
- 1 year intervention + 1 year followup
  - 0.142 vs. 0.048 contacts / week.



Example – Best Design?

- You've just developed the "Matchmaker" – a handheld device that beeps when you are in the vicinity of a compatible person who is also carrying a Matchmaker.
- You evaluate the number of users who are married after six months of use compared to a non-intervention control group.

31

Example – Best Design?

- You've just developed "Reado Speedo" that reads print books using OCR and speaks them to you at twice your normal reading rate. You want to evaluate your product against the old fashioned way on reading rate, comprehension and satisfaction.

32

Example – Best Design?

- You've developed a new web-based help system for your email client. You want to compare your system to the old printed manual.

33

Example – Best Design?

- You are evaluating a new customer support ticketing system and want to handle some customer calls with the new system to compare it to the old one.

34

Example – Best Design?

- Want to evaluate skype instead of face-to-face for sales calls among your international B2B salesforce
- 10x productivity difference among salespeople
- A salesperson makes 1-2 sales calls per month

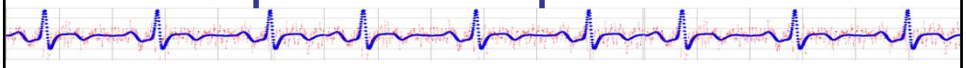
35

Study of Novice Programmers using Eclipse & Gild

- Critique?

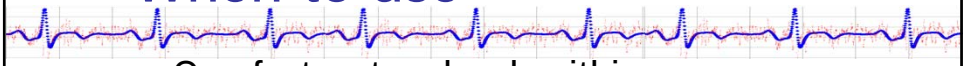
36

t-test for dependent means
aka “paired sample t-test”



37

t-test for dependent means
When to use



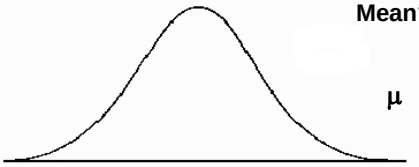
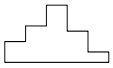
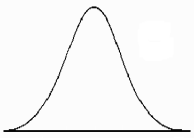
- One factor, two-level, within-subjects/repeated measures design

  -or-

- One factor, two-level, between-subjects, matched pair design
- *In general, a bivariate categorical IV and numeric DV when the DV scores are highly correlated.*
- Assumes
  - Population distribution of individual scores is normal

38

Review - Sampling

		Mean?	Variance?
Population		μ	σ^2
Sample of size N		$M = \frac{\sum X}{N}$	$SD^2 = \frac{\sum (X - M)^2}{N}$
Mean values from all possible samples of size N aka "distribution of means"		μ	$\sigma_M^2 = \frac{\sigma^2}{N}$

39

Wanted: a statistic for differences between paired individuals

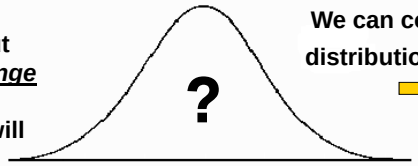
- In a repeated-measures or matched-pair design, you directly compare one subject with him/herself or another specific subject (not groups to groups).
- So, start with a sample of change (difference) scores:

Sample 1 = Mary's wpm using Wizziword – Mary's wpm using Word

42

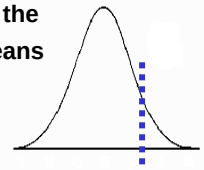
Hypothesis testing with paired samples

Given info about population *of change scores* and the sample size we will be using (N)



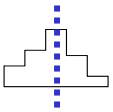
$\mu = 0$
est σ^2 from sample

We can compute the distribution of means



and finally determine the probability that this mean occurred by chance

Now, given a particular sample *of change scores* of size N



We compute its mean

$$t = \frac{M}{S_M}$$

df = N-1

Effect Size & Power

- d = M / S
  - M = mean of difference scores
  - S = std dev of population of individual's difference scores
  - 0.2 = small; 0.5 = medium; 0.8 = large

N required for 80% power, two-tailed, α=.05:

Effect Size		
Small (d=0.2)	**Medium (d=0.5)**	**Large (d=0.8)**
196	**33**	**14**

For Comparison
Power calcs for t-test for indep means

TABLE 8–5 Approximate Number of Participants Needed in Each Group (Assuming Equal Sample Sizes) for 80% Power for the *t* Test for Independent Means, Testing Hypotheses at the .05 Significance Level

	Effect Size		
	Small (.20)	Medium (.50)	Large (.80)
One-tailed	310	50	20
Two-tailed	393	64	26

46

R

Subject	Condition1	Condition2	var
1	104	212	
2	210	415	
3	157	127	
4	321	302	
5	98	309	
6	129	742	
7	205	489	
8	137	425	
9	291	321	
10	91	81	

47

R

#two-sided by default

- t.test(d$Condition1, d$Condition2, **paired=TRUE**)

Paired t-test

data: d$Condition1 and d$Condition2
t = 4.1849, df = 16, p-value = 0.0003501
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.233975   Inf
...

paired t(16)=4.18, p<.05

48

Effect Size

```
> diffs <- abs(d$Condition1 - d$Condition2)

#Effect size
> mean(diffs)/sd(diffs)
```

