Beyond *p* value, the importance of Effect Size

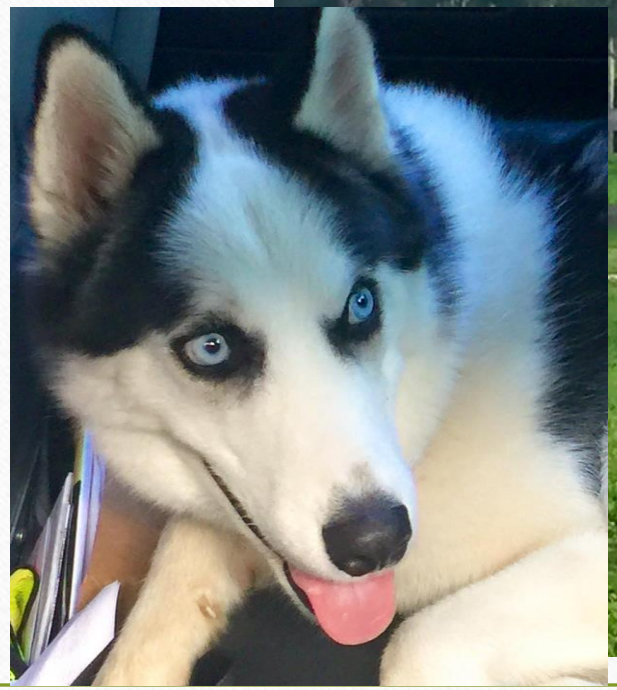
Michael A. DeDonno Ph.D.
www.michaeldedonno.com

