Distributions and Sampling Variance for Five Effect Size Measures of Variance Overlap

Erin M. Buchanan

John E. Scofield

Missouri State University

Abstract

Distributions and Sampling Variance for Five Effect Size Measures of Variance Overlap

Effect Sizes

Effect size measures play an integral role regarding the outcomes of empirical studies. Effect sizes give clues to whether differences in treatments or experimental conditions are meaningful, and can yield measures of how big, strong, or important an effect really is (Lakens, 2013). Technically speaking, an effect size measure is a quantity or index that estimates an effect in a population that theoretically remains independent of details from experiments such as sample size. These measures quantify the magnitude or strength of differences between different interventions or conditions (Keppel & Wickens, 2004; Olejnik & Algina, 2003). The precision of effect size measures does increase as a positive function of sample size. Larger samples tend to be more precise in terms of effect size estimates. Broadly speaking, lower variances are associated with an increase of sample size, which can yield more precise estimates (Borenstein, Hedges, Higgins, & Rothstein, 2006). The variance of a given effect size estimate not only indicates precision, but also determines confidence interval bands. With smaller sample sizes, confidence intervals of effect sizes can be both unstable and larger than experiments with larger samples (Lakens & Evers, 2014). If a given sample is inadequately small, experimenters can, as Lakens and Evers describe it, “sail the seas of chaos”. Effect sizes, confidence interval bandwidth, and significance levels can be volatile, especially considering the comparisons across studies with small samples.

The d family

While a variety of effect size indices exist, estimates are mainly binned into two families. Briefly, the first family is the *d* family, consisting of mean differences (Lakens, 2013; Rosenthal, 1994). The difference between means is the simplest measure, being most appropriate when two groups are being examined. While mean differences are advantageous in a sense that estimates are reported in original dependent variable units, a single universally acceptable dependent measure may be hard to pin down in the social sciences (Maxwell & Delaney, 2004). To increase the comparability of mean differences, measures are transformed into a standardized scale (i.e. standardized mean differences). Cohen’s *d* represents these mean differences in terms of standard deviation units, which is useful when interpreting effects across experiments. Cohen’s *d* is indeed a popular effect size estimate for two reasons. First, is its simplicity, being used with basic experimental designs with only two conditions. Second, it is an intuitive estimate, with the strength of an effect acting as a function of the degree to which two groups are different (Keppel & Wickens, 2004).

Measures of association strength

The second family, in which this article is focused on, deals with measures of association strength. The social sciences often relate measures of interest (probabilistically) to experimental manipulations. Maxwell and Delaney (2004) express that is can be easy to lose sight that factors investigated or manipulated in experiments are often only a small component of a larger pool of determinants relating to a dependent measure. It is especially easy to lose sight of this, Maxwell and Delaney discuss, when researchers rely solely on hypothesis testing. To avoid interpreting the importance of a statistically significant finding, measures of association can act as a balancing force against this. Measures of association are interpreted as proportions, ranging from zero to one, which measures the variation in the dependent measure that shares a relation with the variation observed within the various levels of the independent variable (Lakens, 2013; Maxwell & Delaney, 2004). Proportion of variance measures act as an alternative effect size measure to standardized mean differences, being more appropriate when three or more levels or groups are present. Proportion of variance effect sizes are modeled by how variation is separated into systematic and unsystematic portions (Keppel & Wickens, 2004). The strength of the effects between levels within an independent variable can be estimated via eta squared, epsilon squared, and omega squared (Hays, 1963; Kelley, 1935; Olejnik & Algina, 2000; Pearson, 1905). These three effect size measures indicate the size or strength of contrasts or an omnibus test, as seen in statistical tests like analysis of variance (ANOVA, Olejnik & Algina, 2003).

