TALKING ABOUT EFFECT SIZES WITH YOUR STATISTICIAN

Adam Ciarleglio, PhD
Biostatistics, Epidemiology and Research Design (BERD) Lecture

December 11th, 2017

Outline

• Motivating Examples
• P-value Is Not Enough
• Effect Sizes for Two Group Comparison
• Other Effect Sizes
• Summary

The content of these slides are intended for educational and not-for-profit use only. You are free to share and adapt, but must give appropriate credit to the author, Adam Ciarleglio, PhD, who is affiliated with the Irving Institute for Clinical and Translational Research and the Department of Biostatistics at the Mailman School of Public Health, and indicate whether any changes were made to the source material.
MOTIVATING EXAMPLES

Two motivating studies

• **Study 1**: A randomized controlled trial (RCT) is conducted to assess the difference between Citalopram and placebo in treating major depressive disorder (MDD). 250 subjects with MDD are randomized to Citalopram and 250 are randomized to placebo. After 8 weeks, Hamilton Rating Scale for Depression (HRSD) scores are measured and compared between the two groups.

  - **Treatment** = Citalopram or Placebo
  - Two independent groups
  - **Outcome** = HRSD score
    - 17 items
    - Scores range from 0 to ~50
    - Higher scores indicate greater severity of depression

• Note that:
  - Large overall sample size
  - HRSD score is probably not interpretable to most people
Two motivating studies

• **Study I**: An RCT is conducted to assess the difference between Relenza and placebo in treating flu symptoms. 11 subjects with flu are randomized to Relenza and 11 are randomized to placebo. Subjects are followed until the flu symptoms are gone. Investigators are interested in comparing the number of days subjects suffered from flu symptoms under the different treatments.
  
  - **Treatment** = Relenza or Placebo
  - Two independent groups
  - **Outcome** = Number of days with flu symptoms
  - Fewer days with flu symptoms is better

• **Note that:**
  - SMALL overall sample size
  - Number of days with flu symptoms is probably interpretable to most people

Going right to the p-values

• **Study I – MDD**

![R code output]

Two Sample t-test

data: 0, mdd by T_mdd
t = 3.4433, df = 498, p-value = 0.0001277
alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval:
  0.438745 1.655168
sample estimates:
mean in group 0 mean in group 1
  0.245970 8.191949

• **Study II – Flu**

![R code output]

Two Sample t-test

data: 0, flu by T_flu
t = 2.3508, df = 28, p-value = 0.0291
alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval:
  0.1965211 3.2915784
sample estimates:
mean in group 0 mean in group 1
  0.547649 4.805404

Wow! Eh? Ok! Wow!
Some observations

• We see statistically significant results in both studies.
• Based on p-value
  – Results of MDD Study seem more compelling
  – Are they really?
• Looking at the means in the treatment groups
  – There’s a small (?) difference in the MDD Study
  – There’s a large (?) difference in the Flu Study

P-VALUE IS NOT ENOUGH
What the p-value tells us

• P-values come from hypothesis tests
• In a hypothesis test, we test
  – Null hypothesis: $H_0$ vs.
  – Alternative hypothesis: $H_A$
• $H_0$ typically corresponds to “no treatment effect” in the population
  – We assume $H_0$ is true and use our data to decide if there is **enough evidence** to reject $H_0$ or **not enough evidence** to reject $H_0$
• The p-value summarizes how much evidence we have against $H_0$

What the p-value tells us

• More formally:
  – A p-value is the probability of observing our data or data that are less consistent with $H_0$, assuming $H_0$ is true.
• Example: MDD Study
  – $H_0: \mu_{cit} = \mu_{pla}$
  – Use 2-sample t-test and obtain $T = 3.44$

Interpretation: There is a 0.06% chance that we would see the data that we saw or data more inconsistent with $H_0$ if $H_0$ were actually true.

