Formulas Used by the “Practical Meta-Analysis Effect Size Calculator”

David B. Wilson
George Mason University

April 27, 2017

1 Introduction

This document provides the equations used by the on-line *Practical Meta-Analysis Effect Size Calculator*, available at:

`http://www.campbellcollaboration.org/resources/effect_size_input.php`

and

`http://cebcp.org/practical-meta-analysis-effect-size-calculator`

The calculator is a companion to the book I co-authored with Mark Lipsey, titled “Practical Meta-Analysis,” published by Sage in 2001 (ISBN-10: 9780761921684).

This calculator computes four effect size types:

1. Standardized mean difference (d)
2. Correlation coefficient (r)
3. Odds ratio (OR)
4. Risk ratio (RR)

In addition to computing effect sizes, the calculator computes the associated variance and 95% confidence interval. The variance can be used to generate the weight for inverse-variance weighted meta-analysis ($w = 1/v$) or the standard error ($se = \sqrt{v}$).

2 Notation

Notation commonly used through this document is presented below. Other notation will be defined as needed.

d	Standardized mean difference effect size
n_1	Sample size for group 1
n_2	Sample size for group 2
N	Total sample size
s_1	Standard deviation for group 1
s_2	Standard deviation for group 2
s	Full sample standard deviation
se_1	Standard error for the mean of group 1
se_2	Standard error for the mean of group 2
s_{gain}	Gain score standard deviation
$\overline{X}_1$	Mean for group 1
$\overline{X}_2$	Mean for group 2
$\overline{X}_{ij}$	Mean for subgroup i of group j
$\overline{\Delta}_1$	Mean gain score for group 1
$\overline{\Delta}_2$	Mean gain score for group 2
v_d	Variance of d
se_d	Standard error of d
r	Correlation coefficient
Z_r	Fisher's Z_r transformation of r
t	t-value from a Student's t-test
F	F-value from an ANOVA
f_{ij}	Frequency for the ith row and jth column of a contingency table
χ^2	Chi-squared value
b	Unstandardized regression coefficient
β	Standardized regression coefficient
df	Degrees-of-freedom
SS	Sum-of-squares

3 The Standardized Mean Difference Effect Size

The standardized mean difference effect-size (Cohen's d or Hedges' g)[1] is widely used in meta-analysis and more generally as a descriptive statistic in primary studies. The fundamental relationship represented by this effect-size is a dichotomous independent variable and a continuous (scaled) dependent variable. For example, d would be an appropriate effect size index for a study of the effectiveness of a cognitive-behavioral program with two experimental conditions (treatment versus control) and a scaled measure of depression. Alternatively,

[1] Cohen's d and Hedges' g are conceptually the same but Hedges' g is more precise for effect sizes based on small sample sizes. Cohen's d is upwardly biased in absolute value when based on a small sample size. Hedges' g corrects for this with the equation presented in 3.1.3. The term Cohen's d however is often applied to both versions of this effect size. This online calculator does not apply Hedges' sample size size bias correction. Future version may build in this correction.

it would be appropriate for representing the difference between naturally occurring groups, such as boys and girls, on a scaled dependent variable such as reading comprehension. The standardization allows for the comparison of effect sizes across studies with different operationalizations of the same construct. In these two examples, the measures of depression and reading comprehension. Note that the term “effect” does not necessarily imply causation. Effect sizes are merely an index of the empirical relationship of interest, causal or otherwise.

Although d is defined by 4, there are numerous ways to compute d depending on the statistical information available in a given manuscript. Some of these estimation methods for d are algebraically equivalent to 4 whereas others represent an approximation.

3.1 Some Preliminary Equations

Below are the equations for computing the variance of d, the confidence intervals around d, and for the small sample size bias correction.

3.1.1 Variance of d

The variance of the standardized mean difference effect size (d) is:

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2\,(n_1 + n_2)} \,. \tag{1}$$

This estimate of v_d is used for all methods of computing d unless otherwise noted. In general, it is used for all computations of d based on means or on statistics derived from means, such as a t-test. In cases where d is based on a binary (dichotomous) dependent variable, an alternative estimate of v_d is used that is specific to that method of approximating d.

3.1.2 95% Confidence Interval of d

The 95% confidence interval of d is computed in the standard fashion using the standarized normal distribution. The lower-bound of the interval is computed as

$$d_{lower_{95}} = d - 1.959964\sqrt{v_d} \;,$$

and the upper-bound is computed as

$$d_{upper_{95}} = d + 1.959964\sqrt{v_d} \;.$$

3.1.3 Small Sample Size Bias Correction

The standardized mean difference effect size d is slighly upwardly biased in absolute value when based in small sample sizes. This bias is effectively removed by multiplying d by the correction factor J.

$$J = 1 - \frac{3}{4N - 9} \tag{2}$$

Thus, Hedges’ g is computed as

$$g = d \times J. \tag{3}$$

The use of this adjustment is recommended if you are using these effect sizes for the purpose of meta-analysis. Future versions of the calculator will help automate this process. However, it is fairly trivial to apply this correction to all effect sizes that are part of a data file through a transformation statement (such as `compute` in SPSS or `generate` in Stata).

