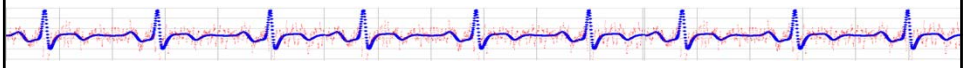
# Empirical Research Methods in Information Science

## IS 4800/CS 6350

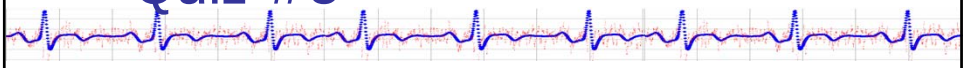


Lecture 15

Power and Effect Size

1

## Quiz #5



**1. What is true for a between-subjects experiment? Check all that apply.**

- ☐ Each subject sees all conditions of the study.
- ☐ Each subject sees only one condition of the study.
- ☐ You use Pearson correlation if the assumptions are met.
- ☐ You use $X^2$ test for independence if the assumptions are met.
- ☐ You use independent-samples t-test if assumptions are met.
- ☐ You use randomization to control extraneous variables.
- ☐ You use confounds to control extraneous variables.
- ☐ You are trying prove causality.
- ☐ You are trying to prove correlation.

2

## Quiz #5

**2. What is true of a distribution of means? Check all that apply (N=sample size, $\mu$=underlying population mean, $\mu_M$=mean of distribution of means, $\sigma^2$=underlying population variance, $\sigma_M^2$=variance of distribution of means).**

- ☐ It is the same as the underlying population distribution.
- ☐ It is based on the population distribution and the sample size.
- ☐ It has thicker tails than the population distribution (for N>1).
- ☐ It is skinnier than the population distribution (for N>1).

| | | |
|---|---|---|
| ☐ $\mu_M = \mu/N$ | ☐ $\mu_M = \mu$ | ☐ $\mu_M = \mu/\sqrt{N}$ |
| ☐ $\sigma_M^2 = \sigma^2/N$ | ☐ $\sigma_M^2 = \sigma^2$ | ☐ $\sigma_M^2 = \sigma^2/\sqrt{N}$ |

3

## t-test for independent means
### *What you need to know*

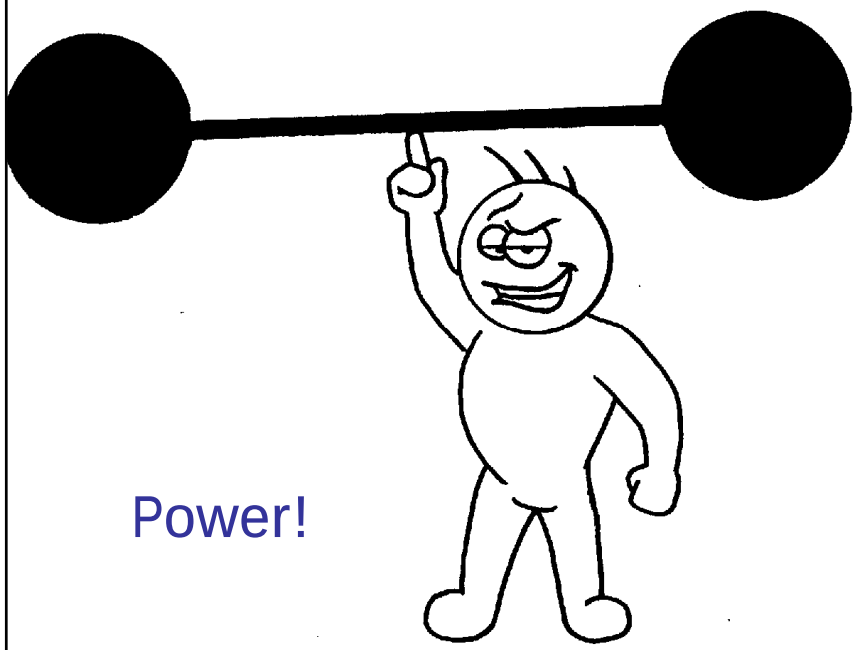
- When do you use it (kind of study design)?
- Assumptions?
- What is 't'?
- What is the 't distribution'? How is it parameterized?
- How do you interpret results?
- How do you report results?

