Statistical Significance & Effect Size

Michael A. DeDonno Ph.D.
www.michaeldedonno.com

I WANT YOU TO PROMOTE STATISTICS

Why study statistics?

- Because at some point in your career, you will see statements such as the following:

  - A one way ANOVA showed that the differences in quiz scores between the control group (n = 3, M = 4.00, SD = 1.00, the first experimental group (n = 3, M = 8.00, SD = 1.00), and the second experimental group (n = 3, M = 9.00, SD = 1.00) were statistically significant, $F(2,6) = 21.00$, $p = .002$, $\eta^2 = .88$.

Why study statistics?

- Because you are 3x more likely to meet someone on Match.com than not using Match.com

- Because three out of four dentists recommend Crest toothpaste

Why should we be concerned with Effect Size?

- Because Psychologist / Statistics author, Jacob Cohen (1923-1998) said so:

  - *"The primary product of a research inquiry is one or more measures of effect size, not p value."*

Agenda

- Random "chance" factors in research

- *p* value defined

- Factors that impact *p* value

- Beyond the *p* value ~ effect size

- Common indices of effect size

- Recommendations for future research

The Influence of Random / Chance Effects

- Random factors influence our lives
  - Overhearing a job opportunity
  - Meeting someone at a coffee shop
  - Hearing an old favorite song
  - Finding a stray dog on the streets
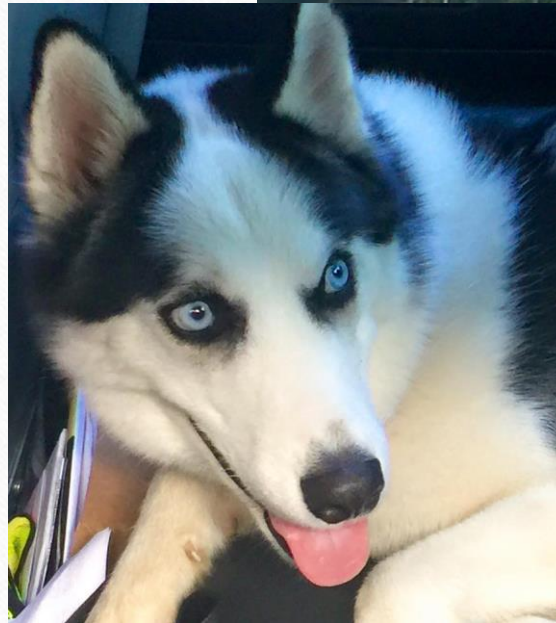
The Influence of Random / Chance Effects

- Similar to life, randomness has an influence on the outcome of research
- However, unlike life where randomness can have a positive impact, it's typically an unwelcome factor in research
  - Randomness in place of an Intervention, can be the cause of a certain result
    - May cause the researcher to believe the Intervention worked, when in reality it did not
  - A *p* **value** is used to help determine the probability of results occurring by chance

Research Example

- Students generally want to do well on tests.
  - Measure of ability / performance
  - Authorization to continue in program
- It would be of value to uncover a "safe and legal" supplement that could enhance test performance

Research Question

- Does drinking caffeine improve performance on an cognitive test?

- The study: n = 56.
  - 28 students drink 100mg of caffeine before a test
  - 28 ingest a placebo before a test