Eta squared is calculated by taking the sum of squares for the effect and dividing that by the total sum of squares. However, eta squared is not comparable between studies due to the total variability within a study being dependent on particular characteristics of a study, like the number of variables or levels included in statistical analyses (Lakens, 2013). Using partial eta squared improves comparability across different experiments, and can be calculated by taking the sum of squares for the effect divided by the sum of squares of the effect in addition to the sum of squares associated with the error of a given effect (Keppel, 1991). Eta squared has been observed to have a positive bias, which tends to overestimate the population effect size. (Olejnik & Algina, 2003; Okada, 2013 The extent of the positive bias eta squared holds has been observed to be a decreasing function of sample size (Maxwell & Delaney, 2004). Given the bias of eta squared, Maxwell et al. (1981) explain eta squared as an estimator of the variance associated with a dependent measure shared with the levels within an independent variable, given a respective *sample*. Eta squared has also been criticized by some researchers in terms of its correlational aspect being less appropriate for experimental designs as well as the number and range of levels within an independent variable creating significant differences in effect size estimates (Abelson, 1985; Rosenthal & Rubin, 1982, Okada & Hoshino, 2016). Maxwell and Delaney also have pointed out that these types of estimates can lead to misinterpretation due to differing variance calculations with experimental manipulations. Okada and Hoshino have observed that effect sizes can be vulnerable to the type of research design implemented. Lakens also explains how partial eta squared has problems with comparability across different research designs. Olejnik and Algina further proposed the use of a generalized eta squared to increase comparability across research designs. Generalized eta squared excludes variance from other factors from calculations while including variance associated with individual differences. Alternative, less biased versions of eta squared include omega squared and epsilon squared (Okada, 2013, 2016).

When considering effect size interpretation guidelines, like those of Cohen (1988), researchers should take caution with the evaluations from the terms “small”, “medium”, or “large”. Keppel and Wickens (2004) express how large effects are more commonly known, leaving less to be gained by an increase in reliability. Whereas large effects can bring progress to an existing theoretical basis, Keppel and Wickens explains how these are often a byproduct of incremental advancements in technology or innovative paradigms. In older or well established fields of research, the majority of effects commonly investigated tend to be smaller effects. Effect sizes, whether small or large, also can have increased informative value as presented in tandem with descriptive and inferential statistics, compared to being a sole replacement for hypothesis testing.

Other Stuff

The use of effect size measures, such as those from the *d* and *r* family can be extremely useful when aggregating and synthesizing data, especially with meta-analyses (Keppel & Wickens, 2004). By using indices like effect sizes, researchers can make inferences about the size and direction of effects in a given field of research. However, researchers need to address comparability problems associated with different types of effect size measures, as meta-analyses can be inconclusive unless the type of research design is considered and controlled for (Okada & Hoshino, 2016). Effect sizes can also be used from prior research when planning new studies. *A-priori* power analyses can provide suggested sample sizes given an effect in a field of research (Lakens, 2013).

**Problem statement**

The distribution of effect sizes is relatively unstudied, as most assume they match the distribution of the test statistic. Therefore, *d* values are often calculated with *t* tests because the formulas are mathematically similar (as *d* usually is only a transform of *t* by removing the square root of *N*). However, several researchers have shown that the distribution of *d* is not normal, and actually follows a non-central *t* distribution, even as the central limit theorem approximates normal for the sampling distribution of *t*.

This study focuses on several related measures of variance overlap that are traditionally paired with ANOVA (linear models). Previous research has two implications: 1) the calculation of confidence intervals should not be based on the normal approximation, and 2) meta-analytic techniques that require the estimate of the variance of the effect size sampling distribution are unexplored for this statistic.

**Method**

***Datasets*.** Datasets were simulated using the *rmvnorm* function in the *mtvnorm* package in *R* (citation here). This function simulates multivariate normal data with a given set of means, variances, and correlations between levels. We used this function over the *rnorm* function in *R* to be able to manipulate the relationship between levels and maintain multivariate normal structure. To simulate data that might occur in a real study, we then rounded each data point to a whole number within the confines of a typical 1-7 Likert scale. Therefore, all points below one were truncated to one, and all points above seven were truncated to seven. This procedure did not appear to change the relationship between levels because the data generate was generally within this range because of the choice of means (see below). 1000 datasets were simulated for each of the manipulated variables, resulting in over one million data points.