4

## Some ‘formal’ terminology

- “Between-subjects, N factor, M level {univariate|multivariate}, experimental design with a {*yourcontrolhere*} control”

5



Power!

## Power

- The “power” of a statistical test is its ability to detect differences in data that are inconsistent with the null hypothesis.
  - p(rejecting H0|H1)
- aka – the ability to find a significant result, if your hypotheses are actually true.
- Why is doing an ‘underpowered’ test a bad idea?

7

## Effect size

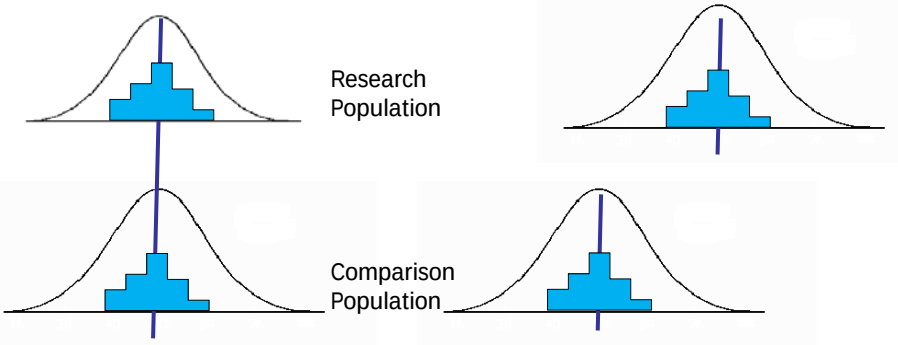
- The *amount* of measured difference between study conditions.
- The greater the effect size, the easier it is to show there is a significant difference in your study (ie, the greater the power).
- Calculation is different for each hypothesis test procedure.
- Tabulated standard values for “small”, “medium”, and “large” effect sizes.
- Only talk about effect size IF significance is established – but then DO present it in your results.

