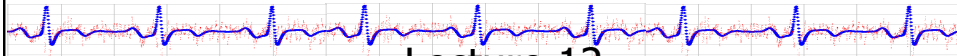


Empirical Research Methods in Information Science

IS 4800 / CS6350

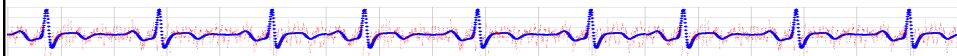


Lecture 12

Between Subjects Experimental Designs
The Normal Curve
The Single-Observation Test
Randomization & Control Groups

1

But first...



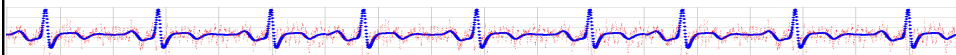
Wrapup Correlation!

2

Example: Leashes & Attachment



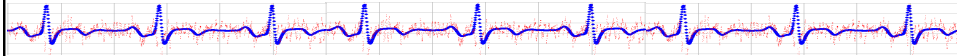
Example: Leashes & Attachment



- You want to see if toddlers who grow up leashed have better attachment scores.
- You recruit 30 parents of toddlers, and randomly give half of them leashes and sign contracts agreeing to leash their toddler every time they leave the house.
- After one year you administer the strange situation protocol to classify the toddler attachment as secure, avoidant, or resistant.
- What kind of study is this?
- What statistic would you use to evaluate results?
- What is df ?
- Assuming $X^2(df) = 32.4$, what would you conclude?
- Assuming $X^2(df)=0.2$, what would you conclude?

4

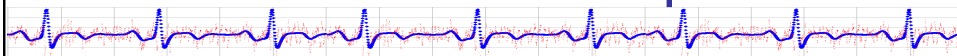
Example: Net Latency & Satisfaction



- NU ITS wants to save money by switching to slower wireless routers, and wants to assess the impact this will have on student satisfaction. You want to know how slow things have to get before students start complaining.
- You have ITS implement a program that randomly chooses a network latency (between 0s and 10s) every time a student logs into NUwave, then adds that latency to every network access from them. After 10 minutes of use a web form pops up asking students to rate their degree of satisfaction with NU ITS (10-item).
- What kind of study is this?
- What statistic would you use to evaluate results?
- Assuming $r = -0.8$, $p = .021$, what would you conclude?
- Assuming $r = 0.1$, $p = .342$, what would you conclude?

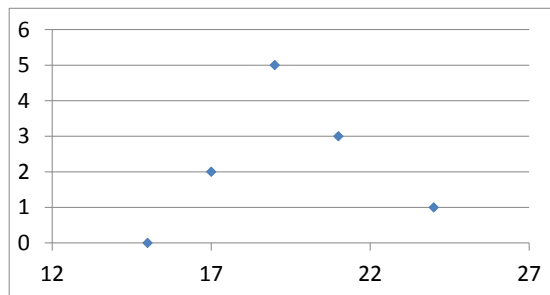


What do you do if your data is clearly not unimodal & symmetric OR there is a clear non-linear relationship?



Sample Data for Spearman rho

| | | | | | |
|--------------|----|----|----|----|----|
| Monitor Size | 21 | 24 | 17 | 19 | 15 |
| Productivity | 3 | 1 | 2 | 5 | 0 |

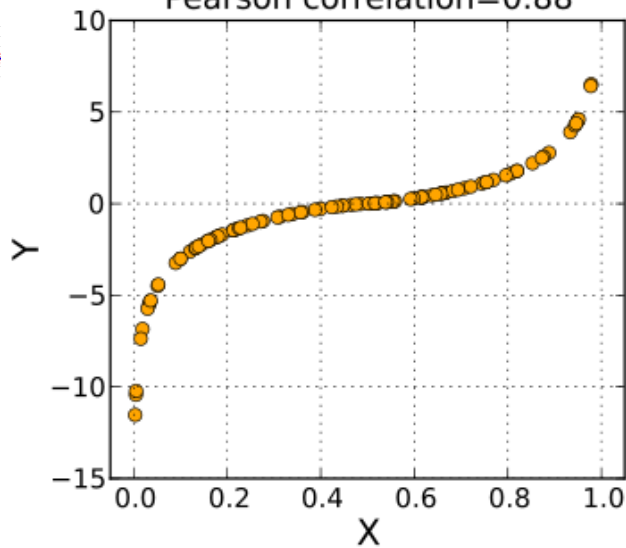


$\rho = 0.3$

$*r = 0.18$

7

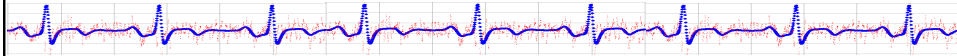
Spearman correlation=1
Pearson correlation=0.88



Spearman
measures
degree of
monotonicity.