So it seems that our data are very unlikely to have been observed if $H_0$ is true. We should reject $H_0$. 
p-value limitations

• p-values don’t tell us about
  – How big a difference there is between the two treatments
  – Which treatment yields larger/smaller mean outcome

• Furthermore, p-value depends on sample size
  – Using same sample means and variances from MDD Study

<table>
<thead>
<tr>
<th>n per Group</th>
<th>P-value</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.50</td>
<td>Fail to reject H₀</td>
</tr>
<tr>
<td>25</td>
<td>0.28</td>
<td>Fail to reject H₀</td>
</tr>
<tr>
<td>50</td>
<td>0.13</td>
<td>Fail to reject H₀</td>
</tr>
<tr>
<td>150</td>
<td>0.008</td>
<td>Reject H₀</td>
</tr>
<tr>
<td>250</td>
<td>0.0006</td>
<td>Reject H₀</td>
</tr>
</tbody>
</table>

This was our data

p-value Summary

• P-value measures strength of evidence against H₀

• With large enough sample size, you can almost always find statistically significant differences (small p-values)

• Even if you find statistically significant differences, your findings may not be clinically meaningful
  – MDD Study results were highly statistically significant (p = 0.0006) but are they clinically meaningful?
    • Sample mean HRSD in Citalopram group = 8.19 (sd = 3.37)
    • Sample mean HRSD in Placebo group = 9.25 (sd = 3.47)

• And larger p-values don’t necessarily mean that there is less clinically meaningful difference
  – Flu Study results were also statistically significant (p = 0.03) though not as much as in the MDD Study, yet:
    • Sample mean flu days in Relenza group = 4.80 (sd = 1.98)
    • Sample mean flu days in Placebo group = 6.56 (sd = 1.46)
Final thought on p-value

• Consider one more study
  – RCT for subjects with endometrial cancer
  – Outcome: Survival time after surgery
  – Treatments: EARLY surgery vs. DELAYED surgery
  – p < 0.001

• What can you say?
  – The data are very inconsistent with the H₀
  – No info on size or importance of the difference

• Would like to know:
  – Which treatment better?
  – By how much?
    • Don’t want this to depend on the sample size
    • Want this to take variability of the measurements into account

EFFECT SIZE FOR TWO GROUP COMPARISON
Effect Size

• Effect Size (ES)
  – In a two group setting, ES is the magnitude of the difference between the groups
  – There are several ways that ES can be presented
    • Different ES measures highlight different characteristics
• Note: ES can also be computed for settings with
  – More than two groups
  – No groups
  – Will mention later

Simple idea

• Take the absolute difference.
• Notation:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td>O_A</td>
<td>O_B</td>
</tr>
</tbody>
</table>

• If outcomes are binary (e.g., disease/not disease)
  – Might use $p_A - p_B$
  – Where
    • $p_A = P(\text{Disease} | \text{Treatment A})$
    • $p_B = P(\text{Disease} | \text{Treatment B})$

• If outcomes continuous (e.g., HRSD score)
  – Might use $\mu_A - \mu_B$
  – Where
    • $\mu_A = \text{mean outcome under Treatment A}$
    • $\mu_B = \text{mean outcome under Treatment B}$
ES for two groups

- If outcome is normally distributed:
  - Let’s assume that large values of outcome are good

<table>
<thead>
<tr>
<th>Treatment</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td>OA</td>
<td>OB</td>
</tr>
<tr>
<td>Population Mean</td>
<td>$\mu_A$</td>
<td>$\mu_B$</td>
</tr>
<tr>
<td>Population SD</td>
<td>$\sigma_A$</td>
<td>$\sigma_B$</td>
</tr>
</tbody>
</table>

- Possible ES is:

  \[ d = \frac{\mu_A - \mu_B}{\sigma_{Avg}} \]

  - $d > 0 \leftrightarrow A$ is better than $B$
  - $d = 0 \leftrightarrow A$ same as $B$
  - $d < 0 \leftrightarrow B$ is better than $A$

- Interpretation: Average outcome in group $A$ is $d$ standard deviations larger/smaller than average outcome in group $B$

Estimating $d$

- To estimate $d$: \( \hat{d} = \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{Avg}} \)

- Example 1: MDD Study
  - Sample mean HRSD Citalopram = 8.19 (sd = 3.37)
  - Sample mean HRSD Placebo = 9.25 (sd = 3.47)
  - \( \hat{\sigma}_{Avg} = 3.42 \)

- Example 1I: Flu Study
  - Sample mean flu days Relenza = 4.80 (sd = 1.98)
  - Sample mean flu days Placebo = 6.56 (sd = 1.46)
  - \( \hat{\sigma}_{Avg} = 1.74 \)

\[ \hat{d} = \frac{8.19 - 9.25}{3.42} = -0.31 \]

Mean HRSD in Citalopram Group is 0.31 standard deviations lower than Mean HRSD in Placebo Group

\[ \hat{d} = \frac{4.80 - 6.56}{1.74} = -1.01 \]

Mean number of flu days in Relenza Group is 1.01 standard deviations lower than Mean number of flu days in Placebo Group
ES for two groups