Randomly have students count off by 2 to determine treatment and control groups

X   X   X   X   X   X   X

X   X   X   X   X   X   X

X   X   X   X   X   X   X

X   X   X   X   X   X   X

X   X   X   X   X   X   X

X   X   X   X   X   X   X

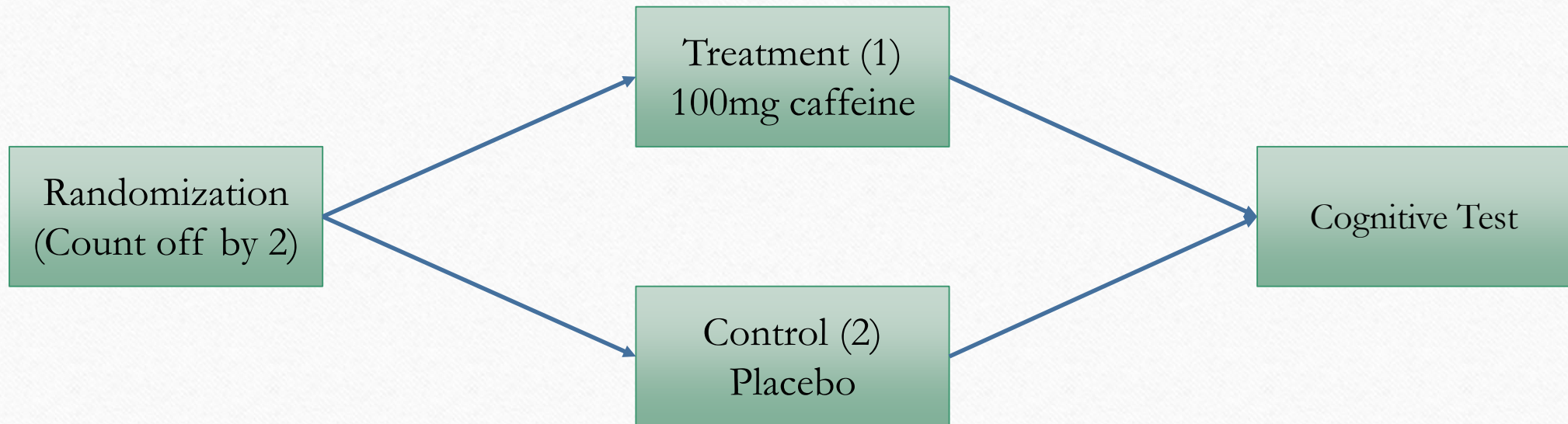X   X   X   X   X   X   X

X   X   X   X   X   X   X

Breakdown of Treatment and Control groups

1 = Treatment Group

2 = Control Group

1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2

Conduct Study

Results

- Results revealed a statistically significant difference in test performance. Treatment group had higher scores on the test than the Control group. Gets published in American Journal of Education

- FAU institutes a "Drink coffee before a test" campaign

  - Orders and places signs throughout the campus

  - Covers the cost of Starbucks coffee on test days

The Influence of Random / Chance Effects

- However, the Semester test results show a minimal change in test performance.

  - FAU invested time & money in an attempt to improve test scores of its students

  - How could this have occurred, particularly when results showed a p value of $< .05$?

Let's first look at the sample of participants

1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2

There is evidence that the best performing students tend to sit in the first row and the entire center row

1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2

There is evidence that the best performing students tend to sit in the first row and the entire center row

10 of the 28 students in Treatment group

1	1	1	1	1	1	1
2	2	2	2	2	2	2
1	1	1	1		1	1
2	2	2	2			
1	1	1	1			
2	2	2	2	2	2	2
1	1	1	1	1	1	1
2	2	2	2	2	2	2

4 of the 28 students in the Control group

The seemingly random selection process may have influenced the results of the study

The Influence of Random / Chance Effects

- Other random (or chance) factors that could have influenced the results
  - A greater number of the Treatment group were new caffeine drinkers making the influence of caffeine more significant
  - A greater number of the Control group were advanced caffeine drinkers making the influence of caffeine less pronounced

The Influence of Random / Chance Effects

- General random factors that can influence results
  - Temperature in one room (treatment) vs. another room (control)
  - Randomness of test questions
  - Time of day
  - Mood of specific participants

p-value

- We need to verify randomness or chance factors are not overly influencing our results

At the Biometric Laboratory, University College, London.

TABLES FOR STATISTICIANS AND BIOMETRICIANS. Edited by KARL PEARSON, F.R.S.

The third edition of this book will consist of two Parts

PART I issued in 1930 embraces the Second Edition revised. It may be obtained direct from the Biometric Laboratory, University College, London, price 15*s*. net, plus 1*s*. postage to any address, or through any bookseller.

PART II will contain all the Tables issued in *Biometrika* during the last sixteen years together with a number of Tables not yet published, but at present being computed. It is hoped to issue it this year.

PRESS NOTICES OF THE FIRST EDITION

"To the workers in the difficult field of higher statistics such aids are invaluable. Their calculation and publication was therefore as inevitable as the steady progress of a method which brings within grip of mathematical analysis the highly variable data of biological observation. The immediate cause for congratulation is, therefore, not that the tables have been done but that they have been done so well....The volume is indispensable to all who are engaged in serious statistical work."—*Science*

"The whole work is an eloquent testimony to the self-effacing labour of a body of men and women who desire to save their fellow scientists from a great deal of irksome arithmetic; and the total time that will be saved in the future by the publication of this work is, of course, incalculable....To the statistician these tables will be indispensable."—*Journal of Education*

"The issue of these tables is a natural outcome of Professor Karl Pearson's work, and apart from their value for those for whose use they have been prepared, their assemblage in one volume marks an interesting stage in the progress of scientific method, as indicating the number and importance of the calculations which they are designed to facilitate."—*Post Magazine*

(vii)

What is a p value?

