Concepts from previous lectures

• effect size
• confidence interval
• Type I error
• Winner’s curse

Reproducibility Crisis

Lykken DT, Psychol Bulletin, 1968

“Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published.” - David Lykken

# Cleaning Up Science

A list of scientists have been forced recently for making up data and falsifying statistics. The case involves a Harvard professor who three years and worked with another of which avoid psychology who made up results by the

Meta-Analysis

- Replication is essential for scientific progress
- But replications often yield different and even conflicting results
- How do we integrate & compare findings across studies?
- Meta-analysis: statistical methods/procedures for integrating and contrasting results from different studies
- Immediate problem: how do we integrate studies that may have used different dependent measures, sample sizes, and/or statistical methods to evaluate effects?
Importance of Measures of Effect Size

- Effect size is a measure of the magnitude of effects that is relatively invariant to changes in sample size
- Moreover, measures of effect size typically are dimensionless:
  - express "effect" relative to error variance
  - makes it possible to compare different dependent measures
    - (e.g., accuracy vs. response latency)
- But there are many different kinds of effect size measures...

Effect Size for 2 Means

\[ d = \frac{X_1 - X_2}{s_{pooled}} \]

for independent groups

\[ d = \frac{X_1 - X_2}{s_{control}} \]

for independent treatment & control groups

\[ d = \frac{X_{post} - X_{pre}}{s_{diff} \sqrt{2(1 - r)}} \]

for paired measures (pre vs post treatment)

Effect Size for more than 2 means

Cohen’s \( f^2 \) and \( f \)

for \( a \) independent groups \((a > 2)\)

\[ f^2 = \frac{\sigma_{\mu}^2}{\sigma_{error}^2} = \frac{\left(\sum_{j=1}^{a} (X_j - \bar{X}_{gm})^2\right)}{a} \]

ratio of variance of population means divided by population error variance

\[ f = \sqrt{f^2} \]

Effect Sizes for Contingency Tables

<table>
<thead>
<tr>
<th></th>
<th>Recovered</th>
<th>Died</th>
<th>N</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated Group</td>
<td>A</td>
<td>B</td>
<td>A+B</td>
<td>( \frac{A}{B} )</td>
</tr>
<tr>
<td>Control Group</td>
<td>C</td>
<td>D</td>
<td>C+D</td>
<td>( \frac{C}{D} )</td>
</tr>
</tbody>
</table>

\[ OR = \frac{A}{C} / \frac{B}{D} = \frac{A \times D}{B \times C} \]

Log Odds Ratio = ln(OR)

Relative Risk = Risk Ratio

\[ RR = \frac{A}{A + B} / \frac{C}{C + D} = \frac{A(C + D)}{B(A + B)} \]
Effect Sizes for Correlations

Fisher’s transformation:

\[ r' = 0.5 \times \ln \left( \frac{1 + r}{1 - r} \right) \]

\( r \) itself is often used as a measure of effect size, but the sampling distribution of \( r \) is skewed whereas the sampling distribution of \( r' \) is approximately normal.

Integrating Different Effect Size Measures

- Different studies report different measures of effect size
- Fortunately, it is easy to convert one measure of effect size to another:

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

- Hence, results from various studies can be translated into a common metric (i.e., a single type of effect size)
  - which makes it possible to integrate (average) and/or compare results across studies

Example: Pygmalion Effect

Robert Abelson, Statistics as Principled Argument, pp 150-152

- Pygmalion Effect: Higher expectations lead to higher performance
- Rosenthal & Jacobson (1968)
  - told elementary school teachers that some students had been identified with a special test as being likely to excel
  - in fact, students were selected at random (there was no test)
- hypothesis: teachers would have higher expectations for these students which would initiate a self-fulfilling prophecy
  - teachers would be more attentive to these students who in turn would exhibit greater confidence & involvement in class
- at end of year, found that identified students scored higher on tests of mental ability (e.g., IQ)
  - the study inspired many follow-up studies
  - is the effect “real”?

