**Mathematical details of effect sizes used in this paper**

**x.1 Standardized effect size typology**

Over the last 30 years an increased focus has been placed on the reporting and interpretation of effect sizes as an important part of the development of a cumulative and interpretable research literature {Kruschke, 2017 #105;e.g.`, \Cumming, 2013 #158;Wilkinson, 1999 #566;Hedges, 1981 #786}. Effect sizes can be expressed in standardised or unstandardized units. Unstandardized effect sizes (e.g., mean differences) are presented in the units the measured variables, and may be particularly useful when the units of analysis are directly interpretable (e.g., income, IQ scores, measures of height or weight). Standardised effect sizes (e.g., Cohen’s *d* for mean differences) have several distinct uses. Some measures may be useful for facilitating interpretation when the units of measurement are not themselves interpretable (e.g., a newly developed measure), as they express observed patterns in the data in a way interpretable without reference to the units of measurement. Effect sizes are also useful in meta-analysis, allowing for studies’ effects to be compared and collapsed. Standardised effect sizes are useful in power analysis as they are a succinct way of proving a great deal of information about the alternative hypothesis.

In order to perform formal sample size determination, researchers must specify an alternative hypothesis in sufficient detail to determine the sampling distribution of the test statistic under the alternative hypothesis. For relatively simple designs (e.g., for a comparison of the mean scores of two independent groups or correlational analysis) the specification of a single standardised effect size characterises the sampling distribution under the alternative hypothesis adequately for power analysis {Cohen, 1988 #562}. For more complex designs (e.g., when covariates are to be included or when repeated measures designs are used) additional parameters may need to be specified. One of the major difficulties often cited by researchers in performing a power analysis is that they have trouble developing appropriate parameters for use in power analysis [cite interviews and survey]. This chapter outlines the three main types of effect sizes, outlines some commonly maligned benchmarks that have been proposed for power analysis.

This chapter provides the definitions of the different standardised effect sizes that are often used in power analysis, using the notation and terminology that is followed in this dissertation. This chapter groups standardized effect sizes into three main categories; effect sizes for group differences (e.g., Cohen’s *d* and Hedge’s *g*), variance explained effect sizes (e.g., r, R2, eta2, partial eta2, omega2), and probability effect sizes (e.g., odds ratios, hazard ratios).

**Effect sizes for Mean differences**

Cohen’s *d*, was originally proposed as an measure of the size of effect in Cohen’s first power survey, and was explicitly developed to aide in sample size determination (Cohen, 1962). There are a number of estimators for the population parameter $\delta$, the difference between groups divided by the pooled standard deviation. The estimates produced by all of these estimators are commonly called “Cohen’s *d*”, and all use equation x.1.

C:\Users\fsingletonthorn\Downloads\CodeCogsEqn (1).gif x.1

$d = \frac{\bar{x}\_1 - \bar{x}\_2}{s\_p}$

(adapted from McGrath & Meyer, 2006, p. 386)

Where $\bar{x}\_1$ is the mean of sample 1, and $\bar{x}\_2$ is the mean of sample 2, and $s$ is the pooled standard deviation. The pooled standard deviation is most often calculated for samples as:

https://latex.codecogs.com/gif.latex?%24%24s%20%3D%20%5Csqrt%7B%5Cfrac%7B%5Csum%28x_1%20-%20%5Cbar%7Bx%7D_1%29%5E2%20&plus;%20%28x_2%20-%20%5Cbar%7Bx%7D_2%29%5E2%7D%7Bn_1%20&plus;%20n_2%20-%202%7D%7D%24%24 x.2

$$s\_p = \sqrt{\frac{\sum(x\_1 - \bar{x}\_1)^2 + (x\_2 - \bar{x}\_2)^2}{n\_1 + n\_2 - 2}}$$

(Cohen, 1977, p. 67)

Or equivalently as equation x.3[[1]](#footnote-1).