3.2 Means and Standard Deviations

The definitional equation for the standardized mean difference (d) effect size is based on the means, standard deviations, and sample sizes for the two groups being contrasted. The equation is:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}} , \tag{4}$$

where s_{pooled} is

$$s_{pooled} = \sqrt{\frac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{n_1 + n_2 - 2}} . \tag{5}$$

All other equations for d are either algebraic alternatives or approximations of this quantity.

3.3 Student’s t-test (Two Independent Samples)

The standard formula for an independent t-test is equation 4 with an additional term in the denominator based on group sample sizes. Thus, d can easily be computed from t. For unequal sample sizes, the equation is:

$$d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}} . \tag{6}$$

For equal sample sizes, the equation is:

$$d = \frac{2t}{\sqrt{N}} . \tag{7}$$

Note that this equation cannot be used for a dependent or paired t-test. The two means being differenced must be from two independent samles. Also note that this equation cannot be used for t-values from other statistical procedures, such as the t associated with a regression coefficient from a multivariate regression model.

3.4 Significance level (p-value) from a Student's t-test

In cases where the t-value for an independent t-test is not reported the exact p-value is reported, the t-value associated with that p-value and sample size can be determined. This is done by using an asymptotic approximation to the quantile function for the inverse of the two-tailed Student's t distribution. This is implemented using Algorithm 396 (see Hill, 1970). This method is accurate to six decimal places so long as the t has 5 or more degrees-of-freedom. The algorithm is as follows with k representing the degrees of freedom and p representing the p-value:

$$
\begin{aligned}
a &= \frac{1}{k - .5} \\
b &= \frac{48}{a^2} \\
c &= ((20700\frac{a}{b} - 98)a - 16)a + 96.36 \\
d &= ((94.5/(b + c) - 3)/b + 1)\sqrt{(a\frac{\pi}{2})k} \\
x &= dp \\
y &= x^{\frac{2}{k}}
\end{aligned}
$$

If $y \leqslant (a + .5)$ then

$$
\begin{aligned}
y &= ((1/(((k + 6)/(ky) - 0.089d - 0.822)(k + 2)3)+ \\
&\quad 0.5/(k + 4))y - 1)(k + 1)/(k + 2) + 1/y \\
t &= \sqrt{ky}
\end{aligned}
$$

Else if $y > (a + .5)$ then

$$
\begin{aligned}
x &= z\left(\frac{p}{2}\right) \quad \text{Standard normal deviate of } p \text{ divided by 2} \\
y &= x^2 \\
c &= (((.05dx - 5)x - 7)x - 2)x + b + c \\
y &= (((((0.4y + 6.3)y + 36)y + 94.5)/c - y - 3)/b + 1)x \\
y &= ay^2
\end{aligned}
$$

If $y > .002$ in previous line, then

$$t = \sqrt{(\exp^{y} - 1)\, k}$$

Else

$$t = \sqrt{(.5y^2)\, k}$$

Equation 6 or 7 is then used to compute d.

3.5 One-way ANOVA (F-test) with 2 Independent Groups

A one-way ANOVA contrasting the means of two groups produces an F that is equal to t^2. The d effect size for a one-way ANOVA with a single degree-of-freedom in the numerator (i.e., two means) and unequal sample sizes is computed as:

$$d = \sqrt{\frac{F\,(n_1 + n_2)}{n_1 n_2}}\ .$$

This formula simplifies if the group sample sizes are equal:

$$d = 2\sqrt{\frac{F}{N}}\ .$$

This equation is not appropriate for F-values based on three or more groups or categories (i.e., three or more means). See section 3.20 for a method of handling one-way ANOVAs with 3 or more groups or categories.

3.6 Means and Standard Errors

The standard error of a mean can be converted into a standard deviation of the raw data as follows:

$$s = se\sqrt{n-1}\ .$$

Equation 3.2 can then be used to compute d.

3.7 2 by 2 Frequency Table (Contingency Table)

A study with two conditions, such as treatment and control, and a dichtomous dependent variable often reports the findings in a 2 by 2 frequency table (also called a contingency table). A standardized mean difference effect size can be estimated from such data. Note that if the outcome of interest for a meta-analysis is typically dichotomous then one should use the odds ratio or risk ratio as the effect-size of choice. The conversions presented in this section are intended for situations in which most studies measure the construct of interest on a scale but a few do not. The methods below allow for comparison of findings across such studies and for the synthesis of the scaled and dichotmous findings in a single meta-analysis. It is important to note, however, that these methods are approximations to what would have been observed hand the dependent variable been measured on a scale.