- **Notation:**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td>O_A</td>
<td>O_B</td>
</tr>
</tbody>
</table>

- Let’s assume that large values of outcome are good

- One possible ES is:

\[
\text{AUC} = P(O_A > O_B) + 0.5P(O_A = O_B)
\]

- **AUC is between 0 and 1 (inclusive)**
  - \(AUC > 0.5 \leftrightarrow A \text{ is better than } B\)
  - \(AUC = 0.5 \leftrightarrow A \text{ same as } B\)
  - \(AUC < 0.5 \leftrightarrow B \text{ is better than } A\)

- Think of AUC as a probability of superiority

---

Relationship between AUC and d

- When \(O_A\) and \(O_B\) are both normally distributed:

\[
\text{AUC} = P\left(Z \leq \frac{d}{\sqrt{2}}\right)
\]

- Recall, AUC is probability of superiority
  - AUC gives the probability that a randomly selected person from Group A scores higher than a randomly selected person in Group B
Two More ES Measures

• Closely related to AUC
  – Success Rate Difference
    • SRD = \( P(O_A > O_B) - P(O_A < O_B) \)
    • Between -1 and 1 inclusive
      – SRD > 0 \( \leftrightarrow \) A is better than B
      – SRD = 0 \( \leftrightarrow \) A same as B
      – SRD < 0 \( \leftrightarrow \) B is better than A
  – Number Needed to Treat
    • NNT = \( 1 / SRD \)
    • Between -1 and \( \infty \)
    • See next slide for examples

SRD and NNT Example

• Suppose \( P(O_A > O_B) = 0.52 \) so \( P(O_A < O_B) = 0.48 \)
• SRD = 0.52 – 0.48 = 0.04
• NNT = \( 1 / 0.04 = 25 \)
  – This means I have to treat 25 with Treatment A and 25 with Treatment B to observe 1 additional better outcome in Group A than in Group B
    • CHECK:
      – Number expected better on A: 0.52 x 25 = 13
      – Number expected better on B: 0.48 x 25 = 12
  \{ \text{Difference} = 1 \} 

• Note that SRD and therefore NNT can be negative
  – Example: \( P(O_A > O_B) = 0.48 \) so \( P(O_A < O_B) = 0.52 \)
    – SRD = 0.48 – 0.52 = -0.04
    – NNT = \( 1 / -0.04 = -25 \)
  \{ \text{4 percentage points lower success rate in Group A than in Group B} \}
  \{ \text{It's a bit awkward but the sign just tells you B is better than A} \}
ES measures so far

- So far we have
  - Cohen’s d, AUC, SRD, NNT
- All related
  - If you know one, you know them all
  - Each provides a different and complementary understanding of the difference between Group A and Group B
  - Help to understand clinical relevance of differences

Examples

- Consider Study 1: MDD Study
  - Recall $p = 0.0006$
  - ES Estimates:

  $d = -0.31$
  $AUC = 0.41$
  $SRD = -0.18$
  $NNT = -5.56$

  Mean HRSD in Citalopram Group is 0.31 standard deviations lower than Mean HRSD in Placebo Group.

  Probability that a randomly selected subject from the Citalopram Group has a higher HRSD score than a randomly selected subject from the Placebo Group.

  Alternatively, there is a 59% probability that a randomly selected subject from the Citalopram Group will have a lower HRSD score compared to a randomly selected subject from the Placebo Group.

  There is an 18 percentage points greater chance of a randomly selected outcome from the Placebo group being higher than a randomly selected outcome from the Citalopram Group.

  Calculation: $0.41 - 0.59 = -0.18$

  Need to treat 5.56 people in each Group to see one more larger outcome value in the Placebo Group compared to the Citalopram Group.

  Check:
  - Citalopram: $0.41 \times 5.56 = 2.28$
  - Placebo: $0.59 \times 5.56 = 3.28$
### Examples

- **Consider Study II: Flu Study**
  - Recall \( p = 0.0291 \)
  - ES Estimates:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d )</td>
<td>(-1.01)</td>
<td>Mean number of flu days in the Relenza Group is 1.01 standard deviations lower than mean number of flu days in the Placebo Group.</td>
</tr>
<tr>
<td>( AUC )</td>
<td>(0.24)</td>
<td>Probability that a randomly selected subject from the Relenza Group has higher number of flu days than a randomly selected subject from the Placebo Group.</td>
</tr>
<tr>
<td>( SRD )</td>
<td>(-0.52)</td>
<td>There is a 52 percentage points greater chance of a randomly selected outcome from the Placebo group being higher than a randomly selected outcome from the Relenza Group.</td>
</tr>
<tr>
<td>( NNT )</td>
<td>(-1.91)</td>
<td>Need to treat 1.91 people in each Group to see one more larger outcome value in the Placebo Group compared to the Relenza Group.</td>
</tr>
</tbody>
</table>