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

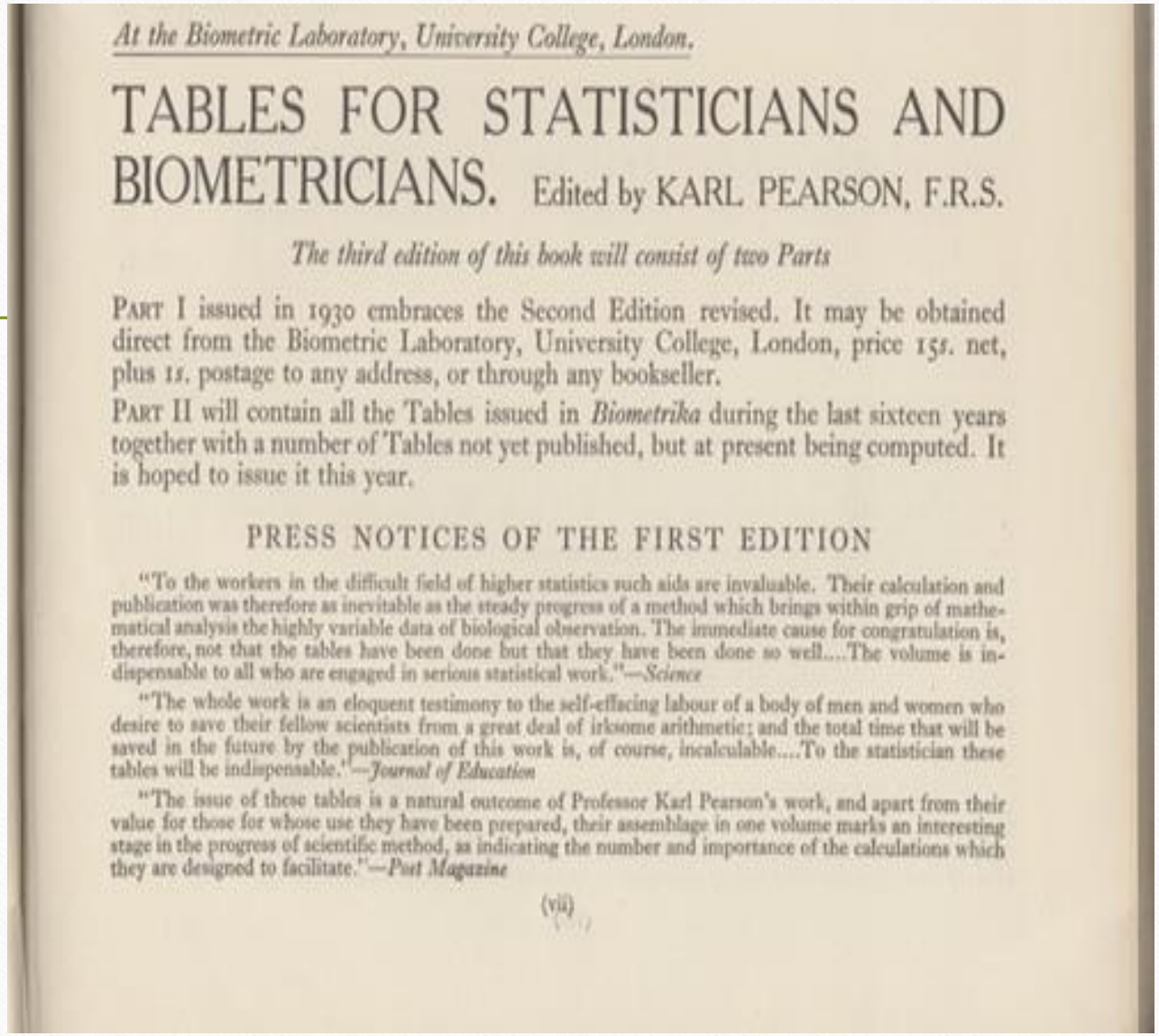


The Influence of Random / Chance Effects

- 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

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*

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

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

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

What impacts a *p*-value?

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

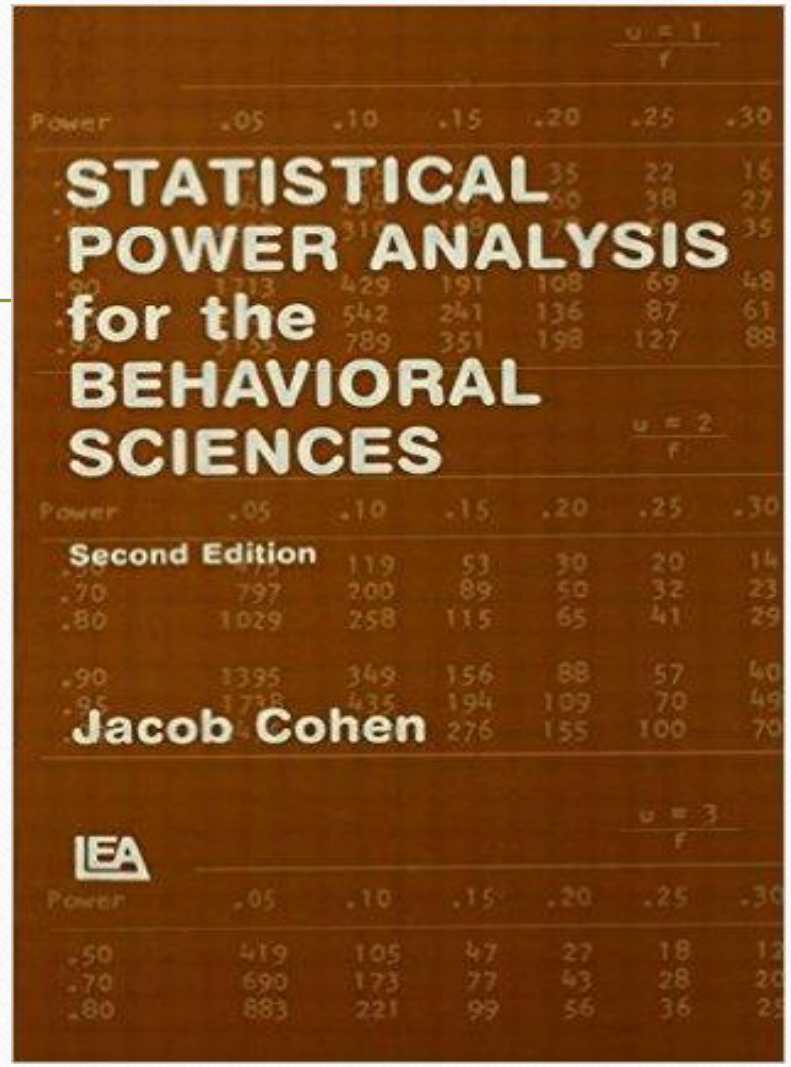
What if it cost $200.00

Beyond *p*-value

- 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?

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."*



Why should we be concerned with Effect Size?

