Effect Size and Statistical Power

Joseph Stevens, Ph.D., University of Oregon
(541) 346-2445, stevensj@uoregon.edu
© Stevens, 2007

An Introductory Problem or Two:

Which Study is Stronger?

Answer: Study B

Study A: \( t(398) = 2.30, p = .022 \)
\( \omega^2 \) for Study A = .01

Study B: \( t(88) = 2.30, p = .024 \)
\( \omega^2 \) for Study B = .05

Examples inspired by Rosenthal & Gaito (1963)

Study C Shows a Highly Significant Result

Study C: \( F = 63.62, p < .0000001 \)
\( \eta^2 \) for Study C = .01, \( N = 6,300 \)

Study D: \( F = 5.40, p = .049 \)
\( \eta^2 \) for Study D = .40, \( N = 10 \)

Correct interpretation of statistical results requires consideration of statistical significance, effect size, and statistical power

Three Fundamental Questions Asked in Science

Is there a relationship? Answered by Null Hypothesis Significance Tests (NHST; e.g., \( t \) tests, \( F \) tests, \( \chi^2 \), \( p \)-values, etc.)

What kind of relationship? Answered by testing if relationship is linear, curvilinear, etc.

How strong is the relationship? Answered by effect size measures, not NHST’s (e.g., \( R^2 \), \( r^2 \), \( \eta^2 \), \( \omega^2 \), Cohen’s \( d \))

The Logic of Inferential Statistics

Three Distributions Used in Inferential Statistics:

- **Population**: the entire universe of individuals we are interested in studying (\( \mu, \sigma, \infty \))

- **Sample**: the selected subgroup that is actually observed and measured (\( \bar{x}, s, N \))

- **Sampling Distribution of the Statistic**: A theoretical distribution that describes how a statistic behaves across a large number of samples (\( \bar{x}, s, \infty \))

The Three Distributions Used in Inferential Statistics

I. Population

II. Sample

III. Sampling Distribution of the Statistic
The NHST Decision Model (based on the sampling distribution of the statistic)

<table>
<thead>
<tr>
<th>Statistical Decision</th>
<th>True State</th>
<th>H₀ True</th>
<th>H₀ False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail to Reject H₀</td>
<td>Correct Decision, (1 – α)</td>
<td>Type II Error (β), False Negative</td>
<td></td>
</tr>
<tr>
<td>Reject H₀</td>
<td>Type I Error (α), False Positive</td>
<td>Correct Decision (1 – β), Statistical Power</td>
<td></td>
</tr>
</tbody>
</table>

True State

The value of α is set by convention which also determines 1 - α

And, if H₀ is really true, then β = 0

Reject H₀

But if H₀ is false, what are the values of β and (1-β)?

What if H₀ is False?
- If the null hypothesis is false, the sampling distribution and model just considered is incorrect
- In that case, a different sampling distribution describes the true state of affairs, the noncentral distribution
  - In fact there is a family of sampling distributions when the null is false that depend on just how large an effect is present
  - The size of the difference between the central and noncentral distributions is described by a noncentrality parameter

Central and Noncentral Distributions

The noncentrality parameter represents the lack of overlap or displacement of the two distributions that results from a true difference between groups or nonzero relationship between variables

Assume an example using the t distribution with Cohen’s d = .4

Note the disparity between the central and noncentral sampling distributions

Note: Sampling distributions are called Central Distributions when H₀ is true.
The Relationship Between Effect Size and Statistical Significance

- It should be apparent that statistical significance depends on the size of the effect (e.g., the noncentrality parameter).
- And, statistical significance also depends on the size of the study (N).
- Statistical significance is the product of these two components.

<table>
<thead>
<tr>
<th>Test Results</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>= Effect Size</td>
<td>X Size of Study</td>
</tr>
</tbody>
</table>

\[ t = \frac{r}{\sqrt{1 - r^2}} \times \sqrt{df} \]

\[ t = \frac{(\bar{X}_1 - \bar{X}_2)}{s} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \]
Significance Test Results = Effect Size X Size of Study

\[ F = \frac{r^2}{1 - r^2} \times \text{df} \]

\[ F = \frac{\eta^2}{1 - \eta^2} \times \frac{\text{df}_{\text{error}}}{\text{df}_{\text{means}}} \]

To make correct interpretations, additional information beyond statistical significance is needed.