8

# The typical situation

H0 Actually True

H1 Actually True

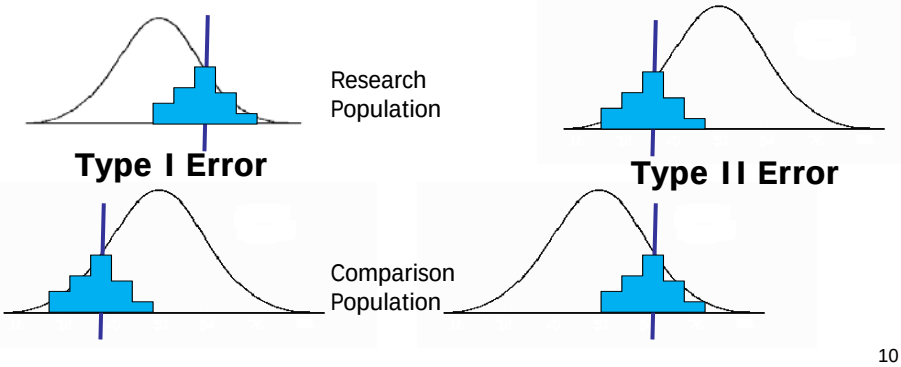



9

# The unlucky situations

H0 Actually True

H1 Actually True



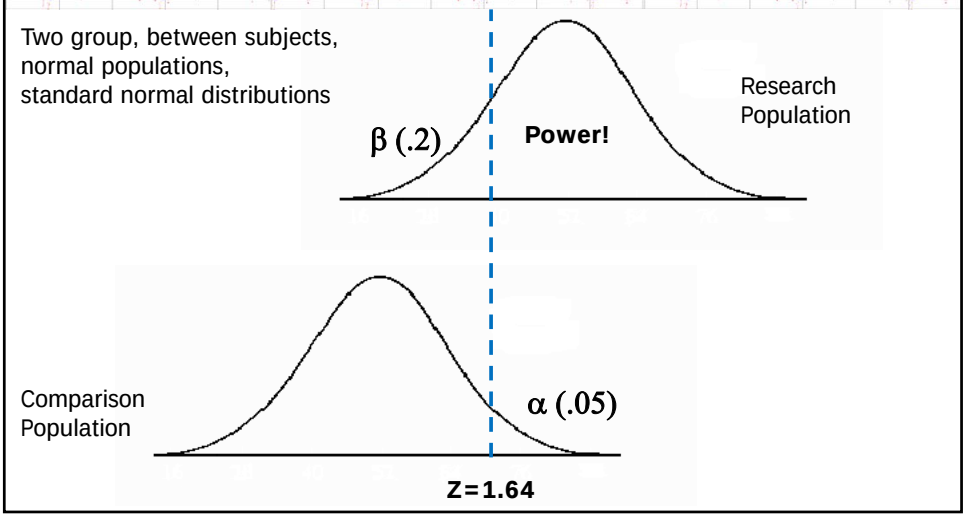

10

# Relationship between alpha, beta, and power.

| | "The Truth" H1 True | "The Truth" H1 False |
|---|---|---|
| **Decide to Reject H0 & accept H1** | Correct p = power | Type I err p = $\alpha$ |
| **Do not Reject H0 & do not accept H1** | Type II err p = $\beta$ | Correct p = 1-$\alpha$ |

11

# Relationship between power and effect size



## Power Analysis

- Power
  - Increases with effect size
  - Increases with sample size
  - Decreases with alpha
- Should determine number of subjects you need ahead of time by doing a 'power analysis'
- Standard procedure (part of your study plan):
  - Determine statistic you will use
  - Fix alpha and beta (1-power) (and number of tails if appropriate)
  - Estimate expected effect size from prior studies
  - Then: Determine number of subjects you need

14

## Power analyses are different depending on the statistical test you are using...

Rest of this lecture assumes we are discussing t-test for independent means.

15

# Effect Size

$$d = \frac{(\mu_1 - \mu_2)}{\sigma}$$

Parameters for population of individuals.
(so, use SD-pooled for t-test of indep means)

Cohen:
d~0.2 small
d~0.5 medium
d~0.8 large

16

# Power table

**TABLE 8–4** Approximate Power for Studies Using the *t* Test for Independent Means Testing Hypotheses at the .05 Significance Level

| Number of Participants in Each Group | Effect Size | | |
|---|---|---|---|
| | Small (.20) | Medium (.50) | Large (.80) |
| One-tailed test | | | |
| 10 | .11 | .29 | .53 |
| 20 | .15 | .46 | .80 |
| 30 | .19 | .61 | .92 |
| 40 | .22 | .72 | .97 |
| 50 | .26 | .80 | .99 |
| 100 | .41 | .97 | * |
| Two-tailed test | | | |
| 10 | .07 | .18 | .39 |
| 20 | .09 | .33 | .69 |
| 30 | .12 | .47 | .86 |
| 40 | .14 | .60 | .94 |
| 50 | .17 | .70 | .98 |
| 100 | .29 | .94 | * |

17

## More Useful and Concise
(for practical purposes use a power calculator)

**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 |

18

## Power Analysis Exercise

- Based on related research, we expect that there will be a medium effect size in our study of an LED sign in the Food Court affecting wait times.
- How many subjects do we need for a two-group, two-tailed test at $\alpha$=0.05, 80% power?

19

# More Useful and Concise
## (for practical purposes use a power calculator)

**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 |

20

# But, I can't study 786 subjects!

- Increase effect size
  - Increase difference in population means
  - Decrease population variance
  - *Hard to do in a principled way*
  - Redesign study to collect many trials of measures per subject
- Relax criteria for Type I error
  - Increase $\alpha$ threshold
  - Change from Two-tailed => one-tailed test
  - *Decreases credibility of your findings*
- Decrease power
  - *Decreases likelihood of getting a significant result*
- Use a different statistic
  - *If possible, maybe consult a statistician*
- Practically
  - usually, redesign experiment so that we have increased effect size
  - OR, call it a "pilot study"

21

## Interpreting results: Significance & Effect Size

- Significance
  - Just indicates that it is likely there is a non-zero difference between populations.
  - Says nothing about how big the difference is.
- Effect Size
  - Only meaningful if result is significant.
  - Indicates how big the difference is (usually normalized to number of std-deviations)

22

## Interpreting results: Significance & Effect Size

- Significant & small effect => ?
  - Real difference, but slight.
  - Probably not of practical importance.
- Significant & large effect => ?
  - Real difference, likely meaningful.
- Significant & small sample => ?
  - Significant & probably important.
- Non-significant & small sample => ?
  - Inconclusive
- Non-significant & large sample => ?
  - Evidence there really is no difference

23

# Group Exercise

- Compute effect size for the study
- Characterize as small/medium/large

- You are now going to do a follow-up study using similar interventions and measures (ie assume same effect size)
- Do a power analysis to determine how many subjects you would need for a two-group between-subjects experiment with 80% power, alpha=0.05, two-tailed test.