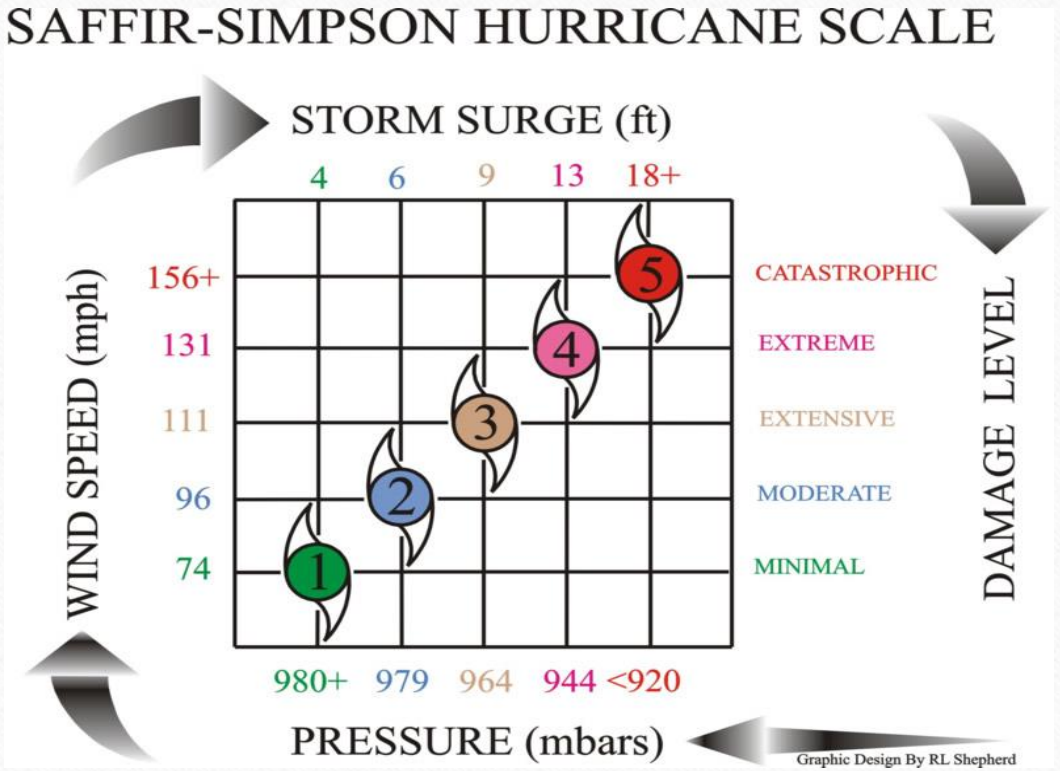
AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

Beyond *p-value*, Effect size!

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



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: $\boldsymbol{p < .00001}$.
  - 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.

Common Indices of Effect Size

- Comparison Studies
  - Cohen's *d*
  - Odds ratio (OR)
  - Relative risk or risk ratio (RR)
  - Number Needed to Treat (NNT)
- 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)