When results are statistically significant, it is very important to estimate effect size to determine the magnitude of results.

Two Kinds of Metric for Measuring the Magnitude of Effects

- Standardized Difference Measures – Express the size of group difference in standard deviation units (e.g., Cohen’s \( d \))
- Strength of Association Measures – Express magnitude of effect as a proportion or percentage (e.g., \( r^2 \), \( \eta^2 \), \( \omega^2 \))

Strength of Association Measures

- Pearson’s \( r \)
- Multiple \( R \)
- Multivariate
  - Canonical \( r \)
  - Wilk’s Lambda \((1 - \Lambda)\)
- Effect size can be interpreted in units of \( r \) (see BESD below) or after squaring and multiplying by 100 as Percent Shared Variance (PSV)

\[ \text{PSV} = r^2 \times 100 \]

Cohen also uses \( f^2 \) as a metric of effect size

This is easily expressed as \( R^2 \) or \( \eta^2 \):

\[ f^2 = \frac{R^2}{(1 - R^2)} \]

\[ f^2 = \frac{\eta^2}{(1 - \eta^2)} \]
Strength of Association Measures: $\omega^2$

**Omega Squared for an independent $t$-test:**

$$\omega^2 = \frac{(t^2 - 1)}{(t^2 + N_1 + N_2 - 1)}$$

Example:

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>65.50</td>
<td>69.00</td>
</tr>
<tr>
<td>Variance</td>
<td>20.69</td>
<td>28.96</td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

$t = \frac{65.5 - 69}{1.29} = -2.71$

$$\omega^2 = \frac{(-2.71)^2 - 1}{((-2.71)^2 + 30 + 30 - 1)} = 0.096,$$  about 10% shared variance

---

**Omega Squared for a one-factor ANOVA:**

$$\omega^2 = \frac{[SS_{between} - (a-1)(MS_{Residual})]}{(SS_{Total} + MS_{Residual})}$$

Example:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3.61</td>
<td>1</td>
<td>3.61</td>
<td>2.76</td>
<td>.101</td>
</tr>
<tr>
<td>B</td>
<td>13.94</td>
<td>3</td>
<td>4.65</td>
<td>3.55</td>
<td>.019</td>
</tr>
<tr>
<td>AB</td>
<td>12.34</td>
<td>3</td>
<td>4.11</td>
<td>3.14</td>
<td>.030</td>
</tr>
<tr>
<td>Residual</td>
<td>94.30</td>
<td>72</td>
<td>1.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>757.00</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Omega Squared for a two-factor ANOVA:**

$$\omega^2 = \frac{[SS_A - (a-1)(MS_{Residual})]}{(SS_{Total} + MS_{Residual})}$$

$$\omega^2 = \frac{[SS_B - (b-1)(MS_{Residual})]}{(SS_{Total} + MS_{Residual})}$$

$$\omega^2 = \frac{[SS_{AB} - (a-1)(b-1)(MS_{Residual})]}{(SS_{Total} + MS_{Residual})}$$

---

**Strength of Association Measures: $\eta^2$**

$$\eta^2 = \frac{SS_{Effect}}{SS_{Total}}$$

An alternative measure is partial eta squared:

$$\eta^2_p = \frac{SS_{Effect}}{(SS_{Effect} + SS_{Residual})}$$

Note: Partial eta may sum to more than 100% in multifactor designs
Strength of Association Measures: $\eta_p^2$

An alternative formula using only $F$ and df:

$$\eta_p^2 = \frac{(F)(df_{\text{effect}})}{(F)(df_{\text{effect}}) + df_{\text{residual}}}$$