- What does a *p* value of less than .05 mean?
  - It's significant
- But what does the .05 mean?
  - It's significant

What is a *p*-value?

- Statistical significance - Probability *p-value*: Identifies the likelihood a particular outcome may have occurred by chance

 - Value range from 0 (0%) to 1.00 (100%)

What is *p*-value

- Statistical significance:
  - *p* = .80 = There is a 80% probability the findings occurred by chance.
    - Evidence that chance factors (i.e. randomness) influenced the results
  - *p* < .05 = There is less than a 5% probability the findings occurred by chance

# $p < .05$ = There is less than a 5% probability, the findings occurred by chance

But what impacts a p value?

What impacts a *p*-value?
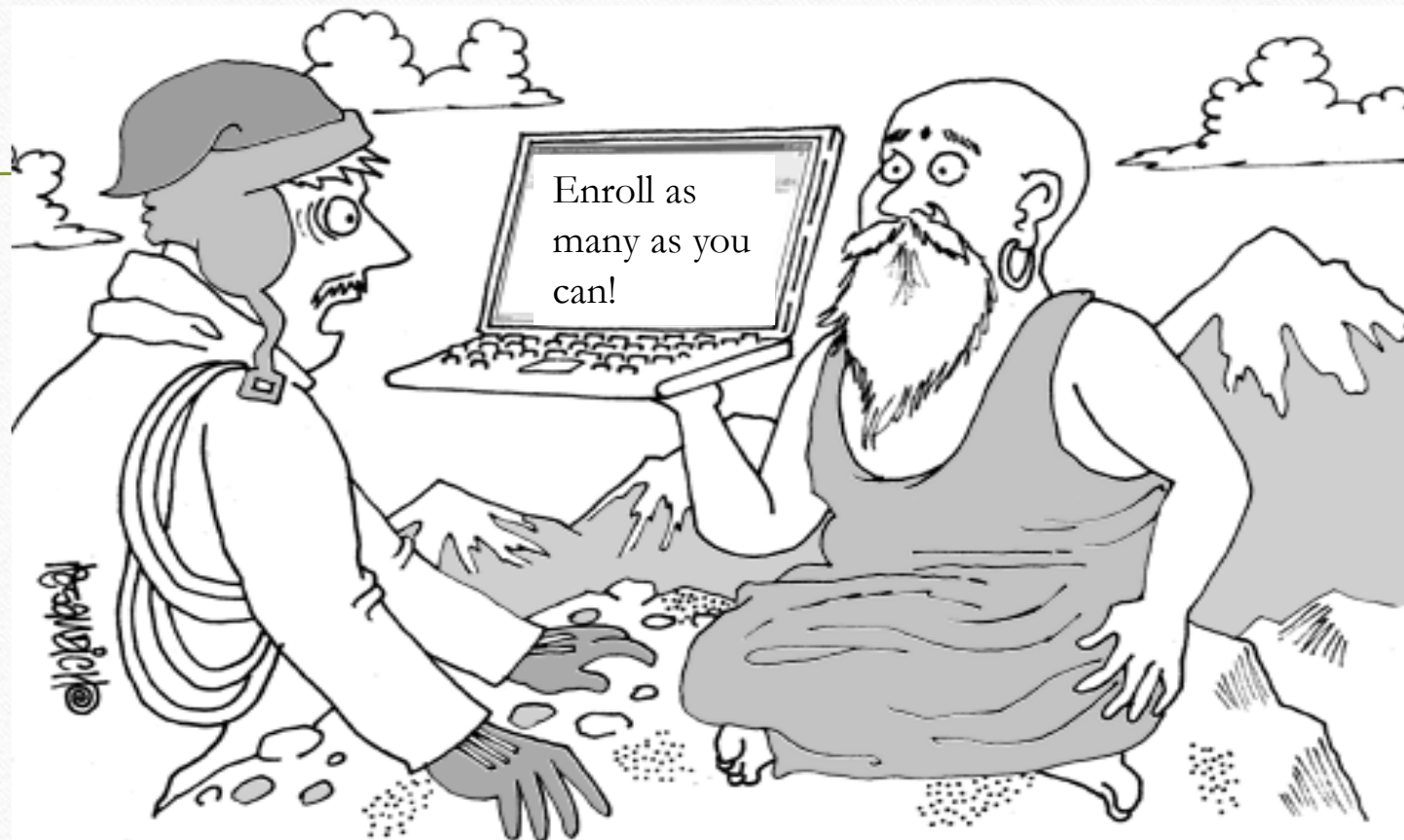
"I climbed all this way, and you tell me That's the meaning of p-value?"