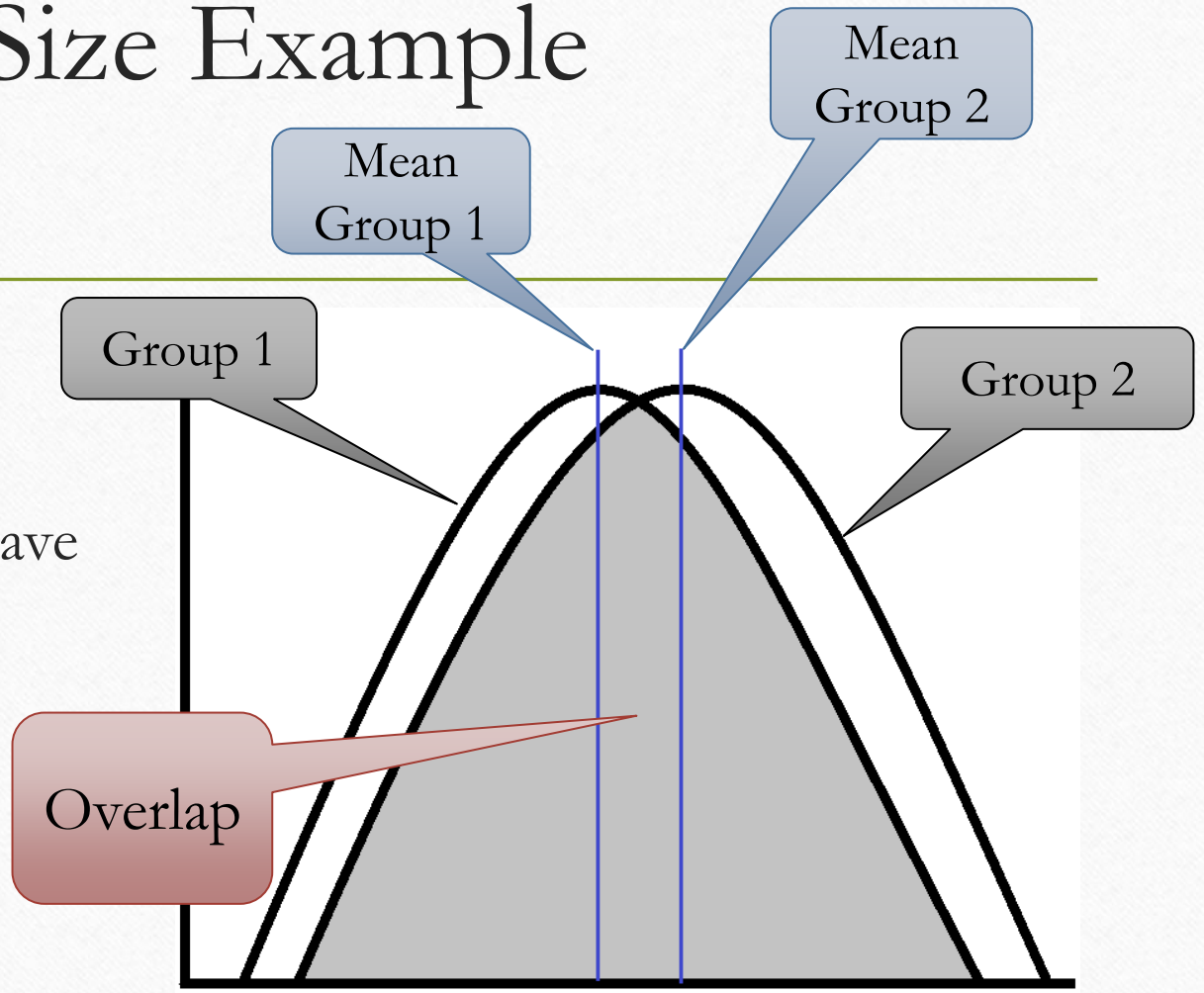
Cohen's *d*

- 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.