Example using the interaction effect from above:

$$\eta_p^2 = \frac{(3.14)(3)}{(3.14)(3) + 72} = .116$$

Comparing Strength of Association Measures

Note the problems with partials:
- Different denominator for each effect
- Partials may sum to more than 100% in multifactor designs

$$\eta_p^2 \leq \eta^2 \leq \eta^2_p$$

Group Difference Indices

- There are a variety of indices that measure the extent of the difference between groups
- Cohen’s $d$ is the most widely used index (two groups only)
- Generalization of Cohen’s to multiple groups is sometimes called $\delta$, but there is great variation in notation
- Hedges’ $g$ (uses pooled sample standard deviations)
- For multivariate, Mahalanobis’ $D^2$

The Standardized Mean Difference: Cohen’s $d$

$$d = \frac{(X_1 - X_2)}{s_{\text{pooled}}}$$

$$s_{\text{pooled}} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

Example: Group 1 Group 2

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>65.50</td>
<td>69.00</td>
</tr>
<tr>
<td>Variance</td>
<td>20.69</td>
<td>28.96</td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

$$d = \frac{(65.50 - 69.00)}{4.98} = -0.70$$

Interpreting Effect Size Results (How big is big?)

- There is no simple answer to “How large should an effect size be?”
- The question begs another: “For what purpose?”
- The answer does not depend directly on statistical considerations but on the utility, impact, and costs and benefits of the results
Interpreting Effect Size Results

Cohen’s “Rules-of-Thumb”
- standardized mean difference effect size (Cohen’s $d$)
  - small = 0.20
  - medium = 0.50
- correlation coefficient (Pearson’s $r$)
  - small = 0.10
  - medium = 0.30
  - large = 0.50

“If people interpreted effect sizes (using fixed benchmarks) with the same rigidity that $p = .05$ has been used in statistical testing, we would merely be being stupid in another metric” (Thompson, 2001; pp. 82-83).

Are Small Effects Unimportant?

<table>
<thead>
<tr>
<th>Condition</th>
<th>Alive</th>
<th>Dead</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>66</td>
<td>34</td>
<td>100</td>
</tr>
<tr>
<td>Control</td>
<td>34</td>
<td>66</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>


Confidence Intervals for Effect Size

95% Confidence
interval for Cohen’s $d$

Cohen’s $d = -.70$ (same example as slide 35)


42

Statistical Power

- Statistical power, the probability of detecting a result when it is present
- Often the concern is “How many participants do I need?”
- While estimating $N$ is important, a more productive focus may be on effect size and design planning
- How can I strengthen the research?
Factors Affecting Statistical Power

- Sample Size
- Effect Size
- Alpha level
- Unexplained Variance
- Design Effects

Effect of Sample Size on Statistical Power

All things equal, sample size increases statistical power at a geometric rate (in simple designs)

- This is accomplished primarily through reduction of the standard error of the sampling distribution
- With large samples, inferential statistics are very powerful at detecting very small relationships or very small differences between groups (even trivial ones)
- With small samples, larger relationships or differences are needed to be detectable

**As an example, if the estimated population standard deviation was 10 and sample size was 4 then:**

\[ \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{4}} = 5 \]

But if sample was 16 (4 times larger) then the standard error is 2.5 (smaller by half):

\[ \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{16}} = 2.5 \]

Consider the following example with \( N = 10 \), note that power = .21

Versus a second example with \( N = 30 \), note that power = .56

Impact of Sample Size on Statistical Power
Impact of Effect Size on Statistical Power

- One-tailed tests are more powerful than two-tailed tests
  - Require clear a priori rationale
  - Requires willingness to ignore results in the wrong direction
  - Only possible with certain statistical tests (e.g., t but not F)
- Larger alpha values more powerful (e.g., $p < .10$)
  - May be difficult to convince reviewers
  - Can be justified well in many program evaluation contexts (when only one direction of outcome is relevant)
  - Justifiable with small sample size, small cluster size, or if, a priori, effect size is known to be small