(x.3)

$$s = \sqrt{frac{(n\_1-1)s\_1^2 + (n\_2-1)s\_2^2}{n\_1 + n\_2 - 2}}$$

(adapted from Hedges, 1981, p. 110)

Where $s\_j^2$ is the sample variance for each group, calculated as

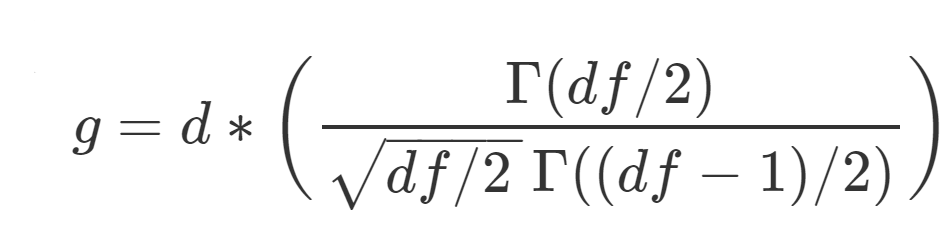
https://latex.codecogs.com/gif.latex?s%5E2_j%20%3D%20%5Cfrac%7B1%7D%7Bn_j-1%7D%20%5Cdisplaystyle%5Csum_%7Bi%3D1%7D%5E%7Bn%7D%20%28x_%7Bj%2Ci%7D%20-%20%5Cbar%7Bx%7D_j%29%5E2 (x.4)

$$s^2\_j\ =\ \frac{1}{n\_j-1}\ \displaystyle\sum\_{i=1}^{n}\ (x\_{j,i}\ -\ \bar{x}\_j)^2$$

The j subscript indicating the group.

The pooled standard deviation should be calculated for populations (i.e., if all possible units of analysis have been collected) using n1+n2 in the denominator as opposed to n1+n2-2, without Bessel’s correction (Cohen, 1977, 1988; McGrath & Meyer, 2006).

Terminology around these effect sizes is remarkably inconsistent, and sometimes Cohen’s d is reserved to describe the estimator that doesn’t use Bessel’s correction, and the estimator outlined in equation x.1 to x.3 is called Hedges’ g (e.g., (Rosenthal, 1991)). However, as Cohen outlined both estimators (e.g., Cohen, 1977) before Hedges (1981), and as the population version is rarely applicable, it seems reasonable to use Cohen’s *d* to refer to the estimator outlined in equations x.1 to x.4. This estimator for Cohen’s d is consistent (that is, as the n increases its expectation increasingly accurately approximates the population parameter), but it is upwardly biased (it tends to overestimate the population parameter, especially when the included sample size is small). Hedges (1981) outlines a correction factor to produce an unbiased estimator:

(x.5)

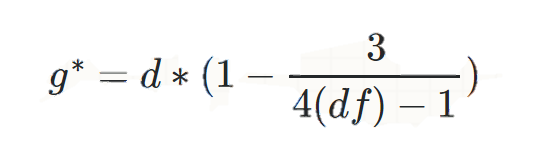
$$g = d(\frac{\Gamma(df/2)}{\sqrt{df/2 \,}\,\Gamma((df-1)/2)})$$

Where $df=n1+n2−2$ for an independent groups design, d is calculated as per equation x.1 and $\Gamma(x)$ is the gamma function.

(Originally from Hedges, 1981; this version adapted from Hedges & Olkin, 1985, p. 104).

However, this correction factor is fairly computationally complex (although trivial on modern computers), so Hedges also provided a computationally simple approximation which performs well for all practical scenarios (Hedges, 1981, p. 114).

Hedge’s approximate bias corrected *g*\* is calculated as:

 (x.6)

$$g^\* = d(1 - \frac{3}{4(df)-1})$$

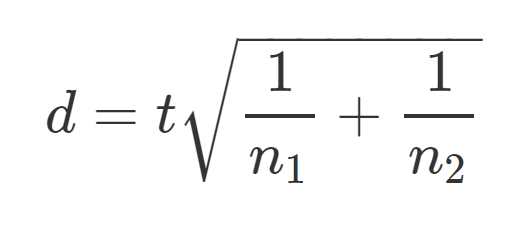
Where $df=n1+n2−2$ for an independent groups design and *d* is Cohen’s d as calculated in equation x.1 using x.2 as the estimator for the variance.