8

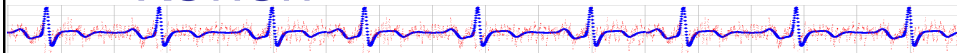
Parametric vs. Non-parametric Statistics



- non-parametric statistics that do not rely on data belonging to any particular distribution
- E.g., Pearson r is a parametric statistic (assumes underlying distributions are normal – can be described using parameters – mean & stddev)
- E.g., Spearman ρ is non-parametric

9

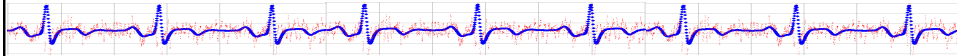
Review



- Kinds of analyses for Descriptive studies?
 - Descriptive
 - Chi-square goodness of fit [nominal]

11

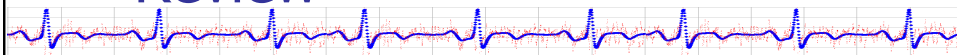
Review



- Kinds of analyses for Correlational studies?
 - Descriptive
 - Chi-square goodness of fit [nominal]
 - Chi-square test for independence [nominal/nominal]
 - Correlation [numeric/numeric]
 - Point-biserial [numeric/nominal]
 - Spearman rho [ordinal/ordinal], [ordinal,numeric], or [numeric,numeric and not linear or not normal]

12

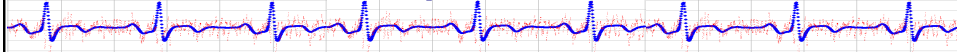
Review



- Kinds of analyses for Experimental studies?
 - Descriptive
 - Chi-square goodness of fit [nominal]
 - Correlation (large number of IV values - *parametric*) [numeric/numeric] (atypical)
 - Chi-square test for independence [nominal/nominal]
 - t-test! [nominal/numeric]

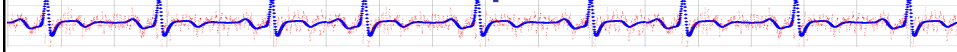
13

t-test concepts



14

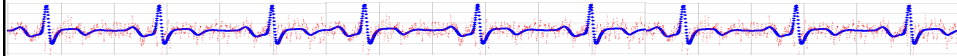
t-test for independent means



- Tests association between binomial IV and numeric DV.
- Examples:
 - WizziWord vs. Word => wpm
 - Small vs. Large Monitors => wpd
 - Wait time sign vs. none => satisfaction

15

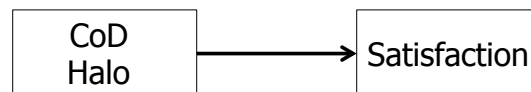
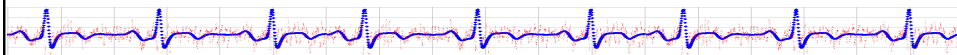
Understanding numeric measures



- Sources of variance
 - IV
 - Other uncontrolled factors ("error variance")

16

Example: Call of Duty vs. Halo



- What variables might affect Satisfaction?
- Typically, one subject's Satisfaction score = **TrueSatisfaction** + var1 + var2 + var3 + ...
- A sum of random variables.

17

Central Limit Theorem

- If (many) independent, random variables with the same distribution are added, the result is approximately a normal curve

18

Why be normal?

Example

- Suppose random variable X has distribution

$$X = \begin{cases} 1 & \text{with probability } 1/3, \\ 2 & \text{with probability } 1/3, \\ 3 & \text{with probability } 1/3. \end{cases}$$

| | | |
|-------|---|---|
| o | o | o |
| ----- | | |
| 1 | 2 | 3 |

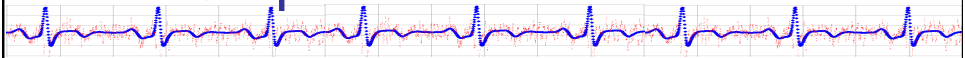


From wikipedia

19

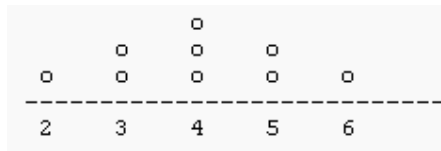
Why be normal?