Three different methods are provided below. All three produce similar results in most situations—differences increase as the event rate approaches 0 or 1. The first two are based on a conversion of the logged odds ratio. The logic of this is based on the similarity of the logistic and standardized normal distributions. The former has a mean of zero and a standard deviation of $\pi/\sqrt{3}$, whereas the latter has a mean of zero and a standard deviation of 1. The shapes are similar but not identical, with the logistic distribution being somewhat leptokurtic. The Hasselblad and Hedges conversion simply rescales the logged odds ratio based on the difference between the standard deviation of these two distributions. The Cox method tweaks the conversion to account for the slightly leptokurtic nature of the logistic distribution, improving the performance of the conversion when the event rate approaches 0 or 1. The probit method is based entirely on the normal distribution. This approach finds the z associated with the area under the left portion of the curve that equals the proportion experiencing the event for each group. For example, a success rate (or failure rate) of .50 would correspond with a z of 0 (half of the distribution is to the left of 0). Similarly, a success rate of .025 would correspond to a z of -1.96. This value is called the probit of a proportion. The standardized mean difference effect size is computed as the difference between these two probits. Each of these methods has a specific equation for the variance of the effect size, provided below.

The Hasselblad and Hedges method starts by computing the odds ratio. Labeling the cell frequencies of a 2 by 2 table as a, b, c, and d, reading from left to right, top to bottom, the odds ratio is

$$OR = \frac{ad}{bc} \; . \tag{8}$$

The odds ratio is then converted to d using

$$d = \ln\left(OR\right)\frac{\sqrt{3}}{\pi} = \ln\left(OR\right) \times 0.551 \; .$$

The variance of d using this method is

$$v_d = v_{\ln(OR)}\frac{3}{\pi^2} = v_{\ln(OR)} \times 0.304 \; ,$$

where $v_{\ln(OR)}$ is

$$v_{\ln(OR)} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}. \tag{9}$$

The Cox method also starts with the odds ratio (equation 8) and then converts to d using

$$d = \frac{\ln\left(OR\right)}{1.65} \; .$$

The variance of d using this method is

$$v_d = \frac{v_{\ln(OR)}}{1.65^2} \; ,$$

where $v_{\ln(OR)}$ is computed as above in equation 9.

Note that the above becomes undefined if any of the four cell frequencies equals zero. To avoid this, .5 is added to any cell frequency equal to zero.

The probit method computes d as

$$d = \text{probit}\left(p_1\right) - \text{probit}\left(p_2\right) \; . \tag{10}$$

The variance of d using this method is

$$v_d = \frac{2\pi p_1 (1-p_1) e^{z_1^2}}{n_1} + \frac{2\pi p_2 (1-p_2) e^{z_2^2}}{n_2} \ .$$

The calculator determines the probit using the Odeh and Evans' (1974) algorithm. This method is accurate to within 5 decimal places (see Brophy, 1985). The algorithm first computes y as:

$$y = \sqrt{-2\ln(p)} \ .$$

Using y, the probit value (z_p) is then approximated as

$$\begin{aligned} z_p = y - (& (((0.0000453642210148y + 0.0204231210245)y \\ & + 0.342242088547)y + 1)y + 0.322232431088)/ \\ & ((((0.0038560700634y + 0.10353775285)y + \\ & 0.531103462366)y + 0.588581570495)y + \\ & 0.099348462606) \ . \end{aligned} \tag{11}$$

3.8 Binary Proportions

A dichotomous or binary dependent variable may also be reported as a simple proportion or percent succeeding (or failing) in each of the two groups or conditions. The frequencies for a 2 by 2 contingency table can be computed from these proportions using the group sample sizes.

$$\begin{aligned} f_{11} &= p_1 n_1 \\ f_{12} &= (1-p_1)\, n_1 \\ f_{11} &= p_2 n_2 \\ f_{22} &= (1-p_2)\, n_2 \end{aligned}$$

The d type effect-size is then computed as in section 3.7 for a 2 by 2 frequency table.

3.9 Point-Biserial Correlation

The point-biserial correlation is a correlation coefficient between a dichotomous variable, such as a condition dummy code, and a continuous or scaled variable. This correlation is conceptually and algebraically very similar to d and can be converted to d with the formulas below. Separate equations are provided for equal and unequal sample sizes between the two conditions.

For unequal sample sizes, d is computed as

$$d = \frac{r}{\sqrt{(1-r^2)\,(p\,(1-p))}} \ . \tag{12}$$

For equal sample sizes, d is computed as

$$d = \frac{2r}{\sqrt{1 - r^2}} \; . \tag{13}$$

A study may report the p-value for a point-biserial correlation coefficient. The significance test on which this p-value is based is the independent t-test. As such, the method presented in section 3.4 is used to find the t-value and then equation 6 or 7 is used to compute the effect-size. Note that the p-value from a correlation based on two continuous variables (a Pearson's correlation coefficient) or two dichotomous variables (a phi coefficient) should not be handled in this way.

The variance for d based on a point-biserial correlation is computed as in equation 1.

3.10 Phi Coefficient

The relationship between a dichotomous independent variable (such as treatment condition) and a dichotomous dependent variable (such as success or failure) might be presented in a study as a phi coefficient. A study may report this simply as r. However, if the underlying data is dichotomous for both the independent and dependent variable, then the r is a phi coefficient and should be treated as such.