(this version adapted from Borenstein, Hedges, Higgins, & Rothstein, 2011, p. 27; originally from Hedges, 1981)

People commonly refer to $d$, $g$ and $g^\*$ as Hedge's g or Cohen's d interchangeably (Lakens, 2013). They are all virtually identical for most practical purposes when *n* > 30, and all can be interpreted in the same way. For the purposes of power analysis, it is important to realise that Cohen’s d is upwardly biased if estimating a $\Delta$ based on a literature that uses the biased estimator. Practically, for the purposes of power analysis, sampling variability and selective reporting are likely to create greater difficulties than the estimator that has been used.

**Summary statistics conversion for two group scenarios**

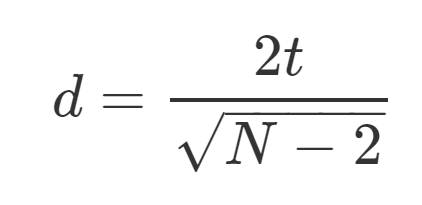
If effect sizes have not been reported, Cohen’s d can be calculated using the results of an independent samples t tests using the formula



$$d = t\sqrt{\frac{1}{n\_1}+\frac{1}{n\_2}} $$

(Lakens, 2013, equation 2)

Where $n\_1$ and $n\_2$ are the sample sizes for groups 1 and two respectively, and *t* is the result of an independent samples t test.



$$d=\frac{2t}{\sqrt{N - 2}}$$

(Rosenthal, 1991, p. 17)

Which is correct if the groups are equal, and will be an underestimate if the groups are unequal. However, although even if the ratio of samples sizes in each group is as extreme as 70 to 30 the underestimation will be no more than 8% (Rosenthal, 1991).

**Standardised mean differences for the comparisons of two repeated measures:**

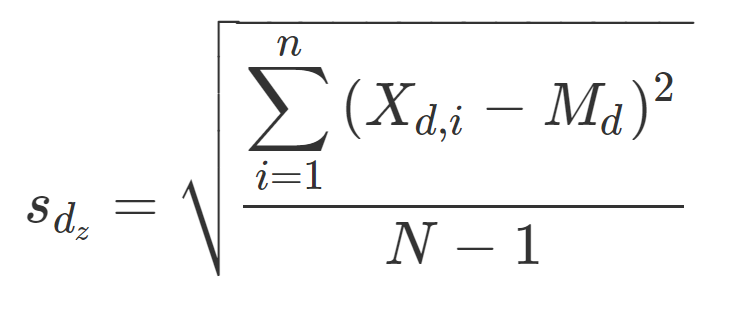
The most common effect size measure for mean difference between repeated measures is also commonly called Cohen’s d, and following Cohen (1977, 1988) I will refer to the repeated measures version as Cohen’s $d\_z$. This effect size follows a similar general form to the independent samples Cohen’s d (x.1), except the denominator is the mean difference between measures,

http://www.sciweavers.org/upload/Tex2Img_1522718895/render.png (x.7)

$$d\_z = \frac{M\_d}{s\_d}$$

(Lakens, 2013) equation 6

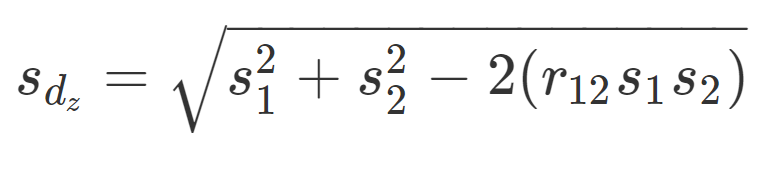
Where $M\_d$ is the mean difference score, and $s\_d$ is the standard deviation of the difference scores calculated as:

 (x.8)

$s\_{d\_z} =\sqrt{\frac{\displaystyle\sum\_{i=1}^{n}{(X\_{d,i} - M\_d)^2}}{N-1}}$

Where $\X\_{d,i}$ is the difference scores for case i, $M\_d$ is the mean difference score, and $ s\_{d\_z}$ is the standard deviation of the difference scores.