Example



■ Now, consider the distribution of $X+X$

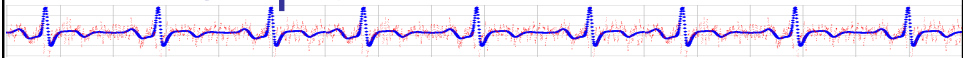
$$\left\{ \begin{array}{l} 1+1 = 2 \\ 1+2 = 3 \\ 1+3 = 4 \\ 2+1 = 3 \\ 2+2 = 4 \\ 2+3 = 5 \\ 3+1 = 4 \\ 3+2 = 5 \\ 3+3 = 6 \end{array} \right\} = \left\{ \begin{array}{l} 2 \text{ with probability } 1/9 \\ 3 \text{ with probability } 2/9 \\ 4 \text{ with probability } 3/9 \\ 5 \text{ with probability } 2/9 \\ 6 \text{ with probability } 1/9 \end{array} \right\}$$



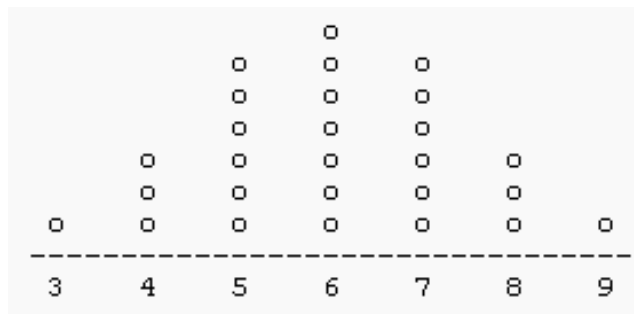
20

Why be normal?