24

# Power Analysis in R
# Use 'pwr' package

```
> require("pwr")    #every session
> pwr.t.test(d=0.5,power=.8)

     Two-sample t test power calculation

              n = 63.76561
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: n is number in *each* group
```

25

## Also in ‘pwr’ package

- pwr.t.test — t-tests (one sample, 2 sample, paired)
- pwr.chisq.test — chi-square test
- pwr.r.test — correlation

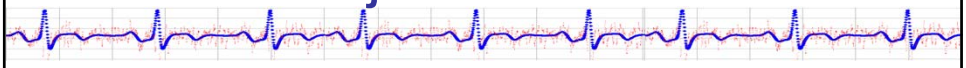
26

## Computing effect size

- Some thoughtless authors do not include means & stddevs (per group) in their article...
- Package ‘compute.es’ contains a variety of methods for computing effect size given other info (e.g., t score, N1, N2)
- Morale: Always include means & stddevs
- Better: Report effect sizes yourself!

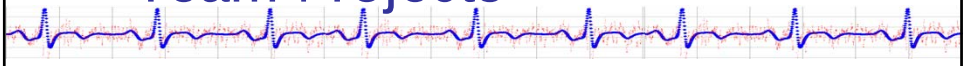
27

## Team Projects



28

## Team Projects



- Goals
  - Give every student 3 cycles of proposal, execution, analysis, reporting of studies
    - Ethnography/Descriptive Study
    - Correlational / Quasi-experimental
    - Experimental
  - Give every student one experience giving oral presentation of study results

29

## Team Projects

- 3 teams per cycle
  - 2-3 members each
  - 8 teams
- On 3/19, 4/2 and 4/16 one member from each team presents results for <10 minutes

30

## Team Projects

- Each project has 2 week duration
- Study proposal
  - Due 2-3 class sessions before deadline (you may submit early)
  - Typically two pages
  - You may not collect data until I approve
- Data collection, analysis, writeup and preparation for oral presentation by due date

31

## Team Projects

- Topic constraints
  - Related in some way to IS
  - Within NU IRB constraints discussed in class
- Writeup
  - At least two pages long
  - Contain both raw data and visualization
  - Analysis (statistics)
  - Discussion (interpretation & implications)
  - More on 3/12

32

## Team Projects

- Oral presentation
  - 10 minutes (hard upper bound)
  - 2 minutes critique
  - Main idea, hypotheses, study design, results, conclusions
  - Visualization of data
  - Either
    - Email ppt to me by 11am, or
    - Bring memory stick with ppt
    - Put on web
    - (best to have a backup, including your laptop)

33

| | Project | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Amirault, Michael | 2 | 1 | 6 |
| Brohman, Joshua M. | 4 | 6 | 2 |
| Brown, Elizabeth L. | 6 | 4 | 8 |
| Chen, Bai-Xiang | 2 | 6 | 1 |
| Chin, Louis | 1 | 1 | 7 |
| Doyle, Benjamin | 6 | 3 | 5 |
| Gallagher, Jonathan | 8 | 2 | 4 |
| Green, Alison J. | 4 | 7 | 6 |
| Guion, Jeffrey J. | 3 | 2 | 5 |
| Jacques, Cody M. | 5 | 5 | 1 |
| Jelly, Christopher D. | 1 | 8 | 2 |
| Kerber, Katharine M. | 8 | 7 | 1 |
| Leibowitz, Alison | 5 | 4 | 4 |
| McGovern, Anna W. | 7 | 8 | 5 |
| Mercier, Matthew | 4 | 1 | 3 |
| Murphy, Michael T. | 7 | 5 | 6 |
| Nakamura, Marika | 1 | 6 | 3 |
| Nashi, Joni | 6 | 2 | 7 |
| Pinto, Christopher J. | 3 | 3 | 4 |
| Scarfo, Joseph J. | 2 | 5 | 3 |
| Schuch, Timothy J. | 3 | 4 | 2 |
| Scorza, Leigh A. | 7 | 3 | 8 |
| Sontag, Trevor A. | 5 | 7 | 7 |