The effect-size d is approximated as

$$d = \frac{2r}{\sqrt{1 - r^2}} \; . \tag{14}$$

Equation 1 for the variance of d is biased for d approximated from a phi coefficient. Given the relationship between phi and χ^2, a variance estimate can be computed that is consistent with the significance test for phi. This is done by first converting phi into a chi^2 using

$$\chi^2 = r^2 * n \; . \tag{15}$$

The variance is then determining as

$$v = \frac{d^2}{chi^2} \; . \tag{16}$$

Note that this estimator for d_v fails when phi equals zero. Also, this approximation of d and d_v should only be used when 2 by 2 frequency or proportion data are not reported or cannot be determined. Computing d from 2 by 2 frequency or proportion data as described in sections 3.7 and 3.8 provides a better estimate.

3.11 Chi-Square (χ^2) based on 2 by 2 Frequency Data

This method is algebraically the same as that used for the phi coefficient (section 3.10). As shown in equation 15, there is a simple algebraic relationship between phi and chi-square. Thus, d is computed from χ^2 as

$$d = 2\sqrt{\frac{\chi^2}{N - \chi^2}}. \tag{17}$$

As with the phi coefficient, the variance is estimated such that it is consistent with the chi-squared test.

$$v = \frac{d^2}{\chi^2}. \tag{18}$$

This estimator fails when chi-squared equals zero. The chi-square must have one degree-of-freedom. If it has two or more degrees-of-freedom, then it is not based on 2 by 2 frequency data.

Also note that this estimator for d should only be used when the 2 by 2 frequency or proportion data are not reported or cannot be determined. Computing d from 2 by 2 frequency or proportion data as described in sections 3.7 and 3.8 provides a better estimate.

3.12 Exact p-value for a Chi-Squared based on 2 by 2 Frequency Data

The chi-squared is the most commonly used significance test for a 2 by 2 frequency table. The chi-squared distribution with one degree-of-freedom is the squared standardized normal distribution. Thus, the chi-squared associated with a p-value from a chi-squared computed on a 2 by 2 frequency table can be determined using equation 11 and squaring the result, that is,

$$\chi^2 = z_p^2 \ .$$

The effect size d is then computed using the methods of section 3.11. Note that the chi-squared must have only 1 degree-of-freedom. Also, this method should only be used when other data are not provided.

3.13 Frequency Distribution

A study may report a frequency distribution by group for an ordinal, interval, or ratio scale but not report means, standard deviations, or a t-test that would allow for the easy computation of d. Assuming equal intervals (and order) for the categories of the frequency distribution, we can compute a mean and standard deviation for each group by assigning the values 1, 2, 3, etc. to the rows of the frequency distribution. For an ordinal distribution, the assumption of equal intervals is not satisfied and as such this method should be used cautiously with recognition that it may result in a biased estimate.

The means for each group are computed as

$$\overline{X}_1 = \frac{\sum f_{i1} i}{\sum f_{i1}},$$

and

$$\overline{X}_2 = \frac{\sum f_{i2} i}{\sum f_{i2}},$$

where f_{ij} are the frequencies for each row, i, and each group, j.

The standard deviations are computed as

$$s_1 = \sqrt{\frac{\sum f_{i1} \sum f_{i1} i^2 - \left(\sum f_{i1} i\right)^2}{\left(\sum f_{i1}\right)^2}},$$

and

$$s_2 = \sqrt{\frac{\sum f_{i2} \sum f_{i2} i^2 - \left(\sum f_{i2} i\right)^2}{\left(\sum f_{i2}\right)^2}}.$$

If only proportions are provided, then the frequencies are computed as

$$f_{ij} = p_{ij}(n_j).$$

The d effect size is then computed using 4.

3.14 Unstandardized Regression Coefficient

The unstandardized regression coefficient for treatment dummy variable (a variable coded as 0 and 1 or 1 and 2 to indicate group membership) represents the mean difference between the two groups adjusted for the other independent variables included in the model. As such, the unstandardized regression coefficient (b) can be used in the numerator of equation 4 in computing d (i.e., it is the covraiate adjusted mean difference). What is needed is an estimate of the denominator, s_{pooled}. The difference between the standard deviation of the the dependent variable (s_y) and the pooled within groups standard deviation (s_{pooled}) is that the latter excludes variability attributable to group membership (treatment). Thus, s_{pooled} can be determined via

$$s_{pooled} = \sqrt{\frac{s_y^2(N-1) - b^2\left(\frac{n_1 n_2}{n_1+n_2}\right)}{N-2}}\ . \tag{19}$$

This equation removes the variance associate with treatment from the total variance. Using this estimate of s_{pooled}, the effect size is computed as

$$d = \frac{b}{s_{pooled}}\ . \tag{20}$$

3.15 Standardized Regression Coefficient

The standardized regression coefficient, (β), can also be used to compute d but it first must be converted into an unstandardized coefficient (b). The relationship between the standardized and unstandardized coefficient is a function of the standard deviations for both the dependent (y) and independent (x) variables. The standard deviation of the dependent variable is often reported but the standard deviation for a treatment dummy variable rarely is. However, because the sample size for each group is generally available, the standard deviation of the dummy variable for x can be computed as

$$s_x = \sqrt{\frac{n_1 - n_1^2/N}{N-1}}. \tag{21}$$

The unstandardized regression coefficient is then computed as

$$b = \beta \frac{s_y}{s_x}, \tag{22}$$

and d is estimated as using the equations of section (3.14).