Example

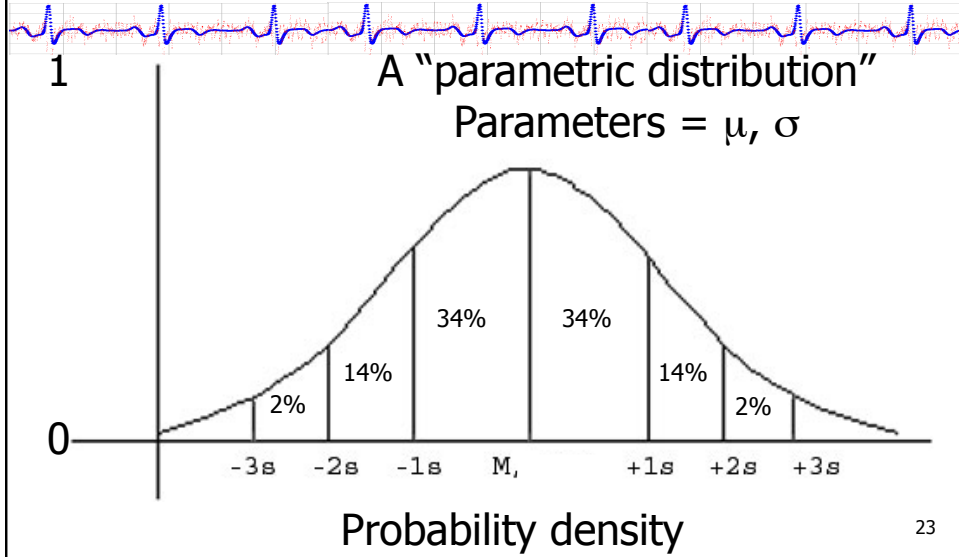


■ Now, consider the distribution of $X+X+X$

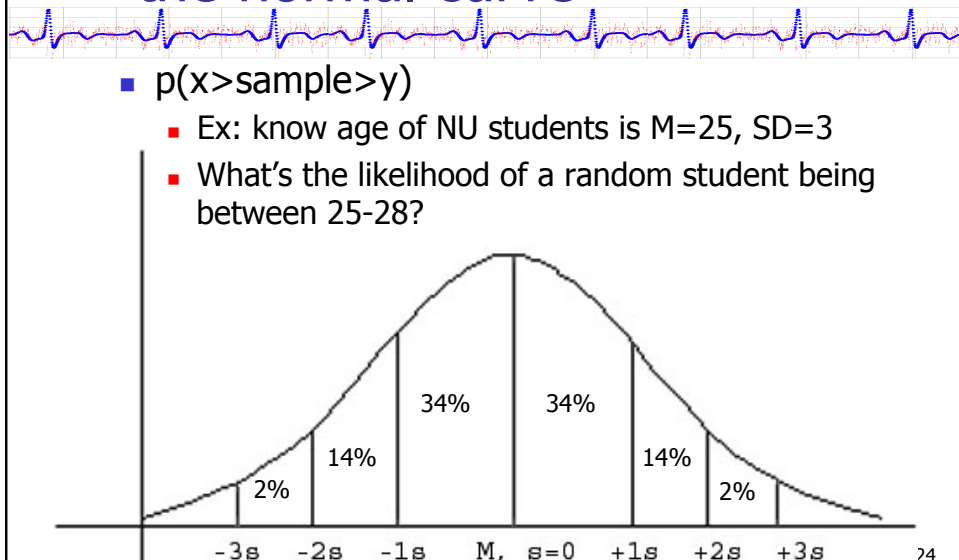


21

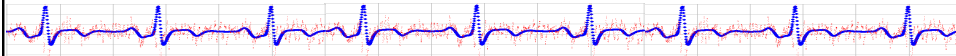
The Normal Curve



Estimating probabilities using the normal curve

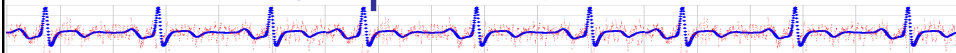


Given what we know so far..
We can already do some
hypothesis testing!



25

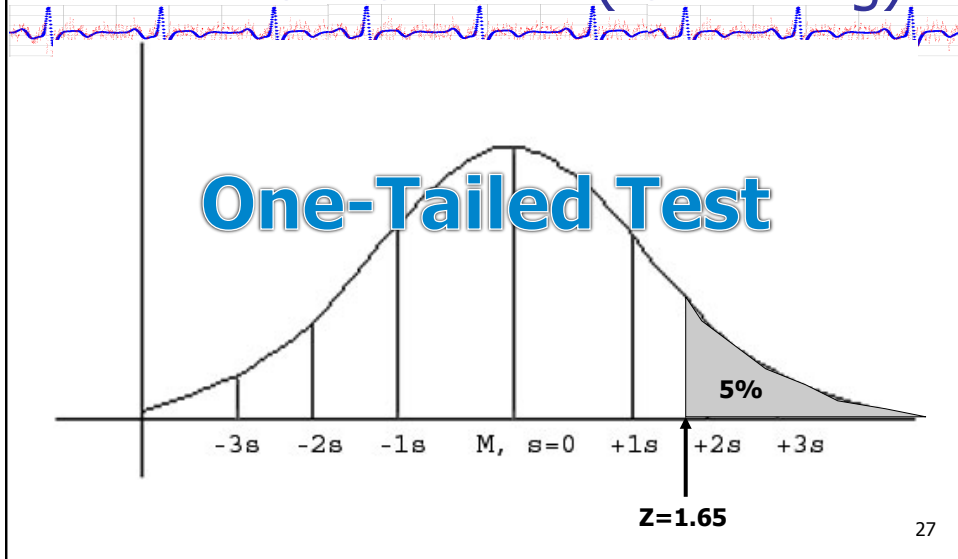
Example: single sample test with sample size = 1



- All your admins use Word
- You measure their typing speed, find it is roughly normal, with
 - $\mu = 150$
 - $\sigma = 25$
- You interview a new admin candidate, his typing speed is 200 wpm
- Is he significantly different than your population of admins?