Beyond *p*-value

What if it cost $200.00

- FAU researchers recently found a statistically significant difference between a "safe and legal" Supplement and a Placebo on test performance

- Would you pay $ 1.00 for this safe and legal supplement that research has shown to result in a statistically significant difference (improvement) in test performance?

Beyond *p-value*, what's important? Effect size!

- **Effect Size** – A name given to indices that measure the relative magnitude of treatment effect.

SAFFIR-SIMPSON HURRICANE SCALE
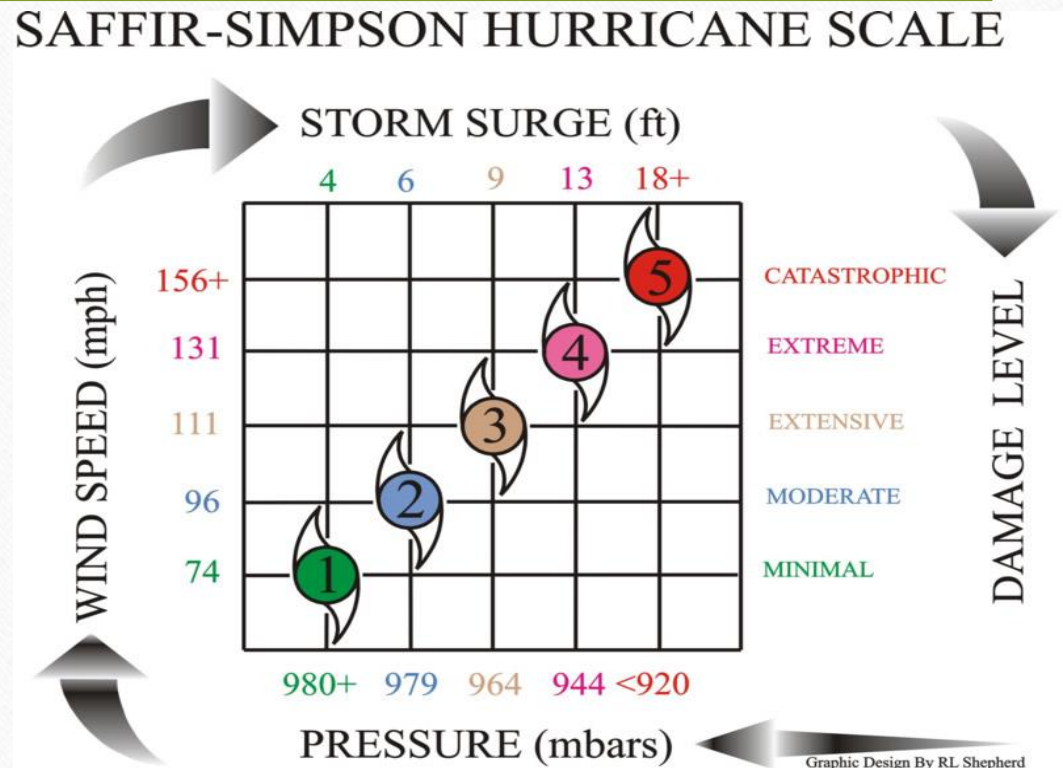
Common Indices of Effect Size

- Comparison Studies
  - Cohen's d
  - Odds ratio (OR)
  - Relative risk or risk ratio (RR)
- Relational studies (all correlations are effect sizes)
  - Pearson's r correlation
  - r^2 coefficient of determination

Cohen's d

- The difference between two means (e.g., treatment mean minus control mean) divided by the standard deviation of the two conditions

$$d = \frac{\overline{X_1} - \overline{X_2}}{s}$$

- What precisely the standard deviation (s) is, was not originally made explicit by Cohen
  - Defined as, the standard deviation of either population (since they are assumed to be equal)

Cohen's *d*

- Identified specific effect size values:
  - .2 = small effect   .5 = medium effect   .8 = large effect