49

Issues with t-test for dependent means

- Use with care on longitudinal data
  - E.g, pre-post design with no control
  - Significant differences (changes) may have been due to something other than your intervention
  - Essentially a demonstration
  - Better to use between-subjects design with control group for comparing changes over longitudinal studies

50

Group Exercise

- For each problem, write
  1. Kind of study design
  2. Two populations being compared
  3. Research & Null hypotheses
     1. English & In terms of Pop means
  4. Test criteria
  5. Test results
     - English & Publication format (requires df)
  6. Determine Effect Size
  7. Determine the number of subjects you would need to do a study with similar effect size in the future

51

Exercise

- You are head of QA at Tomorrow's Technology Tonight. The engineers there have just designed a new PC input device that lets users move their cursor using nasal sounds. You are put in charge of evaluating their claim that this is a faster way of controlling a PC than other conventional methods. You run a series of standardized tests with random users measuring average time click on a screen target (in ms), with a standard mouse and with the nasal sound device.

52

Group Exercise

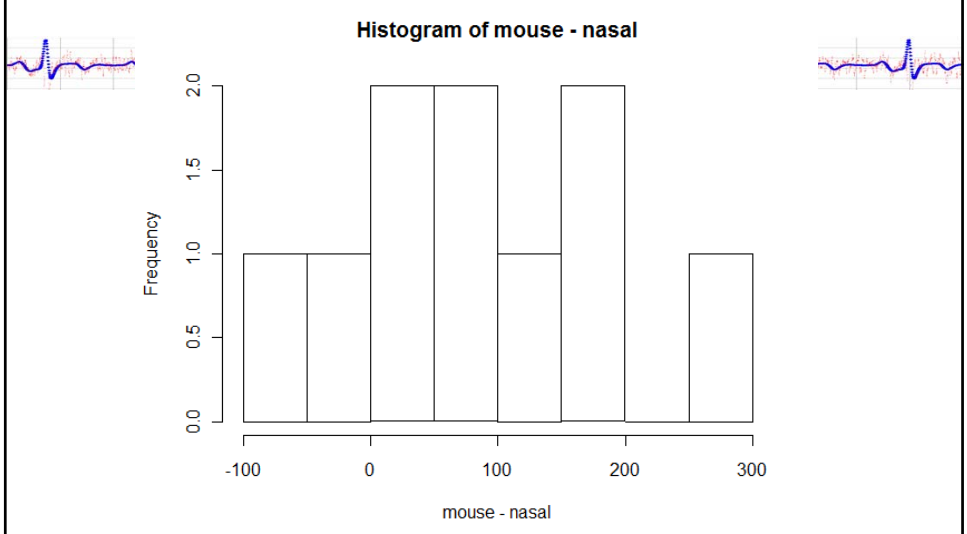
- For each problem, write
  1. Kind of study design
  2. Two populations being compared
  3. Research & Null hypotheses
     1. English & In terms of Pop means
  4. Test criteria
  5. Test results
     - English & Publication format (requires df)
  6. Determine Effect Size
  7. Determine the number of subjects you would need to do a study with similar effect size in the future

53

Data

Subject	Mouse	Nasal	Difference
1	104	52	52
2	210	47	163
3	157	104	53
4	321	63	258
5	98	114	-16
6	129	89	40
7	205	32	173
8	137	191	-54
9	291	142	149
10	91	81	10

54



55

Descriptives?

```
> mean(mouse)
[1] 174.3
> mean(nasal)
[1] 91.5
> sd(mouse)
[1] 80.83048
> sd(nasal)
[1] 48.44527
```

56

Test

```
> t.test(nasal,mouse,paired=TRUE)

      Paired t-test

data:  nasal and mouse
t = -2.6618, df = 9, p-value = 0.02597
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
 -153.16953  -12.43047
sample estimates:
mean of the differences
                  -82.8
```

57

Group Exercise

- For each problem, write
  1. Kind of study design
  2. Two populations being compared
  3. Research & Null hypotheses
     1. English & In terms of Pop means
  4. Test criteria
  5. Test results
     - English & Publication format (requires df)
  6. Determine Effect Size
  7. Determine the number of subjects you would need to do a study with similar effect size in the future

58

Effect Size

```
> mean(mouse - nasal) / sd(mouse - nasal)
[1] 0.8417215
```

- 0.2 = small; 0.5 = medium; 0.8 = large

59

Effect Size & Power

- d = M / S
  - M = mean of difference scores
  - S = std dev of population of individual's difference scores
  - 0.2 = small; 0.5 = medium; 0.8 = large

N required for 80% power, two-tailed, α=.05:

Effect Size		
Small (d=0.2)	**Medium (d=0.5)**	**Large (d=0.8)**
196	**33**	**14**

60

Homework

- Multi-factor between-subjects designs & ANOVA
- Read:
  - B&A Ch 10 from 317
  - B&A Ch 14 442-453
  - Aron Ch 9, [opt Ch 10]
  - Example: Jonsson paper
- T2 & Presentations 2 on 4/2

61