3.16 Means and Full Sample Standard Deviation

The pooled standard deviation used in the denominator of d is simply the total sample standard deviation for the dependent variability with the variance due to group membership removed. Thus, the pooled standard deviation can be computed from the total or full sample standard deviation as follows

$$s_{pooled} = \sqrt{\frac{s_y^2(N-1) - \frac{\left(\overline{X}_1^2 + \overline{X}_2^2 - 2\overline{X_1 X_2}\right)}{N}}{N}}. \tag{23}$$

The d is then computed with equation 4

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}}. \tag{24}$$

3.17 Mean Gain Scores and Gain Score Standard Deviations

Gain scores represent a challenge for computing d. Simply treating the mean and standard deviation of gain scores as simple means and standard deviations and proceeding with equation 4 will produce misleading results. The standardized mean difference version of d is standardized based on the variability across individuals on the dependent variable (i.e., variability in the raw scores). Using the standard deviation of gains scores in place of the standard deviation of the raw dependent variable scores standardizes on variability in change or gains, a rather different thing altogether. Essentially, using the standard deviation of gains produces a version of d that is on a different scale (i.e., standardizing on gains) and is not directly comparable to the standardized mean difference effect size that is standardizing on differences across individuals.

There are two reasons why we might be trying to convert gain score statistics into a standardized mean difference type d. First, this might be the only statistical data reported. Second, we might want to use a mean difference that is adjusted for baseline scores (a difference in differences). A workable solution is to convert the standard deviation of the gain scores into a standard deviation of raw scores. Unfortunately, the correlation between the time 1 and time 2 scores is needed to make this conversion and this is often not available (note that the alpha coefficient for internal consistency that is often reported as a measure of reliability is not an estimate of the time 1 versus time 2 correlation). In some situations you may have reasonable estimates of this correlation from other sources, such as from psychometric studies, and using these approximations may be preferable to omitting the study from the meta-analysis altogether. However, if you do so, sensitivity analyses are recommended to assess the affect of these approximations on overall results.

The conversion of a standard deviation of gain scores (s_{gain}) to a standard deviation of raw scores (s) is

$$s = \frac{s_{gain}}{\sqrt{2\left(1-r\right)}} \tag{25}$$

The pooled standard deviation is then computed using equation 5 and d is computed as

$$d = \frac{\overline{\Delta}_1 - \overline{\Delta}_2}{s_{pooled}}. \tag{26}$$

The variance d based on gain scores is computed as

$$v = \frac{2\left(1-r_1\right)}{n_1} + \frac{2\left(1-r_2\right)}{n_2} + \frac{d^2}{2\left(n_1+n_2\right)}$$

3.18 Mean Gain Scores, Pre and Post-test Standard Deviations, and a Paired t-test or Pre-Post r

A study may present pretest and post-test means, standard deviations, and sample sizes. Using either the pretest or post-test data, d can be computed using 4. However, a mean difference that is adjusted for pretest differences, a difference in differences, effect size may be wanted and can be computed as

$$d = \frac{\overline{\Delta}_1 - \overline{\Delta}_2}{s_{pooled}} \tag{27}$$

where $\overline{\Delta}_1$ and $\overline{\Delta}_2$ are the pre-post mean differences for each group, computed as

$$\overline{\Delta}_j = \overline{X}_{post} - \overline{X}_{pre}\ . \tag{28}$$

The pooled standard deviation is based on all four raw score standard deviations. First, the pooled pre-post standard deviation for each group is determined as

$$s_j = \sqrt{\frac{s_{pre}^2 + s_{post}^2}{2}}\,. \tag{29}$$

Using s_j for each group, the s_{pooled} is computed using equation 5.

Although an estimate of the variance of d can be determined using 1, the variance of d is affected by the precision (or imprecision) of all four means of which it is a function. The proper estimate of v_d must therefore take this into account and can be estimated as

$$v = \frac{2(1-r_1)}{n_1} + \frac{2(1-r_2)}{n_2} + \frac{d^2}{2(n_1+n_2)} \tag{30}$$

where r_1 and r_2 are the correlations between the pretest and post-test scores within each group. These values are rarely reported but can be determined from paired t-tests. The latter are often reported in such situations. The pretest/posttest correlation can be determined for each group as

$$r = \frac{\left(s_1^2 t^2 + s_2^2 t^2\right) - \left(\overline{X}_2 - \overline{X}_1\right)^2 n}{2 s_1 s_2 t^2} \tag{31}$$

where s_1 and s_2 were the pre-test and posttest standard deviations, and t is the paired t-test.

3.19 Means and Standard Deviations with Subgroups

A common situation in conducting a meta-analysis is finding a study that reports means and standard deviations for subgroups within each condition but does not report the overall results. For example, a study may report the results of a treatment and control group comparison separately for boys and girls. The mean for each condition can easily be found by computing a weighted mean. The pooled standard deviation is less straightforward and there are two options, depending on the situation. If the subgroups represent a breakout on a sample characteristic, such as gender or risk-level, then simply pooling the standard deviations, as with formula 5, will underestimate the full within-group standard deviation because variability associated with the subgroup variable has been removed (e.g., variability in the outcome associated with gender). Thus, it is important to compute the variability associated with this subgroup and add it back into the pooled standard deviation. There are situations, however, where the subgroups represent a manipulated variable that adds variability rather than subtracts it. For example, in a two-by-two factorial design where each factor is a true independent variable (i.e., manipulated by the researcher). In such a case, this additional variability associated with a manipulated factor should not be added to the pooled within-groups standard deviation.