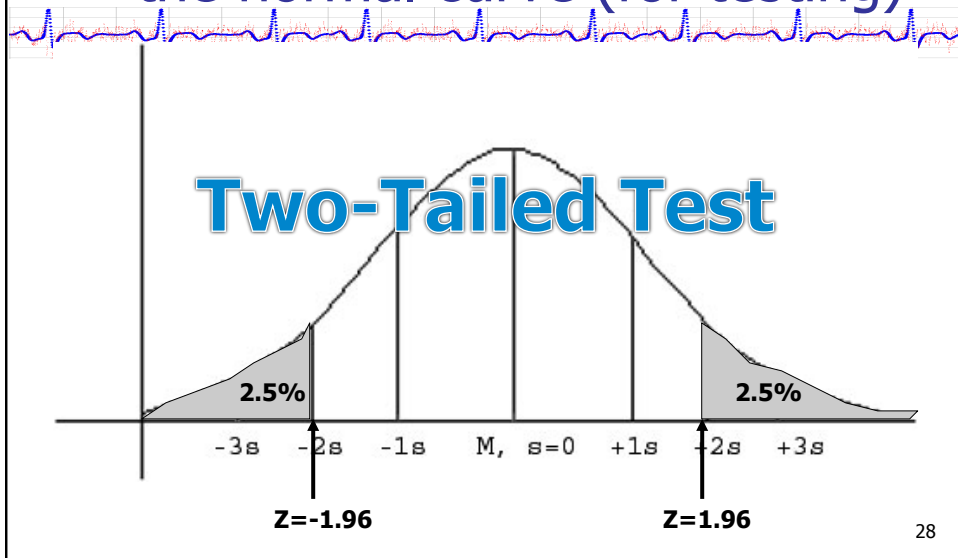
26

The most important parts of the normal curve (for testing)



27

The most important parts of the normal curve (for testing)



28

Hypothesis testing – one tailed

- Hypothesis: sample (of 1) will be significantly greater than known population
 - Population completely known (not an estimate)
- Example – WizziWord experiment:
 - H1: $\mu_{\text{WizziWord}} > \mu_{\text{Word}}$
 - Test criteria: $\alpha = 0.05$, one-tailed
 - Population (Word users): $\mu_{\text{Word}} = 150$, $\sigma = 25$
 - What level of performance do we need to see before we can accept H1?

$$(S-150)/25 \geq 1.65 \quad S = 191.25$$

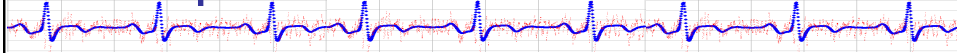
29

Hypothesis testing – two tailed

- Hypothesis: sample (of 1) will be significantly different from known population distribution
- Example – WizziWord experiment:
 - H1: $\mu_{\text{WizziWord}} \neq \mu_{\text{Word}}$
 - $\alpha = 0.05$ (two-tailed)
 - Population (Word users): $\mu_{\text{Word}} = 150$, $\sigma = 25$
 - What level of performance do we need to see before we can accept H1?

$$|(S-150)/25| \geq 1.96 \quad S \leq 101 \text{ OR } S \geq 199$$

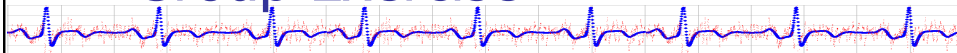
Standard testing criteria for experiments



- $\alpha = 0.05$
- Two-tailed *why?*

31

Group Exercise

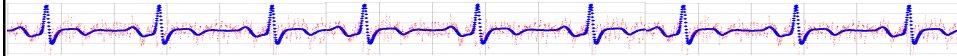


- For each problem, write
 1. What kind of study design is it?
 2. Two populations being compared
 3. Research & Null hypotheses in English
 4. Research & Null hypotheses in terms of Pop means
 5. Test criteria
 6. Test results
 - English
- Critique the study design

32

R

'norm' is the normal distribution



```
#by default, parameters are Z scores
dnorm          #probability density funct
dnorm(1.0)     #height of normal curve at Z=1.0
dnorm(1.0,mean=1.2,sd=0.5)

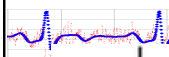
pnorm          #cumulative density funct
pnorm(1.0)     #area under curve for Z<=1.0

qnorm          #critical values (cutoffs)
qnorm(.95)    #the Z score s.t. area below=.95
```

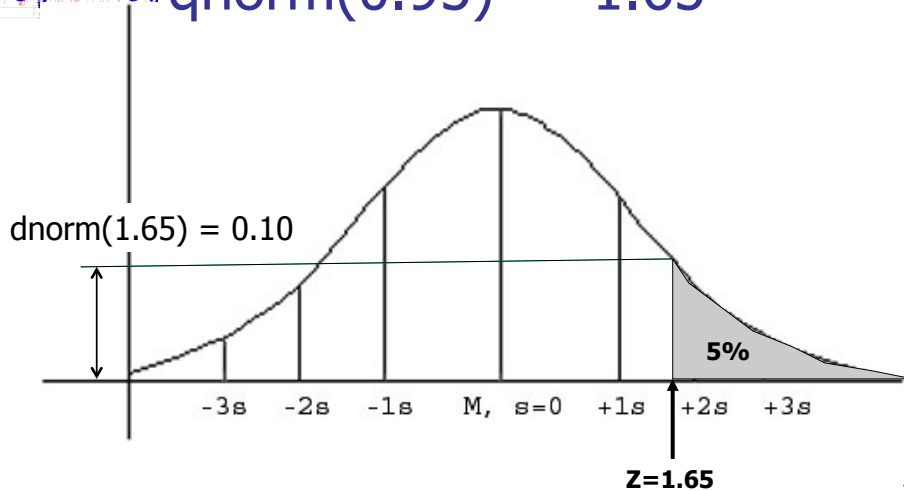
33

Example:

$$\text{pnorm}(1.65) = 0.95$$



$$\text{qnorm}(0.95) = 1.65$$



34

Don't try this at home

- You would never do a study this way.
- Why?
 - Can't control extraneous variables through randomization.
 - Usually don't know population statistics.
 - Can't generalize from an individual.

35

Two Group Between-Subjects Experimental Design

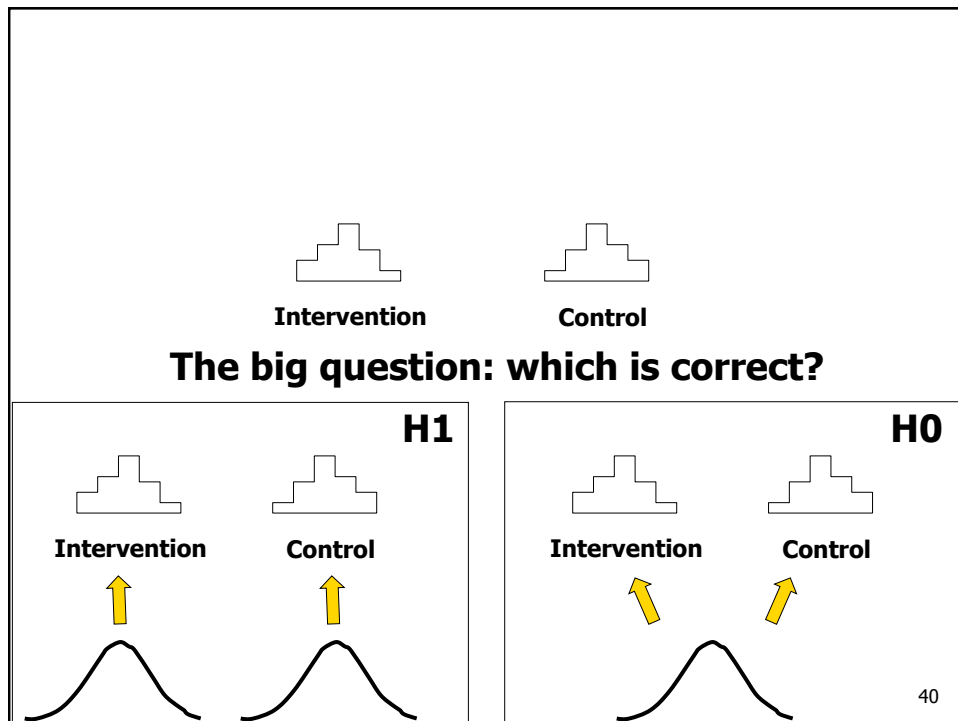
- B&A: "Randomized Two Group Design"
- Have two experimental conditions (treatments, levels, groups)
- Randomly assign subjects to conditions
 - Each subjects sees one condition
- Measure (numeric) outcome in each group

38

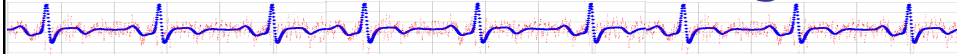
Between-Subjects Design

- Each group is a **sample** from a population
- Big question: are the populations the same (null hypothesis) or are they significantly different?

39

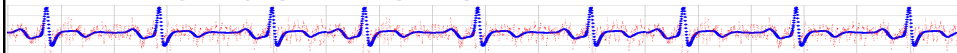


Hold that thought...
More next time on testing this!