Alternatively:
- **Based on ES measures**
  - Are the Flu Study results looking more “impressive” than the MDD Study results?
  - Are there guidelines for interpreting ES values?
    - Yes!
    - Use them critically – not blindly
**Small, medium, large ES**

- These are suggestions

“In this context, Cohen suggested that \( d = 0.2, 0.5, \) and 0.8 are “small,” “medium,” and “large” on the basis of his experience as a statistician, but he also warned that these were only “rules of thumb” (Cohen 1988). An effect size of .2 could be “large” in some contexts (e.g., Salk vaccine to prevent polio in the general pediatric population) or an effect size of .8 “small” in others (e.g., a drug treatment for depression that substantially increases the risk of suicide)."

<table>
<thead>
<tr>
<th>Cohen’s (d)</th>
<th>(r)</th>
<th>AUC</th>
<th>SRD</th>
<th>NNT</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.2</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>0.5</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>0.8</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Size of Treatment Effects and Their Importance to Clinical Research and Practice
Kraemer, Helena Chmura et al., Biological Psychiatry, Volume 59, Issue 11, 990 - 996

**OTHER EFFECT SIZE MEASURES**
Beyond comparing 2 groups

- Many instances in clinical research where you aren’t just comparing 2 groups with respect to a continuous outcome

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Exposure</th>
<th>Test/Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary (e.g., cancer status)</td>
<td>Binary (e.g., smoking status)</td>
<td>Chi-Squared Test</td>
</tr>
<tr>
<td>Continuous (e.g., time to headache relief)</td>
<td>Categorical (e.g., treatment A, B, or C)</td>
<td>ANOVA</td>
</tr>
<tr>
<td>Continuous (e.g., MRI connectivity measure)</td>
<td>Continuous (e.g., age)</td>
<td>Correlation</td>
</tr>
<tr>
<td>Continuous (e.g., weight)</td>
<td>Multiple Exposures (e.g., age, gender, diet program)</td>
<td>Multiple Linear Regression</td>
</tr>
<tr>
<td>Binary (e.g., drug use relapse)</td>
<td>Multiple Exposures (e.g., age, gender, trt program)</td>
<td>Multiple Logistic Regression</td>
</tr>
</tbody>
</table>

Some additional ES measures

<table>
<thead>
<tr>
<th>Index</th>
<th>Description</th>
<th>Effect Size</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohen’s d</td>
<td></td>
<td></td>
<td>Can be used at planning stage to find the sample size required for sufficient power for your study</td>
</tr>
<tr>
<td>Odds ratio (OR)</td>
<td></td>
<td></td>
<td>For binary outcome variables: Comparing odds of outcome occurring from one intervention vs another</td>
</tr>
<tr>
<td>Relative risk or risk ratio (RR)</td>
<td></td>
<td></td>
<td>Comparing probabilities of outcome occurring from one intervention to another</td>
</tr>
<tr>
<td>Measures of association</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson’s correlation</td>
<td></td>
<td></td>
<td>Measures the degree of linear relationship between two quantitative variables</td>
</tr>
<tr>
<td>r² coefficient of determination</td>
<td></td>
<td></td>
<td>Proportion of variance in one variable explained by the other</td>
</tr>
</tbody>
</table>

In addition

- We can also compute confidence intervals for ES measures based on the data that we collect
  - R packages: `effsize`, `compute.es`

- ES plays a critical role in power/sample size calculations needed to propose future studies
  - Relationships between:
    - Significance level ($\alpha$)
    - Sample size
    - Power ($1 - \beta$)
    - Effect Size

\{ knowing any 3 determines the fourth \}

SUMMARY
Summary

• p-values don’t tell the whole story
  – Actually tell us very little
• Size of the p-value tells us nothing about clinical importance
• Effect sizes help us to understand relationships that we see in our data
  – Should be reported along with p-values (and/or confidence intervals)
• With 2 groups and a normally distributed outcome
  – Can use Cohen’s d, AUC, SRD, and NNT
• There are many other useful ES measures for a variety of settings
• Effect sizes are not only useful for understanding associations, they are necessary for power/sample size calculations for future studies