The weighted mean for each condition (i.e., treatment and control) is computed as:

$$\overline{X} = \frac{\sum \overline{X}_j n_j}{\sum n_j},$$

where j represents each subgroup.

The pooled within-group standard deviation for each condition (i.e., treatment and control), ignoring any variance removed due to the subgroup varable, is

$$s_{pooled} = \sqrt{\frac{\sum s_j^2 \left(n_j - 1\right)}{\sum \left(n_j - 1\right)}}.$$

The subgroup variable is added back into the within group variance using the following formula:

$$s_{pooled} = \sqrt{\frac{\sum s_j^2 \left(n_j - 1\right)}{\sum \left(n_j - 1\right)} + \sum \overline{X}_j^2 n_j - \frac{\left(\sum \overline{X}_j n_j\right)^2}{\sum n_j}}.$$

The within-groups pooled standard deviation is then computed using 5.

3.20 F-test (ANOVA) with 3 or More Groups

The F-value from a one-way ANOVA with three or more groups can be used to compute d so long as the means and sample sizes on which the F is based are also available. Thus, this method provides a way to determine the pooled standard deviation when the standard deviation for each group is missing.

In the case of a one-way ANOVA, F is the ratio of the MS-between group means to the MS-within (or MS-error) and the MS-within is the pooled within group variance. We can determine the MS-within as

$$MS_{within} = \sqrt{\frac{MS_{between}}{F}}\ , \tag{32}$$

where $MS_{between}$ is found through

$$MS_{between} = \frac{\sum \overline{X}_j^2 n_j - \frac{\left(\sum \overline{X}_j n_j\right)^2}{\sum n_j}}{k - 1}\ , \tag{33}$$

where k is the number of groups. d is then computed using 4. If two or more means are selected to represent either the treatment or control group, then the weighted mean across the multiple groups is used. For example, if two groups are selected as being part of the treatment condition, the weighted mean for the treatment group is computed as

$$\overline{X}_1 = \frac{\sum \overline{X}_j n_j}{\sum n_j}. \tag{34}$$

3.21 Means and an Analysis-of-Covariance (ANCOVA)

An analysis of covariance tests the effect of treatment or group membership on the dependent variable adjusted for a continuous covariate, such as a pretest or other characteristic of the study participants. The F-value for the treatment effect cannot be converted as described in section 3.5 and using that method will over-estimate d. However, if sufficient information is provided, d can be computed as follows. First, s_{pooled} is computed as

$$s_{pooled} = \sqrt{\left[\frac{MS_{error}}{1 - r^2}\right]\left[\frac{df_{error} - 1}{df_{error} - 2}\right]}, \tag{35}$$

where r is the correlation between the covariate and the dependent variable. In the case of multiple covariates, the multivariate $\sqrt{R^2}$ can be used.

The means and group sample sizes are needed to compute d using equation 4. Thus this method is only useful when the standard deviations missing but the full ANCOVA results are reported along with the correlation between the covariate and dependent variable. In practice, this is rarely the case.

3.22 Means and a Two-way ANOVA

As with the ANCOVA, an F-value from a two-way ANOVA cannot be treated as an F-value from a one-way ANOVA. Doing so will over-estimate d. Computing d requires information from the full sums-of-squares table and uses this information to determine the pooled within groups standard deviation. A two-way ANOVA produces three F-values, one for each main effect (often referred to generically as factors A and B), and one for the interaction between the two facotrs. Assuming that the effect of interest is factor A (e.g., the factor representing the treatment versus control group), you must decide whether to incorporate the variance associated with factor B into the pooled standard deviation. If factor B is a blocking factor (i.e., not manipulated) such as gender, risk-level, etc., then it makes sense to include this variability in the pooled standard deviation. If factor B is a manipulated factor such as another treatment manipulation, then it may make sense to exclude this variability in the pooled standard deviation. The default for the online calculator is to include the variability of factor B, the non-focal factor.

The pooled standard deviation with the variability of factor B is computed as

$$s_{pooled} = \sqrt{\frac{SS_{B} + SS_{AB} + SS_{error}}{df_{AB} + df_{AB} + df_{error}}}. \tag{36}$$

The pooled standard deviation without the variability of factor B is computed as

$$s_{pooled} = \sqrt{\frac{SS_{AB} + SS_{error}}{df_{AB} + df_{error}}}. \tag{37}$$

Using the appropriate estimate of s_{pooled}, d is computed using equation 4.