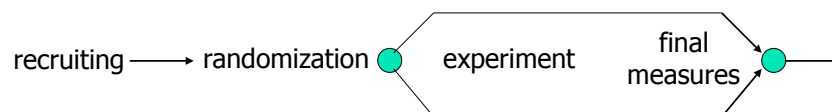


41

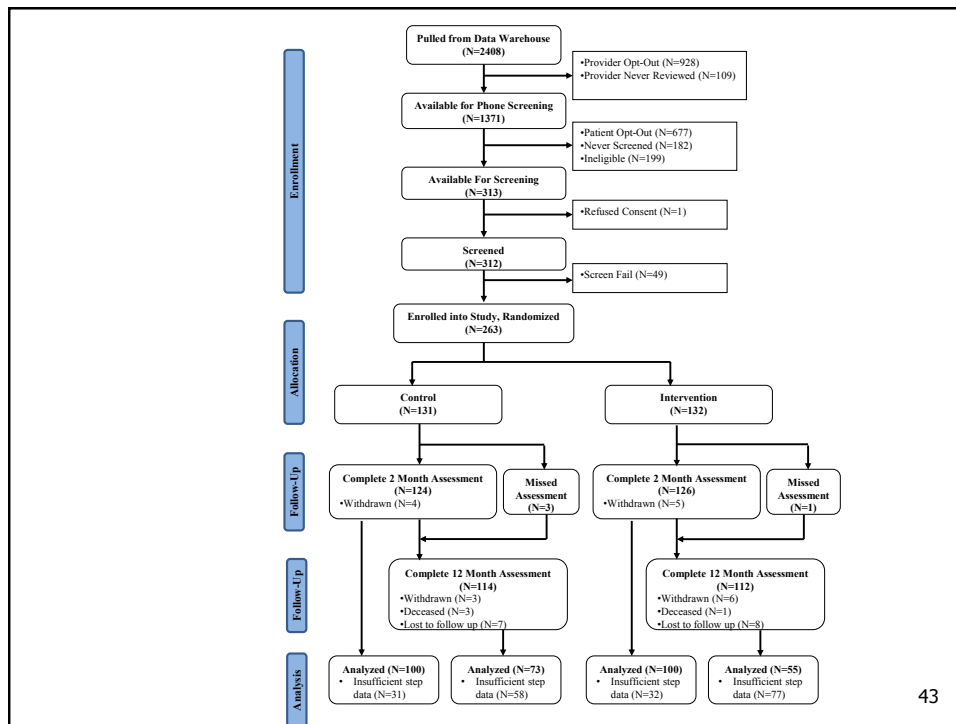
Sidebar: Randomization



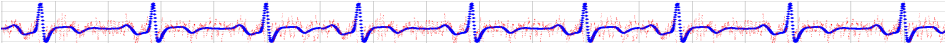
- Crucial: method must not be applied subjectively
- Point in time at which randomization occurs is important



42



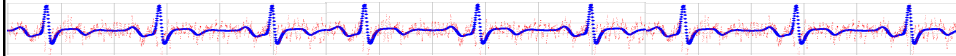
Intent-to-Treat



- You want to test a new support line ticket system.
- You randomize 20 support employees to use the new system, 20 to use the old one, then collect satisfaction and performance measures after one month.
- You discover that 6 of the employees using the new system stopped using it after a week.
- What do you do?

44

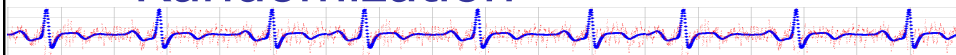
Intent-to-Treat



- Once a subject is randomized, every effort is made to include their outcome measures (DV) in the analysis
 - Even if they did not use the Intervention
 - Even if they went on vacation
 - Even if they died ...
 - Assume worst case for lost data (e.g., intervention did not work)
- Efficacy = IV/DV effect under ideal conditions (e.g., lab study) = "method effectiveness"
- Effectiveness (aka "use effectiveness") = IV/DV effect under real world conditions
- Intent-to-treat assesses "effectiveness"

45

Sidebar: Randomization

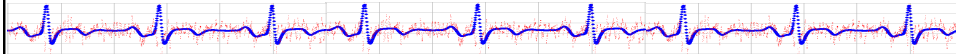


- Simple randomization
 - Flip a coin
 - Random number generator
 - Table of random numbers
 - Partition numeric range into number of conditions

- Problems?