***Manipulated Variables*.** Four variables were manipulated to examine the influence of each on effect size. Sample size was set to range from *n* = 20 in each level to *n* = 110 in each level, increasing by units of six, thus, creating 16 combinations of sample size. The number of levels or groups in a study are known to influence effect size, with increasing levels creating increasing effects (Okada, 2016), and we therefore manipulated levels from three to six, assuming this set to be within the normal range of levels a researcher might include in a typical categorical independent variable study. Means were set to start at *M* = 2.50 and increase by 0.50 for each level, and therefore, for three levels means were 2.50, 3.00, and 3.50, while six levels ranged from 2.50 to 5.00. While SO AND SO has discussed that typical effect sizes are inflated in repeated measures designs, we know of no study that has examined how the correlation between levels affects multiple effect sizes. Correlation between levels was manipulated with the following constraints: 0, .10, .30, .50, .70, and .90. Last, we changed the expected size of the effect by using *SD2* = 5 for small effects, 3 for medium effects, and 1 for large effects (i.e. *SD* = 2.24, *SD* = 1.73, *SD* = 1).

***ANOVA calculations*.** For each simulated data set, we calculated the following ANOVAs using the *ezANOVA* function from the *ez* package in *R* (cite here): 1) one way between subjects, assuming levels were separate groups of people, 2) one way repeated measures, assuming levels were the same people, 3) two way between subjects, assuming all levels were independent, 4) two way repeated measures, assuming all levels were related, and 5) two way mixed design, analyzing the levels as repeated measures. For two way between subjects and mixed designs, we created a second fake independent variable by generating two levels that evenly split groups into even *n* conditions. For two way repeated measures, we used *rnorm* to generate a second random 1-7 Likert ranged variable and generated two levels that evenly split groups into conditions as with between subjects. Therefore, for the two way designs, we only investigate the main effect generated from the original multivariate normal distribution, and ignored the other main effect and interaction, as they were set to be random and close to zero. In the supplemental material, we provide the calculations for all types of ANOVA, using terminology/symbols from Keppel and Wickens (YEAR) for the interested reader, as these symbols are used for the effect size formulas.

***Effect size calculations*.**

**Results**

**Discussion**

References

Table 1.

*Effect Size Formulas for the Study*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA Design | *ges* | *η2* | *ηp2* | *ω2* | *ωp2* |
| One-way Between |  |  |  |  | NA |
| Two-way Between |  |  |  |  |  |
| One-way Repeated |  |  |  |  | NA |
| Two-way Repeated |  |  |  |  |  |
| Two-way Mixed |  |  |  |  |  |

Supplemental Online Material

**One Way Between Subjects ANOVA**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | *SS* | *df* | *MS* | *F* |
| A |  | a – 1 |  |  |
| S/A |  |  |  |  |
| T |  | N – 1 |  |  |

**One Way Repeated Measures ANOVA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Between Subjects  Comparison | Source | *SS* | *df* | *MS* | *F* |
| A | A |  | a – 1 |  |  |
| S/A | S |  | n – 1 |  |  |
| AXS | S/A – S | (a – 1)(n – 1) |  |  |
| T | T |  | an – 1 |  |  |

**Two Way Between Subjects ANOVA**

**Two Way Repeated Measures ANOVA**

**Two Way Mixed ANOVA**

Term Definition:

* A and B are the independent variables, effects.
* S/A, S/B, S/AB is the error term, residual for between subjects.
* AXS, AXB, AXBXS is the error term, residual for repeated measures
* T is the total variance.

Symbol definition:

* a = number of levels.
* N = total sample size.
* n = sample size for level or condition.
* I = individual participant number.
* J = group number.
* Y = dependent variable score.
* T = grand mean.
* Bar = average.