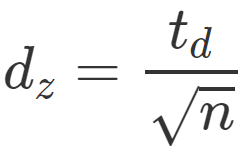
Equivalently, $s\_{d\_z}$ can be calculated as :

 (x.9)

$s\_{d\_z} = \sqrt{s\_1^2 + s\_2^2 - 2(r\_{12}s\_1s\_2)}$

(Cohen 1988 p. 48)

Where $s\_1$ and $s\_2$ are the variances of groups one and two, and $r\_{12}$ is equal to the Pearson correlation between subjects measures on measure one and measure two. Notably, this equation x.9 highlights an important fact Cohen’s $d\_z$, and repeated measures t tests, that the effect size is dependent upon the correlation between scores on repeated measures. The higher the correlation, the greater the $dz$. This can make a large difference to the $d\_z$, for example, taking the example the two groups have equal variance, and there is a one standard deviation difference between groups, $d\_z$ can vary between .707 (when $r\_{12} = 0$) and infinity (when $r\_{12}$ approaches 1). For this reason it has been argued that classical Cohen’s d (equation x.1) should be interpreted in lieu of $d\_z$ for maximum interpretability and comparability across experimental designs {Morris, 2002 #808}. However, for the purposes of power analysis, it is beneficial to use $d\_z$, due to the fact that as the correlation between repeated measures increases the standard error of the difference decreases, or equivalently, the size of the test statistic increases, as can be seen in equation [Rosen]:

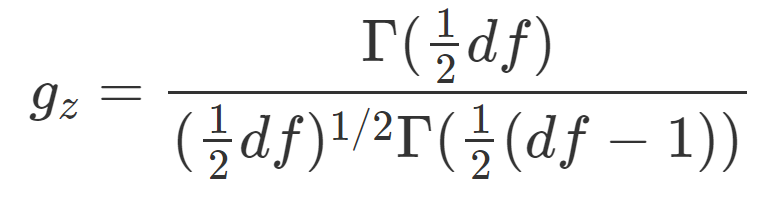
 equation [Rosen]

$$d\_z = {frac{t\_d}{sqrt{n}}} $$

(Lakens, 2013) Equation 7

Where $*t*\_d$ is the *t* statistic calculated as per a repeated measures *t* test, and *n* is the sample size.

$d\_z$ is also biased, and an equivalent to Hedges’ correction can be applied to adjust for this bias similarly to the independents samples Cohen’s d (Gibbons, Hedeker, & Davis, 1993):

 (x.Gibbons)

(equation 7, p. 274 Gibbons, Hedeker & Davis, 1993)   
$$g\_z = \frac{\Gamma(\frac{1}{2}df)}{(\frac{1}{2}df)^{1/2} \Gamma(\frac{1}{2}(df-1))}$$

Where $df =$ degrees of freedom (i.e., $n - 1$ as per repeated measures *t*-tests) and $\Gamma (x)$ is the gamma function.

**Association/variance explained:**

Almost certainly the most commonly used measure of association and one that almost all psychological scientist will be family with is Pearson’s Product Moment Correlation Coefficient, Pearson r. One of the oldest standardised effect sizes commonly used today, r measures the degree of linear association between two variables and was pioneered by Galton and further developed by Karl Pearson {Pearson, 1903 #927}.

Where *x* are the values of x, y are the values of y, and n is the number of pairs of scores.

r2 equals the total variation in one variables that can be predicted through its linear association with another variable. Multiple R2 will be familiar to people from a regression context, describes this same property but with multiple predictor variables. When there are two predictor variables,

Where is the correlation between the dependent variable and the first predictor, and

This estimator is increasingly upwardly biased as more predictors are introduced, so {Theil, 1958 #930} proposed an alternative estimator which adjusts for the sample size and number of included predictor variables often called adjusted R2.

Where *n* is the sample size and *k* is the number of predictor variables {Miles, 2004 #932}.

An equivalent effect size to R2 in the context of ANOVA is η2 (eta squared), the ratio of the sums of squares between groups to the total sums of squares.

Where *SSbetween* is The sums of squares between groups, , where *n* is the sample size in each group, is the kth group’s mean and is the grand mean. *SStotal* is equal to the total sums of squares, , with being the ith item’s value on y.

This is different to another related and commonly reported effect size, or partial eta squared, which describes the proportion of variance that can be attributed to a particular factor after excluding variance explained by other factors in the model.

Equation 2, {Levine, 2006 #934}.