Impact of Sample and Effect Size on Statistical Power

Effect of Alpha Level on Statistical Power

- One-tailed tests are more powerful than two-tailed tests
  - Require clear a priori rationale
  - Requires willingness to ignore results in the wrong direction
  - Only possible with certain statistical tests (e.g., t but not F)
- Larger alpha values more powerful (e.g., $p < .10$)
  - May be difficult to convince reviewers
  - Can be justified well in many program evaluation contexts (when only one direction of outcome is relevant)
  - Justifiable with small sample size, small cluster size, or if, a priori, effect size is known to be small

Effect of Unexplained Variance on Statistical Power

- Terminology: “error” versus unexplained or residual
- Residual variance reduces power
  - Anything that decreases residual variance, increases power (e.g., more homogeneous participants, additional explanatory variables, etc.)
- Unreliability of measurement contributes to residual variance
- Treatment infidelity contributes to residual variance
Effect of Design Features on Statistical Power

- Stronger treatments!
- Blocking and matching
- Repeated measures
- Focused tests ($df = 1$)
- Intraclass correlation
- Statistical control, use of covariates
- Restriction of range (IV and DV)
- Measurement validity (IV and DV)

Effect of Design Features on Statistical Power

- Multicollinearity (and restriction of range)

\[ s_{b,1} = \sqrt{ \frac{s^2_{\text{total}}}{\sum \tau^2_i (1 - r^2_i)} } \]

- Statistical model misspecification
  - Linearity, curvilinearity,…
  - Omission of relevant variables
  - Inclusion of irrelevant variables

Options for Estimating Statistical Power

- Cohen’s tables
- Statistical Software like SAS and SPSS using syntax files
- Web calculators
- Specialized software like G*Power, Optimal Design, ESCI, nQuery

Estimating Statistical Power

- Base parameters on best information available
- Don’t overestimate effect size or underestimate residual variance or ICC
- Consider alternative scenarios
  - What kind of parameter values might occur in the research?
  - Estimate for a variety of selected parameter combinations
  - Consider worst cases (easier to plan than recover)

Recommendations for Study Planning

- Greater attention to study design features
- Explore the implications of research design features on power
- Base power estimation on:
  - Prior research
  - Pilot studies
  - Plausible assumptions
  - Thought experiments
  - Cost/benefit analysis

Power in Multisite and Cluster Randomized Studies

- More complex designs involving data that are arranged in inherent hierarchies or levels
- Much educational and social science data is organized in a multilevel or nested structure
  - Students within schools
  - Children within families
  - Patients within physicians
  - Treatments within sites
  - Measurement occasions within individuals
Power in Multisite and Cluster Randomized Studies

Factors affecting statistical power
- Intraclass Correlation (ICC)
- Number of participants per cluster (N)
- Number of clusters (J)
- Between vs. within cluster variance
- Treatment variability across clusters
- Other factors as discussed above

Intraclass Correlation Coefficient ($\rho$)

Total $\sigma^2_Y = \tau^2 + \sigma^2$

$$ICC = \frac{\text{population variance between units}}{\text{total variance}} = \frac{\tau^2}{(\tau^2 + \sigma^2)}$$

As ICC approaches 0, multilevel modeling is not needed and power is the same as a non-nested design, but even small values of ICC can impact power

Intraclass Correlation ($\rho_i$)

- ICC varies with outcome and with type of group and participants
- Small groups that may be more homogenous (e.g., classrooms) are likely to have larger ICCs than large groups with more heterogeneity (e.g., schools or districts)
- What size of ICCs are common?
  - Concentrated between 0.01 and 0.05 for much social science research (Bloom, 2006)
  - Between 0.05 and 0.15 for school achievement (Spybrook et al., 2006)
- The guideline of 0.05 to 0.15 is more consistent with the values of covariate adjusted intraclass correlations; unconditional ICCs may be larger (roughly 0.15 to 0.25; Hedges & Hedberg, in press)
- “It is unusual for a GRT to have adequate power with fewer than 8 to 10 groups per condition” (Murray et al., 2004)

Intraclass Correlation ($\rho$)