**Team assignments.**
**Highlighted student presents.**

34

# Team Projects

- At completion of each project I will ask you to evaluate your teammates via email

35

## First Team Project

- Ethnographic and/or Descriptive
- Proposals due next class 3/1
  - Only one paragraph for this one
  - I'll give approval within 48 hours
- 1.5 weeks after spring break to conduct, writeup & present on 3/19
- Priorities:
  - Motivation
  - Methodology
  - Descriptive stats
  - Form of writeup & presentation
  - Conclusions/Lessons learned

36

## First Team Project

- This is a descriptive study. It can be as simple as the ethnographic exercise you did in I4, as involved as the behavioral study in I8, or involve the administration of a survey as you did in I9. In any case it should involve some measures that you can do descriptive statistics on.
- Proposals only need to be one paragraph long, but need to describe the general topic you will be investigating, the motivation for the study, and a description of the measures you will collect.
- If you are doing a questionnaire or a structured or semi- structured interview I need to see the exact list of questions you will be asking, so I that I can ensure that you are following the IRB guidelines for non-sensitive topics.

37

# Midterm Review

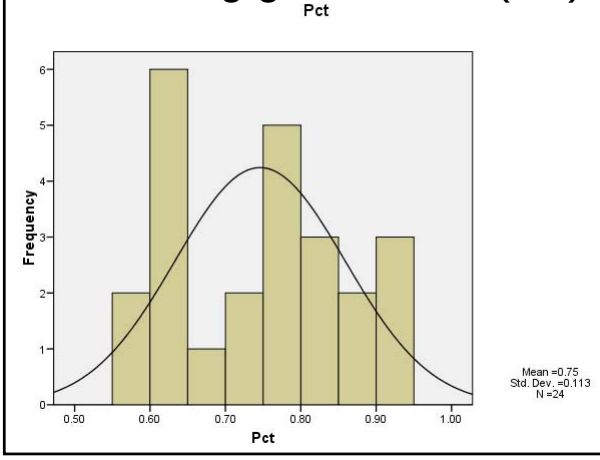
Previous MidTerm is Online

Linked to Exam Review Page

38

# Past MidTerm

- Avg grade: 75% (C+)




| | |
|---|---|
| A+ | 105 |
| A | 100 |
| A- | 95 |
| B+ | 90 |
| B | 85 |
| B- | 80 |
| C+ | 75 |
| C | 70 |
| C- | 65 |

39

## Exam Ground Rules

- NO cheat sheet – closed book, notes
- NO laptops/PDAs/calculators
- I will provide calculators if you need them

40

## All the formulas you need to know

- Required
  - Mean
  - Variance
  - StdDev
  - Z-score
  - Degrees of Freedom
    - Chi^2, t-test for independent means
- Extra credit for anything else
  - Chi-square goodness of fit statistic
  - Correlation
  - t score for t-test for independent means (and supporting formulas)

41

# I5 - Research Designs

• Want to determine cleanliness of houses cleaned with Roomba.

• Design a descriptive study
• Design a demonstration study
• Design a correlational study
• Design an experimental study

• Know which stats to use with each

42

# Measures - Summary

- Reliability
  - Test-retest
  - Internal consistency
- Validity
  - Face
  - Content
  - Criterion-related
    - Concurrent
    - Predictive
  - Construct
    - Convergent
    - Discriminant

- Nominal
- Ordinal
- Interval
- Ratio

43

## I6 – Descriptive Stats

- Type of each measure?
- Type of statistics you can do?
  - Gender(M/F)
  - JobCategory (junior/senior/...)
  - YearsExperience
  - CallAnswerTime
  - CustomerSatisfaction
- Know how to construct/interpret a histogram/frequency distribution

44

## Measures of Center

- *Mode*
  - Used if data are measured along a nominal scale
- *Median*
  - Used if data are measured along an ordinal scale
  - Used if interval data do not meet requirements for using the mean (skewed but unimodal)
- *Mean*
  - Default