Where is the sum of squared residuals, , with being the predicted value of i or equivalently the mean of the group under study. and are equal in one-way ANOVAs as all summed and squared errors are included in the error term, in multiway or repeated measures ANOVA partial eta squared will be larger as the additional factors are not included in the denominator.

Although it has been argued that researchers should favour one over the other (e.g., Levine & Hullett [2006] ague for favouring , whereas Richardson (2011) argues that will be the more meaningful statistic in most cases), both effect sizes are meaningful in different scenarios. If a researcher wants to describe the effect of a particular factor such as a particular manipulation and also has additional factors in their model will be more meaningful, whereas if a researcher wants to describe the total variance explained by all factors in their model they should report . Adding to the confusion between these two statistics is the fact that although was produced in multi-way ANOVAs in SPSS, it was mislabelled as in SPSS versions 7 to 10 (i.e., in some versions of SPSS released before 2001) meaning that it is likely that many of the eta square values in the literature produced using SPSS are in fact partial eta squared {Richardson, 2011 #926}{Levine, 2006 #934}.

can be calculated from F statistics {Richardson, 2011 #926}.

has been criticised in that it will differ between designs when some factors are measured in some designs but not measured in another (e.g., when a covariate is included in some studies, or when a factor that can account for some variance is accounted for such as gender in some analyses but is not included in others). In these cases, partial eta squared will not be comparable across studies or analyses. The variance explained by the covariate or measured factor will be partialled out of the denominator of when the covariate is included in the model, but this variance would be included in the error variance when the measured variable is not included in the model {Olejnik, 2003 #933}.

To make this more concreate take the example of a study on the impact of chocolate advertising on the variable “professed enjoyment of chocolate”. If gender could account for 5% of the variance in “professed enjoyment of chocolate”, the for the impact of chocolate would be higher if it is included in the ANOVA model than if it were not. However, the true impact of the advertisement would not have changed, and the effect sizes from these two studies would not be comparable. Similar issues occur in repeated measure or mixed designs. Correlation between individuals’ scores over levels of a factor (e.g., correlations between individuals’ scores over time) reduce the error sums of squares, decreasing the value of the demominator and inflating the for the factor of interest compared to a between subjects design (i.e., when repeated measures are not taken).

(Generalised eta squared) was developed by {Olejnik, 2003 #933@@author-year} in order to account for this discrepancy. It is identical to except in that it included the measured, non-manipulated factors in the denominator, and the effect itself is only included in the denominator when it is

Or when the effect includes only manipulated factors

With acting as an indicator variable which tales the value of 0 if the effect involves measured factors (e.g., age or sex) or 1 if this is not the case.

WHY IS IT NOT:

However, as R2 is upwardly biased, so is eta squared. ..

Omega square ω2 {Olejnik, 2003 #933}

Epsilon squared too

All estiamotors of the proportion of variance explained

BUT, for sample size determination the most commonly used statistic is **F^2**

**Cohen 1988 uses the f ratio as the main effect size for ANOVA and regression designs. The F ratio**

**Categorical effect sizes:**

References

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to Meta-Analysis*. West Sussex, United Kingdom: John Wiley & Sons.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*, 145-153. doi:10.1037/h0045186

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, New Jersey: Erlbaum.

Cumming, G. (2013). The New Statistics. *Psychological Science, 25*, 7-29. doi:10.1177/0956797613504966

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi:10.3758/bf03193146

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of Effect Size From a Series of Experiments Involving Paired Comparisons. *Journal of Educational Statistics, 18*, 271-279. doi:10.3102/10769986018003271

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics, 6*, 107-128. doi:10.3102/10769986006002107

Hedges, L. V., & Olkin, I. (1985). Statistical Methods for Meta Analysis. San Diego, CA: Academic Press.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-016-1221-4

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. doi:10.3389/fpsyg.2013.00863

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The Other Half of the Story: Effect Size Analysis in Quantitative Research. *CBE Life Sciences Education, 12*, 345-351. doi:10.1187/cbe.13-04-0082

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: the case of r and d. *Psychological Methods, 11*, 386-401. doi:10.1037/1082-989x.11.4.386