- ICC becomes important in research design when:
  - Random assignment is accomplished at the group level
  - Multistage sampling designs are used
  - Group level predictors or covariates are used
- If there is little difference from one group to another (ICC nears zero), power is similar to the total sample size ignoring the clustering of groups
- The more groups differ (ICC is nonzero), effective sample size for power approaches the number of groups rather than the total number of participants

Relationship of ICC and power

- $\rho$ impacts the effective sample size for power calculations

- As ICC increases, the effective sample size decreases

- Important to consider ICC when planning multisite or cluster randomized studies
Relationship of ICC, Effect Size, Number of Clusters and Power

Relationship of ICC, Effect Size, Number of Clusters and Power When J is Small

Effect of Cluster Size (n)

Effect of Number of Clusters (J)

The number of clusters has a stronger influence on power than the cluster size.

Note the difference in power for nj = 500 arranged as 50 per 10 vs. nj = 500 arranged as 25 per 20 clusters.

Difference due to cluster size
Ignoring Hierarchical Structure vs. Multilevel Modeling

Variance of the treatment effect across clusters

\[ y = \frac{(\sigma^2 + \sigma^2/n)}{nJ} \]

The number of clusters has a stronger influence on power than the cluster size as ICC departs from 0

- The standard error of the main effect of treatment is:

\[ SE(\hat{\gamma}_0) = \sqrt{\frac{4(\rho + (1-\rho)/n)}{J}} \]

- As \( \rho \) increases, the effect of \( \sigma \) decreases
- If clusters are variable (\( \rho \) is large), more power is gained by increasing the number of clusters sampled than by increasing \( \sigma \)

The Group Effect Multiplier

<table>
<thead>
<tr>
<th>ICC (( \rho ))</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.01</td>
<td>1.04</td>
<td>1.09</td>
<td>1.22</td>
<td>1.41</td>
<td>1.73</td>
<td>2.48</td>
</tr>
<tr>
<td>0.02</td>
<td>1.09</td>
<td>1.17</td>
<td>1.41</td>
<td>1.73</td>
<td>2.23</td>
<td>3.31</td>
</tr>
<tr>
<td>0.03</td>
<td>1.13</td>
<td>1.25</td>
<td>1.57</td>
<td>1.99</td>
<td>2.64</td>
<td>4.00</td>
</tr>
<tr>
<td>0.04</td>
<td>1.17</td>
<td>1.33</td>
<td>1.72</td>
<td>2.23</td>
<td>2.99</td>
<td>4.58</td>
</tr>
<tr>
<td>0.05</td>
<td>1.20</td>
<td>1.40</td>
<td>1.86</td>
<td>2.44</td>
<td>3.31</td>
<td>5.09</td>
</tr>
<tr>
<td>0.06</td>
<td>1.24</td>
<td>1.46</td>
<td>1.98</td>
<td>2.63</td>
<td>3.60</td>
<td>5.56</td>
</tr>
<tr>
<td>0.07</td>
<td>1.28</td>
<td>1.53</td>
<td>2.10</td>
<td>2.82</td>
<td>3.86</td>
<td>5.99</td>
</tr>
<tr>
<td>0.08</td>
<td>1.31</td>
<td>1.59</td>
<td>2.22</td>
<td>2.99</td>
<td>4.11</td>
<td>6.40</td>
</tr>
<tr>
<td>0.09</td>
<td>1.35</td>
<td>1.63</td>
<td>2.33</td>
<td>3.15</td>
<td>4.35</td>
<td>6.78</td>
</tr>
<tr>
<td>0.10</td>
<td>1.38</td>
<td>1.70</td>
<td>2.43</td>
<td>3.30</td>
<td>4.57</td>
<td>7.13</td>
</tr>
<tr>
<td>0.20</td>
<td>1.67</td>
<td>2.19</td>
<td>3.29</td>
<td>4.56</td>
<td>6.39</td>
<td>10.04</td>
</tr>
</tbody>
</table>

Note: The group effect multiplier equals \( \frac{\rho}{1-\rho} \); table from Bloom (2006).