4 The Odds Ratio and Risk Ratio Effect Size

The odds ratio (OR) and risk ratio (RR) are effect sizes suitable for a two group comparison (e.g., experimental and control) and a dichotomous outcome (e.g., success and failure). These effect sizes are widely used in meta-analyses within public health and medicine and are become increasingly common in the social sciences. These effect sizes are computed from raw frequencies, proportions or percents. It is also possible to transform an r or χ^2 based on a 2 by 2 frequency table into an OR or RR so long as the marginal distributions for the independent (grouping) and dependent (outcome) variables are available.

4.1 Frequency Data

Both the odds ratio (OR) and the risk-ratio (RR) are easily computed from 2 by 2 frequency data as follows

$$OR = \frac{ad}{bc}, \tag{38}$$

$$RR = \frac{a/(a+b)}{c/(c+d)}, \tag{39}$$

where a, b, c, and d, are the cell frequencies of a 2 by 2 table reading from left to right and top to bottom.

Given the asymmetric nature of these effect sizes, meta-analysis is performed on the natural log of each. The variance for the logged OR and logged RR are

$$v_{\ln(OR)} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{40}$$

$$v_{\ln(RR)} = \frac{b/a}{a+b} + \frac{d/c}{c+d} \tag{41}$$

4.2 Proportion or Percent Data

Although OR and RR can be computed directly from proportions, the effect size calculator recreates the 2 by 2 frequencies and uses the equations above. This has the advantage that these frequencies are needed to compute the variance estimates. These equations assume that these are the within group proportions or percentages (i.e., proportion successful in each group). The cell frequencies are computed from proportions as

$$a = \frac{p_1}{n_1}, \tag{42}$$

$$b = \frac{1-p_1}{n_1}, \tag{43}$$

$$c = \frac{p_2}{n_2}, \tag{44}$$

$$d = \frac{1 - p_2}{n_2}. \tag{45}$$

Percents are converted to proportions through division by 100.

4.3 Phi and Chi-squared

The odds ratio and risk ratio can be computed from a phi coefficient (i.e., a correlation based on 2 by 2 frequency data) or a chi-squared with one degree-of-freedom (i.e., chi-square based on 2 by 2 frequency data). However, the marginals of the 2 by 2 table are needed. Generally the sample sizes for the two conditions are known, providing the marginals for the independent variable. The overall event-rate (success or failure rate) is needed to determine the dependent variable marginals.

Using the chi-squared, the top-left cell frequency, a, can be determined as

$$a = N\left[p_x p_y \sqrt{\frac{\chi^2 p_x p_y (1 - p_x)(1 - p_y)}{N}}\right]$$

where p_x is the proportion of the sample in one of the two conditions, p_y is the proportion of the overall sample with a positive (or negative) outcome, and N is the total sample size.

Using the phi coefficient (r), the top-left cell frequency, a, can be determined as

$$a = N\left[p_x p_y + r\sqrt{p_x p_y (1 - p_x)(1 - p_y)}\right] .$$

The remaining cell frequencies can be determined as follows:

$$b = n_1 - a,$$

$$c = N(p_y) - a,$$

$$d = N - (a + b + c).$$

The odds ratio and risk ratio and associated variances are computed using equations 38, 39, 40, and 41 using the frequencies of the 2 by 2 table determined above.

4.4 Odds Ratio Based on d

An odds ratio can be approximated from a standardized mean difference effect size, d. This can be useful when the dominant method of measuring the outcome of interest produces dichotomous data but a subset of relevant studies measure the outcome on a scale. This conversion allows for the combining of effect size based on scale measures along with those based on dichotomous data. Note that this is an approximation and assumes that the outcome measure is normally distributed, although simulations show that it performs reasonably well with moderately skewed data. Two methods are presented below, the Hasselblad and Hedges logit method (denoted simply as $\mathrm{OR}_{\mathtt{logit}}$) and the modified logit method proposed by Cox (denoted as $\mathrm{OR}_{\mathtt{Cox}}$). The Cox method tweaks the conversion to account for the more peaked nature of the logistic distribution. This method performs slightly better when the event rate approaches 0 or 1. The conversions and associated variance estimates are computed as

$$\mathrm{OR}_{\mathtt{logit}} = \frac{d\pi}{\sqrt{3}}, \tag{46}$$

$$v_{\mathtt{logit}} = \frac{v_d \pi^2}{3}, \tag{47}$$

$$\mathrm{OR}_{\mathtt{Cox}} = d1.65, \tag{48}$$

$$v_{\mathtt{Cox}} = \frac{v_d}{.367}. \tag{49}$$

The risk ratio depends on the marginal distribution of the outcomes and as such d cannot be converted into a risk ratio.

5 Correlation Coefficient Effect Size

The correlation coefficient is a commonly used effect size for meta-analysis. It is best suited to research domains in which the two variables of interest are continuous in nature, even if some studies dichotomies one or both of the variables. The equation for computing the Pearson’s correlation coefficient can be ready found in most any introductory statistics book and as such is not reproduced here.