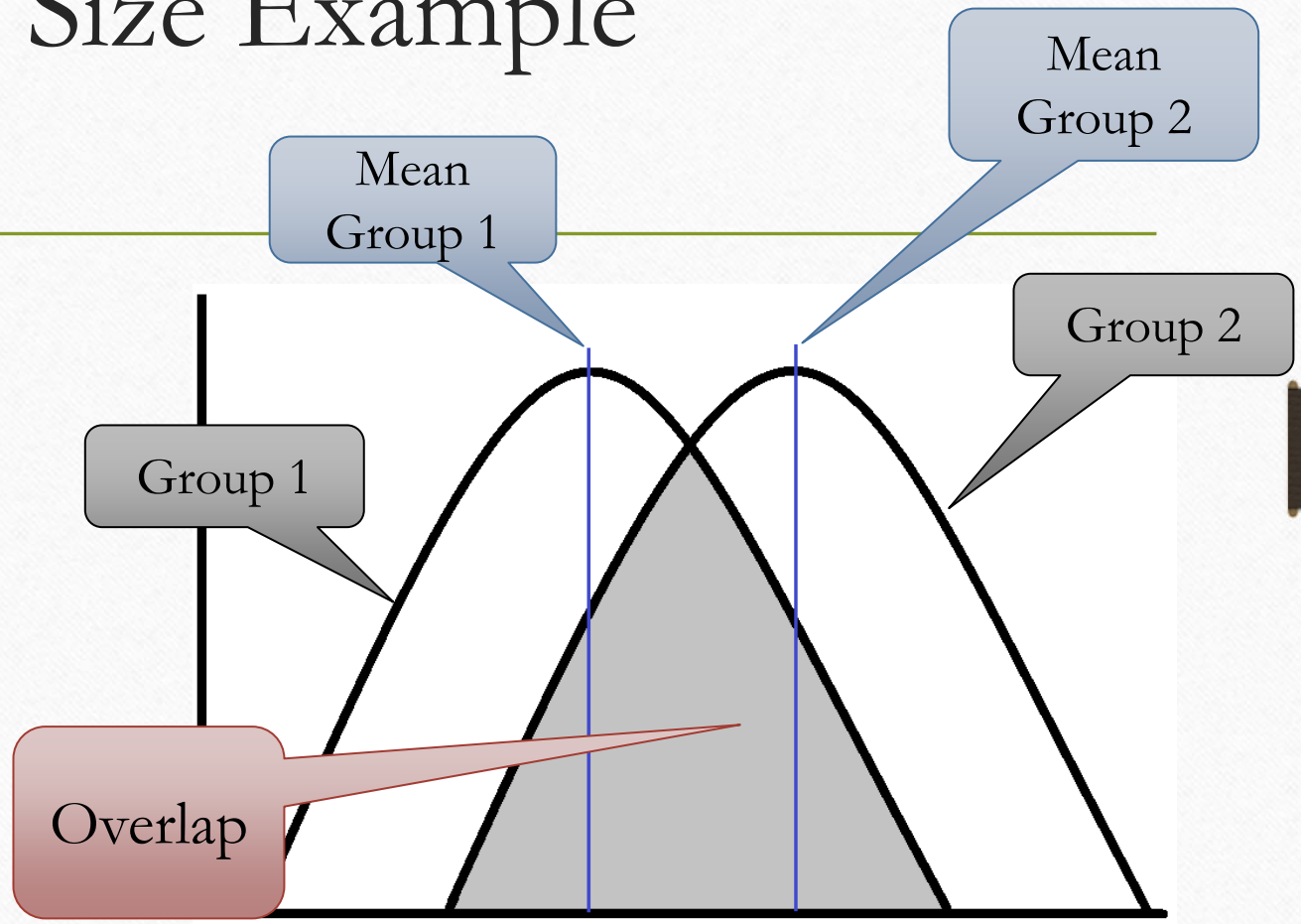
Effect Size Example

- Effect size = 0.2
- 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**



Effect Size Example

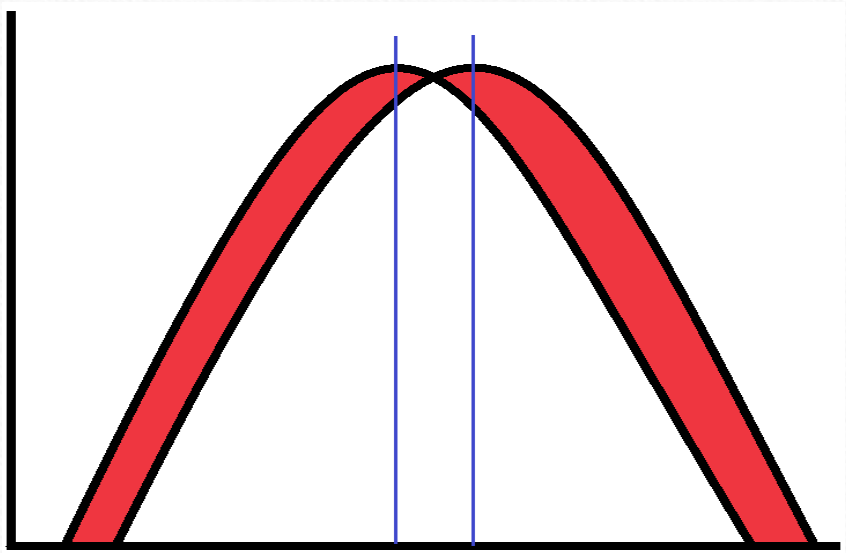
- Effect size = 0.8
- Someone in Group 2 with an average score (ie, mean) would have a higher score than 79% of the people in Group 1
- **53% Overlap of participants**