  NOTE: Ideally, interpretation of results should be grounded in a meaningful context or by quantifying their contribution to knowledge. Where this is problematic, Cohen's effect size criteria may serve as a backup.
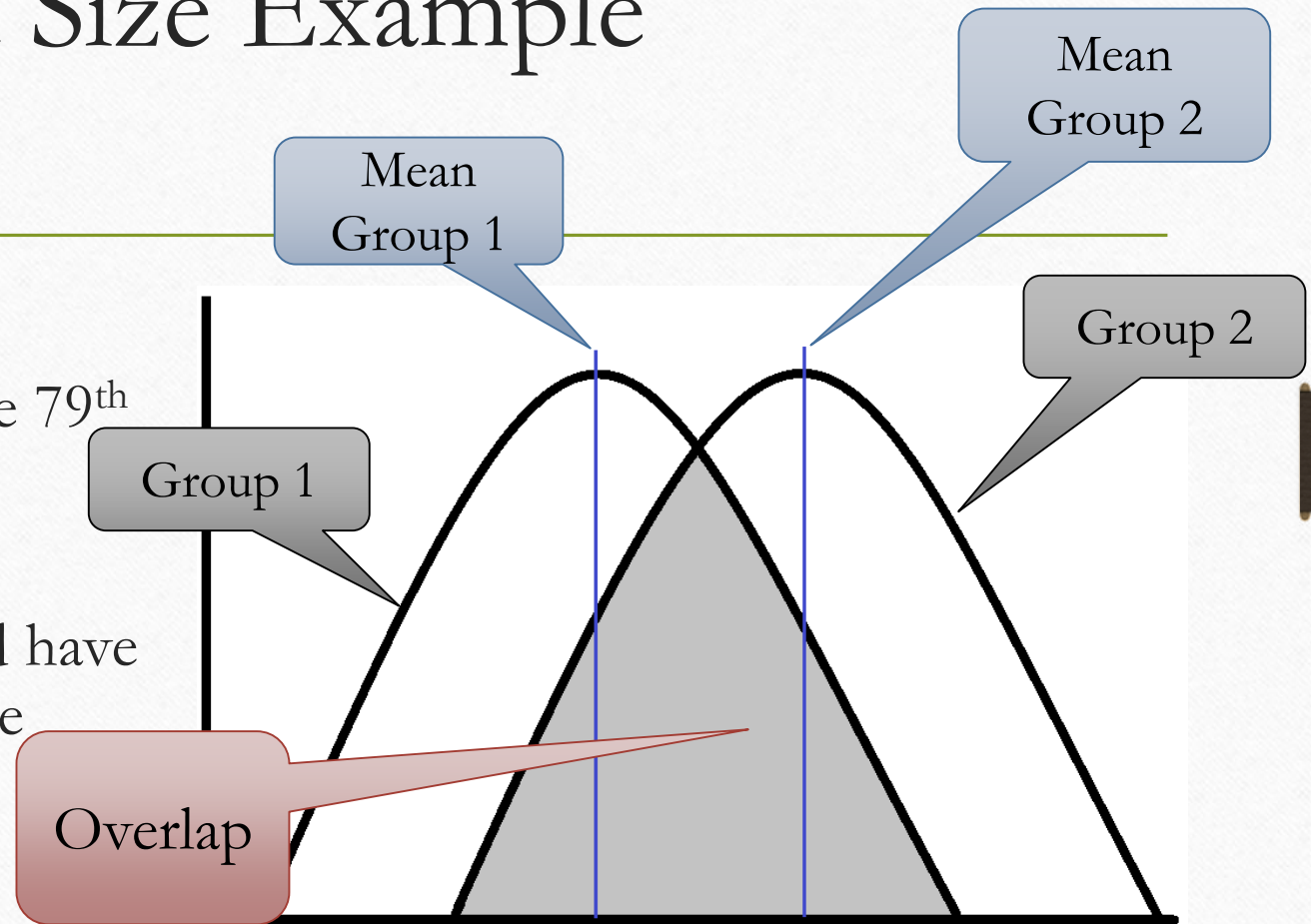
Effect Size Example

- Effect size = 0.2
- The mean of Group 2 is at the 58th percentile of Group 1
- Someone in Group 2 with an average score (ie, mean) would have a higher score than 58% of the people in Group 1
- **85% overlap of participants**