46

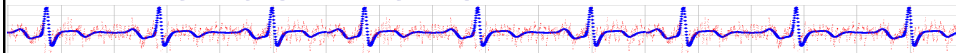
Sidebar: Randomization



- **Blocked randomization**
 - Avoids serious imbalances in assignments of subjects to conditions
 - Guarantees that imbalance will never be larger than a specified amount
 - Example: want to ensure that every 4 subjects we have an equal number assigned to each of 2 conditions => "block size of 4"
 - Method: write all permutations of N conditions taken B at a time (for B = block size)
 - Example: 1122, 1212, 2112, 2121, 2211, 1221
 - At the start of each block, select one of the orderings at random
 - Should use block size > 2, block size = multiple of # arms

47

Sidebar: Randomization



- **Stratified randomization**
 - First stratify Ss based on measured factors (prior to randomization) (e.g., gender)
 - Within each strata, randomize
 - Either simple or blocked

| <u>Strata</u> | <u>Sex</u> | <u>Condition assignment</u> |
|---------------|------------|-----------------------------|
| 1 | M | ABBA BABA... |
| 2 | F | BABA BBAA... |

48

Sidebar: Control groups

- A controlled experiment (“experimental design”) generally compares the results obtained from an experimental sample against a control sample, which is identical to the experimental sample except for the one aspect whose effect is being tested.
- You must carefully select your control group in order to demonstrate that only the IV of interest is changing between groups.
- The control group must also comprise a reasonable comparison.

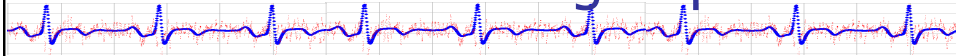
49

Control Groups: Example

- Say you are developing a conversational agent that counsels college students with depression (using CBT) and co-morbid binge drinking (using BMI).
- What is a good control group?



Sidebar: Control groups

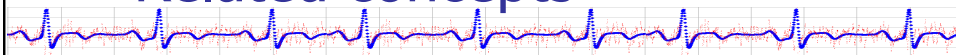


- Standard-of-care control (new vs. old)
- Non-intervention control
- "A vs. B" design (shootout)
- "A vs. A+B" design (e.g., S-O-C vs. S-O-C+intervention)

- Problem: the "intervention" may cause more than just the desired effect
 - Example: giving more attention to intervention Ss in educational intervention
- Some solutions:
 - Attention control
 - Placebo control
 - Wait list control (also addresses measurement issues)

51

Sidebar: Control groups Related concepts



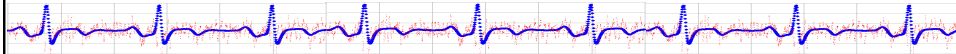
- Blind test – experimenter does not know group
- Double blind test – neither S nor experimenter know

- Manipulation check
 - Test performed just to see if your manipulation is working. Necessary if immediate effect of manipulation is not obvious.
 - "Positive control" test for intervention effect
 - "Negative control" test for lack of intervention effect
 - Example:
 - Student Center Sign: ask students if they saw & read the sign

- Contamination
 - Some control subjects get some of the intervention

52

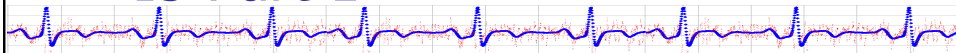
Homework



- Read B&A Ch 14 to 447 (t-test)
- Remedial stats:
 - Distributions of means
 - t distribution
 - Single sample t-test
 - t-test for independent means
- HW I5 (IS4800 only) – due 2/27
 - Work individually on this one.
 - Start Part I Now – Write a complete study proposal

53

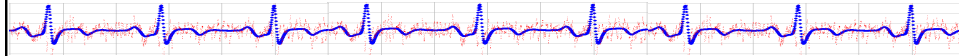
I5 Part 1



- Write a research plan for conducting an experiment comparing WizziWord vs. CoolText word processors using admins from BigBucks, Inc (both of these are new products).
- Outcome measures to include productivity (words per day output during the 8th week after the new word processors are introduced), and satisfaction, using the ILoveWordProcessors 12-item index (Cronbach alpha=0.82, test-retest correlation of 0.93, correlation with the standard 100-item WordProcessorsAreGreat index was 0.72).
- *Power Analysis (covered on 2/24)*: From studies at other sites you expect to see a difference in productivity of approximately 3,000 (SD 1,200) words per day between the products. Assume a 50% response rate to your recruitment ad, and a 85% retention rate for subjects.

54

I5 Part 1 - continued



- Be sure to include the following in your plan:
 - Hypotheses
 - Research model (the boxes and arrows diagram) and description of variables/measures
 - Human subjects issues, including eligibility criteria, recruitment procedures, and the number of potential subjects you need to reach with your recruitment ad.
 - Detailed protocol , including recruitment, sampling and randomization methods
 - Analysis plan
- Refer to sample research plan for inspiration.
- Your complete plan should be 2-3 pages long.