A complication with using the correlation as the effect size in a meta-analysis that relies on the inverse-variance weight method is the inability to compute a variance for r. A solution is Fisher’s z transformation of r (denote as z_r below). This is a variance stabilizing transformation and as r approaches -1 or 1, z_r approaches infinity. The z_r transformation is

$$z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) . \tag{50}$$

The variance of z_r is simply

$$v_{z_r} = \frac{1}{N-3} . \tag{51}$$

Results, such as a mean z_r and associated confidence intervals, can be converted back in r using

$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1} . \tag{52}$$

5.1 Correlation from a j by k Frequency Table

Before computers were widely used to compute statistics there were computationally efficient methods of computing common statistics, such as Pearson's correlation coefficient, from a contingency table, rather than the raw data. This method can be used to compute a correlation from a contingency table (a j by k frequency table, where j is the number of rows and k is the number of columns). This method assumes that each variable is at least ordinal and that the table maintains the ordered nature of the data.

$$r = \frac{N \sum f_{jk} - f_{jk} j f_{jk} k}{\sqrt{\left[N \sum f_{jk} j^2 - \left(\sum f_{jk} j\right)^2\right] \left[N \sum f_{jk} k^2 - \left(\sum f_{jk} k\right)^2\right]}}$$

where j represents the rows and k represents the columns.

5.2 Correlation from d

The standardized mean difference effect size, d, can easily be converted into a correlation coefficient using the following formula:

$$r = \frac{d}{\sqrt{d^2 + 1/(p(1-p))}} \, , \tag{53}$$

where p is defined as the proportion of the total sample in either one of the two groups. This is computed as

$$p = \frac{n_1}{n_1 + n_2} \, . \tag{54}$$

The conversion produces the point-biserial correlation coefficient. If the sample sizes in the two groups are equation, the equation simplifies to

$$r = \frac{d}{\sqrt{4(d^2 + 1)}} \, . \tag{55}$$

The variance for z_r based on d is

$$v_{z_r} = \frac{v_d}{v_d + 1/(p(1-p))} \, , \tag{56}$$

where v_d is the variance of the d being converted.

5.3 Correlation from a 2 by 2 Frequency Table

A correlation coefficient based on a 2 by 2 frequency (contingency) table is a phi coefficient, although using the Pearson's correlation coefficient equation will produce the same result. In the equation below, a, b, c, and d, are the cell frequencies of the 2 by 2 table, reading from left to right and top to bottom. Phi (r) is computed as

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+c)}} \, . \tag{57}$$

Because the variance of z_r based on the above is affected by both the overall sample size and the marginal distributions, a more precise variance estimate is computed by rescaling the variance for a logged odds ratio. Thus, the variance for r based on phi is computed as

$$v_r = v_{\ln(OR)} \frac{z_r^2}{\ln(OR)^2} \tag{58}$$

where $v_{\ln(OR)}$ is computed using equation 40 and $\ln(OR)$ is computed using equations 38. This provides a more accurate estimate of $v_{z_r r}$ then 51.

5.4 χ^2 and Sample Size

A chi-squared for a relationship between two dichotomous variables can be converted into r. The r in this case is a ϕ coefficient as is computed as

$$r = \sqrt{\frac{\chi^2}{N}}. \tag{59}$$

This value is then transformed into z_r using equation 50. The variance of the Fisher’s z_r is estimated as

$$v_{z_r} = \frac{z_r^2}{\chi^2} \tag{60}$$

and produces confidence intervals consistent with the χ^2 test.

Note that this conversion is appropriate only for a χ^2 from a 2 by 2 frequency table. A χ^2 of such a table will have 1 degree-of-freedom. A χ^2 with 2 or more degrees-of-freedom cannot be converted with this equation and does not reflect the linear relationship between the two variables (i.e., you cannot convert it to an r).

5.5 t-test and Exact p-value From a t-test

The t-test and exact p-value from a t-test can easily be converted to r so long as you have the sample size on which the test was based. r based on t is computed as

$$r = \frac{t}{\sqrt{t^2 + N - 2}}. \tag{61}$$

If only the p-value is reported, r can be determined using the equation above after computing the t-value associated with the reported p-value using the method shown in section 3.4.

References

Borenstein, M. (2009). Effect sizes for studies with continuous outcome data. In H. M. Cooper, L. V. Hedges, & J. Valentine (Eds.), *Handbook of research synthesis* (2nd ed.). New York: Russell Sage Foundation.

Brophy, A. L. (1985). Aproximation of the inverse normal distribution function. *Behavior Research Methods, Instruments, & Computers*, *17*, 415–417.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (p. 245-260). New York: Russell Sage Foundation.

Fleiss, J. L., & Berlin, J. A. (2009). Measures of effect size for categorical data. In H. M. Cooper, L. V. Hedges, & J. Valentine (Eds.), *Handbook of research synthesis* (2nd ed.). New York: Russell Sage Foundation.

Hasselblad, V., & Hedges. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*, 167-178.

Hill, G. W. (1970). Algorithm 396: Student’s t-quantiles [S14]. *Communications of the ACM*, *13*, 619–620.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, Calif: Sage Publications.

Odeh, R. E., & Evans, J. O. (1974). Algorithm AS 70: The percentage points of the normal distribution. *Applied Statistics*, *23*, 96–97.

Rosenthal, R. (1994). Parametric measures of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (p. 231-244). New York: Russell Sage Foundation.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003, December). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*, 448-467.