The Minimum Detectable Effect Expressed as a Multiple of the Standard Error

<table>
<thead>
<tr>
<th>Number of groups (J)</th>
<th>Two-tailed test</th>
<th>One-tailed test</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5.36</td>
<td>5.08</td>
</tr>
<tr>
<td>6</td>
<td>5.72</td>
<td>5.07</td>
</tr>
<tr>
<td>8</td>
<td>5.35</td>
<td>2.65</td>
</tr>
<tr>
<td>10</td>
<td>5.20</td>
<td>2.75</td>
</tr>
<tr>
<td>12</td>
<td>5.11</td>
<td>2.69</td>
</tr>
<tr>
<td>14</td>
<td>5.05</td>
<td>2.66</td>
</tr>
<tr>
<td>16</td>
<td>5.01</td>
<td>2.65</td>
</tr>
<tr>
<td>18</td>
<td>2.99</td>
<td>2.61</td>
</tr>
<tr>
<td>20</td>
<td>2.96</td>
<td>2.60</td>
</tr>
<tr>
<td>30</td>
<td>2.90</td>
<td>2.56</td>
</tr>
<tr>
<td>40</td>
<td>2.87</td>
<td>2.54</td>
</tr>
<tr>
<td>60</td>
<td>2.85</td>
<td>2.52</td>
</tr>
<tr>
<td>120</td>
<td>2.83</td>
<td>2.50</td>
</tr>
</tbody>
</table>

Note: The group effect multiplier shown here are for the difference between the mean program group outcome and the mean control group outcome, assuming equal variances for the groups, a significance level of .05, and a power level of .80; table from Bloom (2006).
The Minimum Detectable Effect Size

Intraclass correlation ($\rho_I$) = 0.01

<table>
<thead>
<tr>
<th>Number of groups (J)</th>
<th>Randomized group size (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>1.77</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>0.78</td>
</tr>
<tr>
<td>10</td>
<td>0.67</td>
</tr>
<tr>
<td>20</td>
<td>0.44</td>
</tr>
<tr>
<td>30</td>
<td>0.33</td>
</tr>
<tr>
<td>40</td>
<td>0.30</td>
</tr>
<tr>
<td>60</td>
<td>0.24</td>
</tr>
<tr>
<td>120</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Note: The minimum detectable effect sizes shown here are for a two-tailed hypothesis test, assuming a significance level of .05, a power level of .80, and randomization of half the groups to the program; table from Bloom (2006).

The Minimum Detectable Effect Size

Intraclass correlation ($\rho_I$) = 0.05

<table>
<thead>
<tr>
<th>Number of groups (J)</th>
<th>Randomized group size (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>2.04</td>
</tr>
<tr>
<td>6</td>
<td>1.16</td>
</tr>
<tr>
<td>8</td>
<td>0.90</td>
</tr>
<tr>
<td>10</td>
<td>0.77</td>
</tr>
<tr>
<td>20</td>
<td>0.50</td>
</tr>
<tr>
<td>30</td>
<td>0.40</td>
</tr>
<tr>
<td>40</td>
<td>0.35</td>
</tr>
<tr>
<td>60</td>
<td>0.28</td>
</tr>
<tr>
<td>120</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Note: The minimum detectable effect sizes shown here are for a two-tailed hypothesis test, assuming a significance level of .05, a power level of .80, and randomization of half the groups to the program; table from Bloom (2006).

The Minimum Detectable Effect Size

Intraclass correlation ($\rho_I$) = 0.10

<table>
<thead>
<tr>
<th>Number of groups (J)</th>
<th>Randomized group size (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>2.34</td>
</tr>
<tr>
<td>6</td>
<td>1.32</td>
</tr>
<tr>
<td>8</td>
<td>1.03</td>
</tr>
<tr>
<td>10</td>
<td>0.88</td>
</tr>
<tr>
<td>20</td>
<td>0.58</td>
</tr>
<tr>
<td>30</td>
<td>0.46</td>
</tr>
<tr>
<td>40</td>
<td>0.40</td>
</tr>
<tr>
<td>60</td>
<td>0.32</td>
</tr>
<tr>
<td>120</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Note: The minimum detectable effect sizes shown here are for a two-tailed hypothesis test, assuming a significance level of .05, a power level of .80, and randomization of half the groups to the program; table from Bloom (2006).