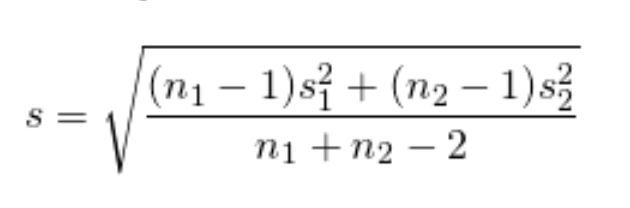
Reiser, B., & Faraggi, D. (1999). Confidence Intervals for the Overlapping Coefficient: the Normal Equal Variance Case. *Journal of the Royal Statistical Society: Series D (The Statistician), 48*, 413-418. doi:10.1111/1467-9884.00199

Rosenthal, R. (1991). Meta-Analytic Procedures for Social Research. Thousand Oaks, California: SAGE Publications, Inc. Retrieved from <http://methods.sagepub.com/book/meta-analytic-procedures-for-social-research>. doi:10.4135/9781412984997

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604. doi:10.1037/0003-066X.54.8.594

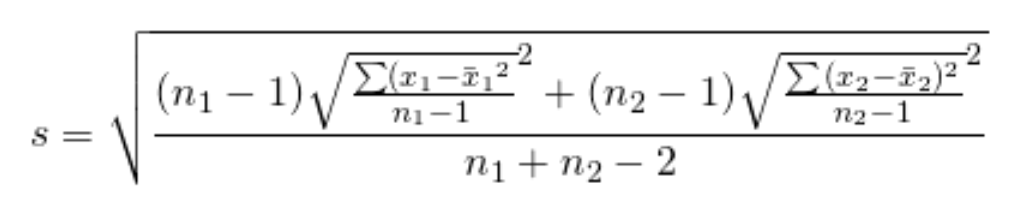
Supplementary material [Cohen’s d/Hedges’ g]

Although the equivalence between formulas x.3 and x.2 is relatively trivial, it seems worth highlighting this equivalence more explicitly as this appears to be a common source of confusion for students and researchers. For example, (Maher et al., 2013) reported that the difference between d and g is that Hedge’s g uses equation x.3 to calculate the pooled standard deviation instead of equation x.2, despite the fact that those formulas are mathematically identical.

 equation x.2

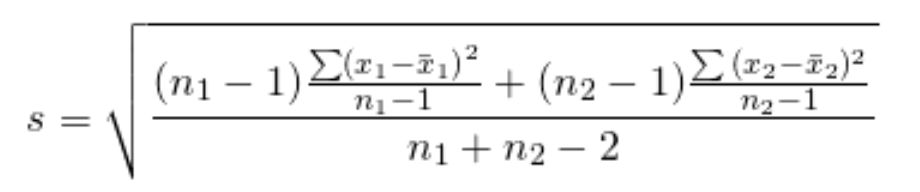
$$s = \sqrt{\frac{(n\_1 -1)s\_1^2 + (n\_2 -1)s\_2^2}{n\_1 + n\_2 - 2}} $$

However, this simplifies to equation x.2 Both “na – 1” and “nb – 1” in the numerator of the fraction cancel out, as can be seen more clearly when the s is replaced with the formula for calculating the standard deviation in [x.2 expanded].

 [x.2 expanded]

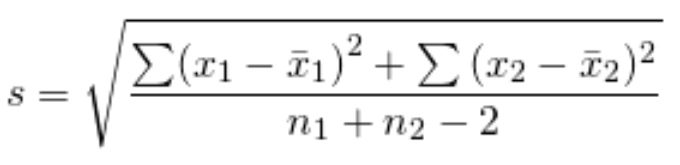
$$s = \sqrt{\frac{(n\_1 -1)\sqrt{\frac{\sum ({x\_1-\bar{x}\_1}^2}{n\_1-1}}^2 + (n\_2 -1)\sqrt{\frac{\sum {(x\_2-\bar{x}\_2)^2}}{n\_2-1}}^2}{n\_1 + n\_2 - 2}}$$

Simple algebra then simplifies this formula to equation [Simplified1].

 [simplified1]

$$ s = \sqrt{\frac{(n\_1 -1)\frac{\sum ({x\_1-\bar{x}\_1)}^2}{n\_1-1} + (n\_2 -1){\frac{\sum {(x\_2-\bar{x}\_2)^2}}{n\_2-1}}}{n\_1 + n\_2 - 2}}$$

Multiplying the elements in the numerator out, we get equation [simplified2], which is identical to equation x.2 above.