Example: Pygmalion Effect

Robert Abelson, Statistics as Principled Argument, pp 150-152

  - nearly half (8 of 18) found negative effect (wrong direction)
  - but distribution of Cohen’s d was positively skewed
    - median \( d = 0.035 \)
    - mean \( d = 0.109 \)
  - mean effect size differed significantly from zero

<table>
<thead>
<tr>
<th>Stem-and-Leaf Plot of Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE 7.3</td>
</tr>
<tr>
<td>A Meta-Analysis of Pygmalion Effects on IQ Scores—</td>
</tr>
<tr>
<td>Stem-and-Leaf Diagram of Effect Sizes in 18 Studies</td>
</tr>
</tbody>
</table>

| 0.0 | 0.25 |
| 0.1 | 0.468 |
| 0.2 | 0.6432227 |
| 0.3 | -0.133 |

*Note: Data adapted from Raudenbush (1984).*
**Example: Pygmalion Effect**

Robert Abelson, *Statistics as Principled Argument*, pp 150-152

- Raudenbush (1984) identified possible important differences between studies
- In some studies (circled in red), teachers had little or no contact with students prior to start of experiment
- In others (not circled) teachers had interacted with students prior to experiment
- Notice difference between effect sizes in 2 groups?
- Meta-analysis identified a variable that probably should be evaluated in future studies

**Example: Child/Adolescent Depression**

- Horowitz & Garber (2006)
  - Intensive search for studies of Child/Adolescent depression
  - Found 30 studies that included control & treatment groups
  - Categorized studies using several criteria [useful for comparing results across studies]
  - Used Cohen’s d as index of difference between control & treatment groups

---

**Stem-and-Leaf Plot of Cohen’s d**

**Table 7.3**

A Meta-Analysis of Pygmalion Effects on IQ Scores—Stem-and-Leaf Diagram of Effect Sizes in 18 Studies

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>Stem-and-Leaf Diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td>5</td>
</tr>
<tr>
<td>.40</td>
<td>4</td>
</tr>
<tr>
<td>.30</td>
<td>3</td>
</tr>
<tr>
<td>.20</td>
<td>2</td>
</tr>
<tr>
<td>.10</td>
<td>1</td>
</tr>
<tr>
<td>.00</td>
<td>0</td>
</tr>
<tr>
<td>-0.10</td>
<td>-1</td>
</tr>
<tr>
<td>-0.20</td>
<td>-2</td>
</tr>
<tr>
<td>-0.30</td>
<td>-3</td>
</tr>
</tbody>
</table>

**Note:** Data adapted from Raudenbush (1984).

---

**Example: Child/Adolescent Depression**

- Target Population
  - U = universal [everyone]
  - S = selective [at risk]
  - I = indicated [sub-clinical]
- N = sample size
- \( d_1, d_2, d_3 \): Cohen’s d at end of intervention, at 6 months follow-up, and at last follow-up

---

**Example: Child/Adolescent Depression**

- Many studies have 95% CI that overlap with d=0
  - Yet, 17 out of 21 show \( d > 0 \)
  - More than we would expect by chance if \( d = 0 \) in population
  - FE Model (weighted average of Cohen’s d across studies) > 0
  - Possible file drawer effect?
Publishing in scientific journals is biased towards studies that find significant effects - experiments that yield null results — i.e., no significant difference among groups/treatments — are difficult to publish.

Thus, averaging published effect sizes probably yields a biased estimate of the true effect size (because zero effect sizes aren’t published).

One solution: calculate the number of null results that would have to be published in order for the significant average effect size to disappear.

If only a few null results would wipe out the significant average effect size, then our meta-analysis provides weak evidence for a true effect.

Meta-analysis summary:

- Meta-analysis consists of statistical procedures for integrating and comparing/contrasting results across studies.
- Meta-analysis relies on measures of effect size.
- Meta-analysis provides a clearer picture of an effect than does any single experiment.
- The file-drawer problem can bias the results of meta-analyses, such that averaged effect sizes might appear bigger than they really are.