Using G*Power

Free software for power estimation available at:
http://www.psych.uni-duesseldorf.de/aap/gpower3/download-and-register

Estimates power for a variety of situations including $t$-tests, $F$-tests, and $\chi^2$

G*Power

Examples using G*Power

Luft & Vidoni (2002) examined preservice teachers' knowledge about school to career transitions before and after a teacher internship. Some of the obtained results were:

<table>
<thead>
<tr>
<th>Knowledge about:</th>
<th>Before</th>
<th>After</th>
<th>t</th>
<th>p</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing</td>
<td>2.92</td>
<td>1.44</td>
<td>-2.23</td>
<td>.06</td>
<td>.59</td>
</tr>
<tr>
<td>Use of Hands-on activities</td>
<td>4.58</td>
<td>2.25</td>
<td>-1.06</td>
<td>.34</td>
<td>.71</td>
</tr>
<tr>
<td>Class assignments</td>
<td>3.67</td>
<td>4.08</td>
<td>-1.42</td>
<td>.10</td>
<td>.56</td>
</tr>
</tbody>
</table>

Twelve students participated in the study and completed the pre and post testing.
Next calculate an effect size based on the supplied table information:

Example 2. Using the same information as example 1, determine the necessary sample size to achieve a power of .80

Using the Optimal Design Software

- The Optimal Design Software can also be used to estimate power in a variety of situations.
- The particular strength of this software is its application to multilevel situations involving cluster randomization or multisite designs.
- Available at: [http://sitemaker.umich.edu/group-based/optimal_design_software](http://sitemaker.umich.edu/group-based/optimal_design_software)
- Optimal Design
Now enter values to produce power estimates.

Use $\alpha = .05$, $n = 10$, $\delta = .5$, and $\rho = .05$.

Note that if you mouse over the power curve, exact values are displayed.

Range and legend for axes can also be modified.

Now explore the use of OD for examining power as a function of $n$, $\rho$, $\delta$, and $R^2$.

The OD software can also be used to determine the best combination of design features under cost constraints.

Enter values of $10,000$ Total budget, $400$ per cluster, $20$ per member, $\rho = .03$, and $\delta = .4$; then compute.

Optimal Design

For the given budget, $n = 21$, $J = 12$ and power is .62.

Note the increase in both $n$ and power.

Note the ratio of $n$ to $J$ given the higher ICC.

One Last Example: Multisite CRT

- The primary rationale in this approach is to extend the idea of blocking to the multilevel situation.
- Clusters are assigned to blocks with other similar clusters and then randomly assigned to treatment.
- Blocking creates greater homogeneity and less residual variance, thereby increasing power.
- For example, schools are collected into blocks based on whether school composition is low, medium, or high SES.
- Schools are within each block are randomly assigned to treatment.
- Between school SES variability is controlled by the blocking.
Multisite CRT

Two additional parameters are used in estimation:

- Number of sites or blocks, K
- The effect size variability, \( \sigma_\delta^2 \)
- \( \sigma_\delta^2 \) represents the variability of effect size from one cluster to another within a site
- This variability represents within site replications of the study

Example:

- 5 cities, 12 schools per city, \( d = .4 \), ICC = .12, \( \sigma_\delta^2 = .01 \), blocking accounts for 50% of the variation in the outcome

Applications

- For the remainder of the workshop you may
  - complete exercises on power estimation
  - calculate power estimates for your own research
- Exercises can be downloaded from:
  - [http://www.uoregon.edu/~stevensj/workshops/exercises.pdf](http://www.uoregon.edu/~stevensj/workshops/exercises.pdf)
- When you finish the exercises, you can obtain answers at:
  - [http://www.uoregon.edu/~stevensj/workshops/answers.pdf](http://www.uoregon.edu/~stevensj/workshops/answers.pdf)
- Discussion as time permits

Bibliography


Cohen, J. (1994). The earth is round \( (p < .05) \). American Psychologist, 49, 597-603.