Mean Group 1

Mean Group 2

Group 1
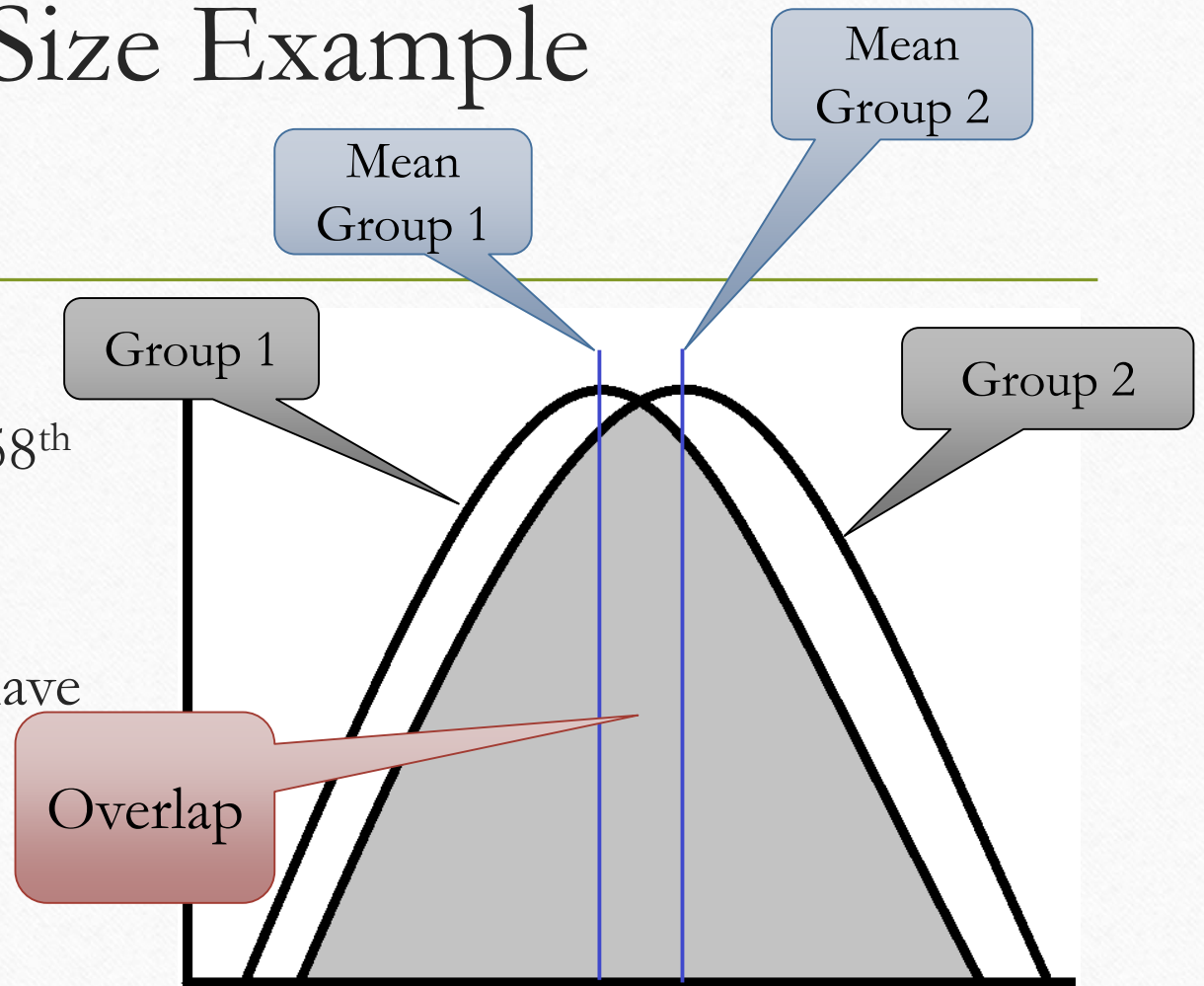
Group 2

Overlap

Effect Size Example

- Effect size = 0.8
- The mean of Group 2 is at the 79th percentile of Group 1
- Someone in Group 2 with an average score (ie, mean) would have a higher score than 79% of the people in Group 1
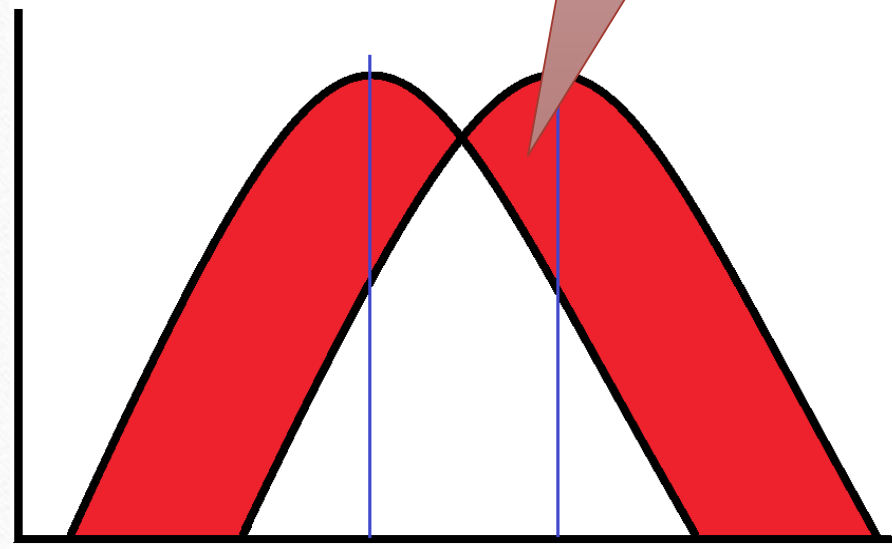- 53% Overlap of participants

Mean Group 1

Mean Group 2

Group 1

Group 2

Overlap

Effect Size Example

Effect Size = 0.2

Effect Size = 0.8

As Effect Size increases, magnitude of difference increases
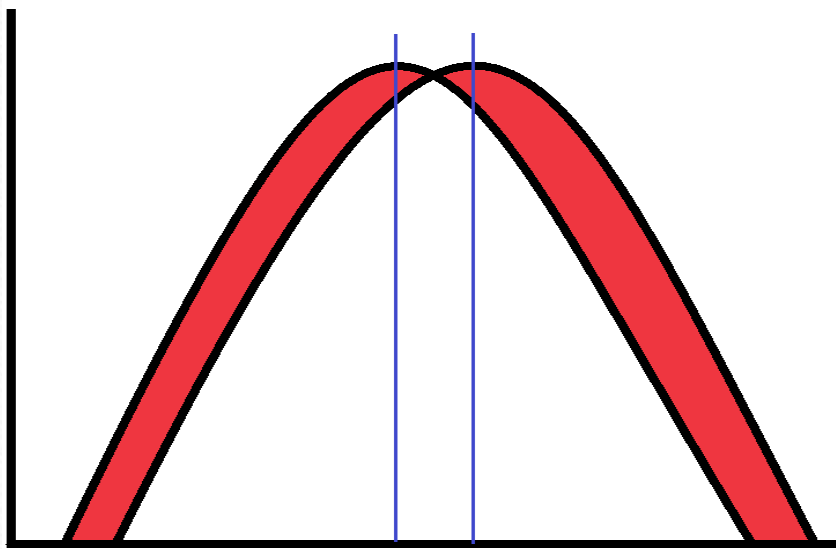
Eta-squared (η^2)

- Eta-squared is a measure of effect size typically for use in ANOVA

- Proportion of variance in Y explained by X

- Interpret η^2 (Cohen):

  - .02 ~ Small

  - .13 ~ Medium

  - .26 ~ Large

- ***Remember! Interpretation of results should be grounded in a meaningful context, or by quantifying their contribution to knowledge.***

Odd Ratio (OR)

- Appropriate when both variables are binary

- Research example - Influence of positive priming on passing a class

  - Control: 2 students pass for every 1 that fails

    - Odds of passing are two to one (or 2/1 = 2)

  - Treatment: 6 students pass for every 1 that fails

    - Odds of passing are six to one (or 6/1 = 6)

  - Effect size computed by noting the odds of passing in Treatment group are three times higher than the Control group (6 ÷ 2 = 3)

  - OR = 3 (Note: not comparable to Cohen's d)

Relative Risk (RR) aka risk ratio

- The risk (probability) of an event relative to some independent variable
  - Different from odds ratio (OR) in that it compares probabilities instead of odds
- Research example - Influence of positive priming on passing a class
  - Control: 2 students pass for every 1 that fails
    - Probability of passing is 2/3 (or 0.67)
  - Treatment: 6 students pass for every 1 that fails
    - Probability of passing is 6/7 (or 0.86)
  - RR = 0.86 / .067 = 1.28 (Note: not comparable to Cohen's d)

Coefficient of determination (R^2 or r^2)

- An output of regression analysis

- Interpreted as the proportion of the variance in the dependent variable (criterion) that is predictable from the independent variable (predictor).