 [simplified2]

$$s = \sqrt{\frac{\sum ({x\_1-\bar{x}\_1)}^2 + {\sum {(x\_2-\bar{x}\_2)^2}}}{n\_1 + n\_2 - 2}}$$

EVERYTHING TOGETHER FOR COHEN’S D:

$$d = \frac{\bar x\_1 -\bar x\_2}{

\sqrt{\frac{\sum(x\_1 - \bar{x}\_1)^2 + (x\_2 - \bar{x}\_2)^2}{n\_1 + n\_2 - 2}}}$$

Stats exchange answer:

It seems when people say Cohen's d they mostly mean:

$$d = \frac{\bar{x}\_1 - \bar{x}\_2}{s}$$

Where $s$ is the pooled standard deviation,

$$s = \sqrt{\frac{\sum(x\_1 - \bar{x}\_1)^2 + (x\_2 - \bar{x}\_2)^2}{n\_1 + n\_2 - 2}}$$

There are other estimators for the pooled standard deviation, probably the most common apart from the above being:

$$s^\* = \sqrt{\frac{\sum(x\_1 - \bar{x}\_1)^2 + (x\_2 - \bar{x}\_2)^2}{n\_1 + n\_2}}$$

Notation here is remarkably inconsistent, but sometimes people say that the the $s^\*$ (i.e., the $n\_1 + n\_2$ version) version is called Cohen's $d$, and reserve the name Hedge's $g$ for the version that uses $s$ (i.e., with Bessel’s correction, the n1+n2−2 version). This is a bit weird as Cohen outlined both estimators for the pooled standard deviation (e.g., $s$ version on p. 67, Cohen, 1977) before Hedges wrote about them (Hedges, 1981).

Other times Hedge's g is reserved to refer to either of the bias corrected versions of a standardised mean difference that Hedges developed. Hedges (1981) showed that Cohen's d was upwardly biased (i.e., its expected value is higher than the true population parameter value), especially in small samples, and proposed a correction factor to correct for Cohen's d's bias:

Hedges's g (the unbiased estimator):

$$g = d \* (\frac{\Gamma(df/2)}{\sqrt{df/2 \,}\,\Gamma((df-1)/2)})$$

Where $df = n\_1 + n\_2 -2$ for an independent groups design, and $\Gamma$ is the gamma function.

(originally Hedges 1981, this version developed from Hedges and Olkin 1985, p. 104)

However, this correction factor is fairly computationally complex, so Hedges also provided a computationally trivial approximation that, while still biased, is only biased to an extremely small extent:

Hedges' $g^\*$ (the computationally trivial approximation):

$$ g^\* = d\*(1 - \frac{3}{4(df) - 1})$$

Where $df = n\_1 + n\_2 -2$ for an independent groups design.

(Originally from Hedges, 1981, this version from Borenstein, Hedges, Higgins, & Rothstein, 2011, p. 27)

But, as for what people mean when they say Cohen's d vs. Hedges' g vs. g\*,

people seem to refer to any of these three estimators as Hedge's g or Cohen's d interchangeably, although I've never seen someone write "$g^\*$" in a non-methodology/stats research paper. If someone says "unbiased Cohen's d", you're just going to have to take your best guess at either of the last two (and I think there might even be another approximation that has been used for Hedge's $g^\*$ too!).

They are all virtually identical if $n > 20$ or so, and all can be interpreted in the same way. For all practical purposes, unless you're dealing with really small sample sizes, it probably doesn't matter which you use (although if you can pick, you may as well use the one that I've called Hedges' g, as it is unbiased).

1. See supplementary material [Cohen’s d/Hedges’ g] for a demonstration of the equivalence between x.2 and x.3. This is explicitly provided in the supplementary material as this appears to be a common point of confusion among students and researchers (e.g., (Maher, Markey, & Ebert-May, 2013) misidentifies equation x.3 as the equation for Hedge’s *g* and contrasts that with equation x.2 of Cohen’s *d*, despite their equality.) [↑](#footnote-ref-1)