45

## Measures of Spread

- *Range*
  - Know what it is
- *Interquartile Range*
  - Know what it is
- *Standard Deviation*
  - Know how to compute
  - Most widely used measure of spread

46

## I8 – Behavioral Measures

- Know sampling methods
  - Time, Individual, Event
- Be able to construct a code book for a given measure
- Know how to measure reliability (inter-rater reliability/Kappa)
  - Don't need to know formula or calculate
- Know how to increase reliability

47

## I9 – Questionnaires

- Difference between reliability and validity (for measures in general)
  - Types of each
- What does internal consistency (Cronbach's $\alpha$) measure?
  - What is a good value?
- Factors vs. Items
  - What are they? (No factor analysis)
- How do you score a composite measure?
- Demographic/Open-ended/Closed-ended/Partially Open-ended/Likert/Semantic Differential items

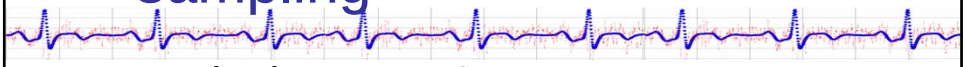
48

## I10 – Chi^2 & Correlation

- Chi^2 goodness of fit
  - What kind of data can you apply to?
  - Observed vs. Expected frequencies?
- Pearson correlation
  - What kind of data can you apply to?
  - Know relationship between typical scatter plots and values for 'r'.
  - Know when you should disregard 'r'
  - Know what significance means for this statistic.

49

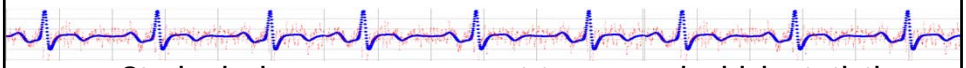
## Sampling



- Why important?
- How different from randomization?
- Convenience, Random, Systematic, Stratified
- How to determine sample size you need?
  - How to determine the number of subjects you need to recruit?

50

## Midterm Review



- Study designs, measurement types, and which statistics to use with each – 50%
  - You may be given descriptions of studies and asked to describe their design and the kind of statistic you would use to analyze their results with.
  - You may be asked to give an example of a particular kind of study design, including a research model and description of variables.
  - You may be given an excerpt of data from a study and asked what kind of statistic you should use.
  - Also important is an understanding of extraneous/confounding variables and how to deal with them, and a conceptual understanding of mediating and moderating variables, proximal vs. distal outcome variables, and independent vs. dependent variables (relative to a research model).

51

# Types of study design

| | Number of Variables | Number of IV Levels | Manipulation |
|---|---|---|---|
| Descriptive | 1 | NA | NA |
| Demonstration | $\geq 2$ | 1 | √ |
| Correlational | $\geq 2$ | NA | NA |
| Experimental | $\geq 2$ | $\geq 2$ | √ |

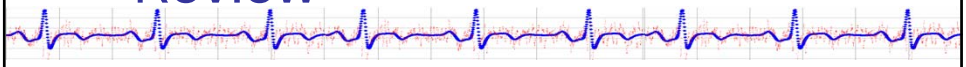
53

# The grand plan

- X^2 tests
  - For nominal measures
  - Can apply to a single measure
- Correlation tests
  - For two numeric measures
- t-test for independent means
  - For categorical IV, numeric DV

54

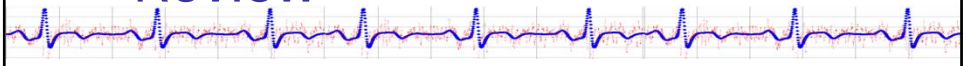
## Review



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

56

## Review



- Kinds of analyses for Demonstration studies?
  - Same as for Descriptive studies

57

## Review

- Kinds of analyses for Correlational studies?
  - Descriptive
  - Correlation [numeric/numeric]
  - Chi-square test for independence [categorical]

58

## Review

- Kinds of analyses for Experimental studies?
  - Descriptive
  - Correlation [numeric/numeric]
  - Chi-square test for independence [categorical/categorical]
  - T-test for independent means [binomial/numeric]

59

## Midterm Review

- Descriptive statistics – 10%
  - You should be able to determine the mean, median, mode, variance, and standard deviation for a data sample (when appropriate), and be able to construct a frequency distribution for it and be able to state whether it is unimodal or bimodal, symmetric or positively or negatively skewed.