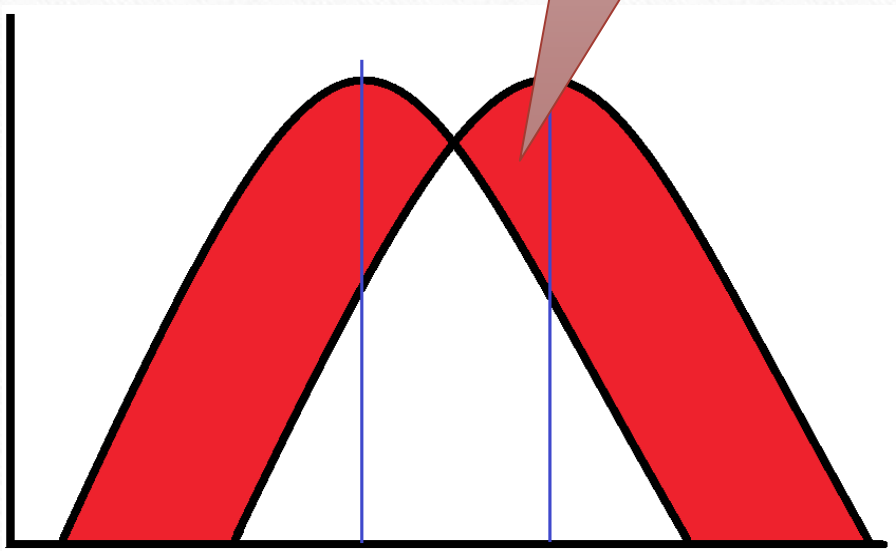
Effect Size Example

As Effect Size increases, magnitude of difference increases

Effect Size = 0.2



Effect Size = 0.8



Cohen's *d* clinical example

- The Research Unit on Pediatric Psychopharmacology Anxiety Study Group (RUPP; 2001). The influence of Fluvoxamine on Pediatric Anxiety
  - DV: Score on the Pediatric Anxiety Rating Scale
  - IV: Medication. Two levels (Fluvoxamine, Placebo)
  - Results (N = 128)
    - Mean SD: Treatment = 9.0 +/- 7.0, Control = 15.9 +/- 5.3, $p < .001$
    - Difference: -6.9 (95% CI: -4.6, -9.2), $d = 1.1$
    - 84% of placebo group had worse scores than the average score of fluvoxamine group.

Be Mindful of Clinical Consequences!

- A small effect (e.g., $d = 0.2$) from study comparing treatments related to mortality rates for two chemotherapies for breast cancer would have greater clinical consequence than a large effect (e.g., $d = 0.80$) from a study of treatment related to ADHD symptom reduction.

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.***

Relative Risk (RR) aka risk ratio

- Cohen's *d* useful for estimating effect sizes from quantitative or continuous measures
- For categorical measures (e.g., Improved vs. Not Improved) consider RR and OR
- RR particularly useful in prospective studies to assess differences in treatments
- Example - Influence of Cognitive Behavioral Therapy (CBT) on children with Asperger's ability to pass a social awareness test
    - 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)

Relative Risk (RR) aka risk ratio

- In the RUPP example
  - 76% of the subjects receiving fluvoxamine were treatment responders according to Clinical Global Impressions (CGI) improvement ratings
  - 29% of the placebo group were treatment responders according to Clinical Global Impressions (CGI) improvement ratings
  - RR = 2.6
  - **Results suggest the patients treated with fluvoxamine for anxiety disorders had almost a threefold greater probability of responding than those on placebo**

Odd Ratio (OR)

- Appropriate when both variables are binary
- Research example - Influence of Cognitive Behavioral Therapy (CBT) on children with Asperger's ability to pass a social awareness test
  - 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)
  - OR = 6 / 2 = 3 (Note: not comparable to Cohen's d)
  - Odds of passing of Treatment group are three times higher than the Control group

Number Needed to Treat (NNT)

- Number of subjects one would expect to treat with agent A to have one or more successes (or one less failure) that if the same number were treated with agent B
- Well suited for binary (success/failure) outcomes
  - NNT = 100 / (% improved on Treatment - % improved on Placebo)
- RUPP study: 76% improved on fluvoxamine, 29% improved on placebo
  - NNT = 100 / (76-29) = 2.1
  - **For every two patients treated with fluvoxamine, at least one will have a better outcome than if treated with placebo**

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
  - $R^2 = 0$ ~ Dependent variable cannot be predicted from the independent variable
  - $R^2 = 1$ ~ Dependent variable can be predicted without error from the independent variable
  - $R^2 = .20$ ~ 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.

Review (1 of 3)

- **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 (2 of 3)

- **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 (3 of 3)

- **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

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/