- Range from 0 to 1

Coefficient of determination (R^2 or r^2)

- Examples
  - $R^2 = 0 \sim$ Dependent variable cannot be predicted from the independent variable
  - $R^2 = 1 \sim$ Dependent variable can be predicted without error from the independent variable
  - $R^2 = .20 \sim 20\%$ of the variance in the dependent variable is predictable from the independent variable.

  - Trait mindfulness explained 28% ($R^2 = .28$) of the variance in test anxiety (Altairi, 2014)
    - Note: Additional factors not explored in the study, explain 72% of the variance in test anxiety.

Large sample size example

- Physicians Health Study (1989) of aspirin to prevent myocardial infarction (MI)

  - In more than 22,000 participants over an average of 5 years, aspirin was associated with a reduction of MI that was highly statistically significant: **$p < .00001$**.

  - Due to the conclusive evidence, aspirin was recommended for general prevention.

  - **However, the effect size was extremely small ($r = .034$, $r^2 = .001$)**

  - As a result, many people were advised to take aspirin who would not experience the benefit, yet were also at risk for adverse effects (e.g., bleeding in the stomach or brain).

  - FDA (2014), revised position on the use of daily aspirin.

Review

- **Statistical significance** - Probability **p-value**: Identifies the likelihood a particular outcome may have occurred by chance

  - $p < .05$ = There is less than a 5% probability the findings occurred by chance

Review

- **Statistical significance** - Probability *p-value*:
  - Considered to be confounded because of its dependence on sample size
    - Sometimes statistical significance means only that a huge sample size was used

Review

- **Effect size** – Measurements that tell us the relative magnitude of the experimental treatment.
  - Tells us the *size* of the experimental *effect*
    - How much did Treatment A improve test performance vs. Treatment B
    - How much did a therapy improve function vs. normal activities
  - Effect size can be used to justify the costs of a new therapy
    - Cost vs. impact

Back to the initial slide

- A one way **ANOVA** showed that the differences in quiz scores between the control group **(n = 3, M = 4.00, SD = 1.00),** the first experimental group **(n = 3, M = 8.00, SD = 1.00)**, and the second experimental group **(n = 3, M = 9.00, SD = 1.00)** were statistically significant, **F(2,6) = 21.00**, **p = .002, η^2= .88.**

Analysis of Variance

n = Sample size, M = Mean, SD = standard deviation

F statistic

P value of .002

Eta squared (effect size) of .88

Closing Remarks

- Perception of statistics
  - "98% of all statistics are made up." Author Unknown
  - "Statistics can be manipulated to prove anything." Pasadena Start - News
  - "Statistics: the mathematical theory of ignorance." Morris Kline

Closing Remarks

- Statistics are based on data
  - In research, data is king! - Accurate collection of data is imperative.
    - A poorly executed study that still yields accurate data can be of value
    - A well executed study with inaccurate data is worthless
  - Statistical calculations based on accurate data is not made up and does not have the flexibility to prove anything. Results are real and fixed! (assuming appropriate, accurate and ethical principals were applied)
  - The challenge comes in a lack of understanding and the interpretation of the data
    - We are becoming an Innumerate society (Paulos, 1989)

Closing Remarks (3 of 3)

- When it comes to developing a research question – Keep it Simple!

- Collect your data with absolute precision

    - Consider a Data Management Plan (DMP) that includes a tasks such as; data input, data transfer, edit checks, source data verifications, etc. (see Prokscha, 2012)

- Take the time to understand your results

    - Results should be grounded in a meaningful context or by quantifying their contribution to knowledge.

    - Remember – We use a sample to Estimate or Predict something relative to a population.

Contact Information

- Michael A. DeDonno Ph.D.
    - College of Education, Office #254
    - mdedonno@fau.edu
    - www.michaeldedonno.com

References

Altairi, M. A. (2014). The impact of mindfulness and test anxiety on academic performance. (Senior Honors Thesis). Paper 39. Retrieved from http://ir.library.louisville.edu/honors/39

Paulos, J. A. (1989). *Innumeracy: Mathematical Illiteracy and its Consequences*. New York, NY: Hill and Wang.

Prokscha, S. (2012). *Practical Guide to Clinical Data Management (3rd ed.)*. Boca Raton, FL: Taylor & Francis.

- http://rpsychologist.com/d3/cohend/