60

## Midterm Review

- Concepts – 10%
  - You should be able to describe the following:
    - Empirical vs. analytic research; scientific explanations; the scientific method; qualitative vs. quantative research methods (and examples of each); primary vs. secondary research literature; meta-analysis; the three ethical principles of human subjects research (from lecture); sampling; generalization of study results; ethnographic research; the typical use of different types of studies during the software development process; internal vs. external validity of a study; reliability vs. validity of a measure; system usability; coding manual (for behavioral/observational measures); inter-rater reliability; retrospective vs. prospective study; demographic measures; between subjects design; distribution of individuals vs. distribution of means; levels/treatments/conditions in a study.

61

# Midterm Review

- Hypothesis testing – 10%
  - You should be able to describe the basic logic of hypothesis testing, research vs. null hypotheses (and how they relate to the populations / population parameters in a study and whether the test is one vs. two-tailed), and significance levels.
  - If provided output from SPSS for Chi-squared goodness-of-fit, Pearson correlation or t-test of independent samples you should be able to state what the results mean in English (relative to the study hypotheses) and reformat the results into publication format.
  - **Understand power, effect size, Type I & II errors**

62

# Review: Basic Process of Hypothesis Testing

- H1: Research Hypothesis:
  - Population 1 is different than Population 2
- H0: Null Hypothesis:
  - No difference between Pop 1 and Pop 2
  - *The difference is "null"*
- Compute *p(observed difference|H0)*
  - 'p' = probability observed difference is due to random variation
- If p<*threshold* then reject H0 => accept H1
  - p typically set to 0.05 for most work
  - p is called the "level of significance"

63

# Know this.

| | "The Truth" H1 True | "The Truth" H1 False |
|---|---|---|
| **Decide to Reject H0 & accept H1** | Correct p = power | Type I err p = $\alpha$ |
| **Do not Reject H0 & do not accept H1** | Type II err p = $\beta$ | Correct p = 1-$\alpha$ |

64

# R – chi-sq goodness of fit

```
> chisq.test(c(9,7,1),p=c(.1,.1,.8))

        Chi-squared test for given probabilities

data:  c(9, 7, 1)
X-squared = 59.5441, df = 2, p-value = 1.175e-13
```

65

## R – chi-squar etest for independence

```
> chisq.test(table(g$Party,g$GunOwner))

Pearson's Chi-squared test with Yates' continuity
correction

data:  table(g$Party, g$GunOwner)
X-squared = 101.8156, df = 1, p-value < 2.2e-16
```

66

## R - correlation

```
> cor.test(c(1,2,3,4),c(5,9,2,7))

      Pearson's product-moment correlation

data:  c(1, 2, 3, 4) and c(5, 9, 2, 7)
t = -0.0612, df = 2, p-value = 0.9568
alternative hypothesis: true correlation is not equal
to 0
95 percent confidence interval:
 -0.9642547  0.9576448
sample estimates:
        cor
-0.04323377
```

67

# R – t-test for indep means

```
> t.test(c(1,2,3,4),c(6,1,8,9),var.equal=T)

        Two Sample t-test

data:  c(1, 2, 3, 4) and c(6, 1, 8, 9)
t = -1.8489, df = 6, p-value = 0.114
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -8.131929  1.131929
sample estimates:
mean of x mean of y
      2.5       6.0
```

68

# Midterm Review
# Misc Topics – 5% each

- Sampling and generalization. You should be able to describe simple random, systematic, and stratified sampling, and what constitutes a representative vs. biased sample and how this affects the generalizeability of your study.
- You should be able to construct a scatter plot from sample data and be able to estimate (using the "eyeball" method) the Pearson correlation coefficient given a scatter plot.
- Measures. You should be able to describe the steps you would take in validating a measure, including test-retest reliability, internal consistency, and face, content, criterion, and construct (convergent and divergent) validity, or identify when these are mentioned in the description of a measure. You will not be asked to calculate any of the figures, but you may be asked to interpret them.
- Survey measures. You should be able to describe and provide examples of: open-ended questions; restricted/closed-ended questions; rating scales (including Likert and semantic differential); and composite measures. If provided with a filled out survey index (composite measure) you should be able to compute the composite score.

69