Exact Solution for Pooled Standard Deviation

Michael K. Ponton
Regent University, School of Education
1000 Regent University Drive, Virginia Beach, VA
United States of America

Alfred P. Rovai
Regent University, School of Education (retired)
1000 Regent University Drive, Virginia Beach, VA
United States of America

Abstract

For meta-analytic purposes, descriptive statistics from multiple studies are often combined in an attempt to provide better estimates for population effect sizes. When comparing means, the pooled estimate of standard deviation is often used as a metric in standardizing mean differences. For some research purposes, however, a reduction of error in estimating the population effect size could be provided if an exact solution to the pooled standard deviation were used instead of an estimate. The purpose of the present article is to outline a method for determining this exact solution.

Keywords: pooled standard deviation, meta-analysis, effect size, descriptive statistics

1. Introduction

Meta-analysis is a statistical method of combining results of independent studies that use different scales of measurement (cf. Schulze, 2004). The solution, first attributed to Gene Glass (1976), is to first estimate a scale-free index for each study. The variation of these indices can then be analyzed across studies. Glass recommended use of effect size when synthesizing research studies that examine group differences. Since quantitative research studies typically report descriptive statistics such as group means and standard deviations, it is a relatively simple procedure for researchers to use these statistics to calculate effect size such as Cohen's *d*. Cohen (1988) defined *d* as the difference between the means of each group, M_1 - M_2, divided by standard deviation, *s*, of either group, assuming homogeneity of variance. Cohen's *d* can be interpreted as standard deviation (*SD*) units. Thus, a small effect size is between .2 and .5 *SD* units, a medium effect size is between .5 and .8 *SD* units, and a large effect size is one that is .8 or more *SD* units (Cohen, 1988; Rosenthal & Rosnow, 1991).

The two groups are considered to be the experimental and control groups. By convention, the subtraction of means is accomplished so that the difference is positive if it is in the direction of improvement or in the predicted direction and negative if in the direction of deterioration or opposite to the predicted direction. In practice, the pooled standard deviation is used to calculate effect size (e.g., Rosnow & Rosenthal, 1996), which represents a weighted average of multiple sample variances; that is, an estimate. The purpose of this paper is to outline a method of determining the exact solution for the pooled standard deviation that can be used instead of an estimate in calculating effect size. *By exact solution we are referring to the standard deviation that would be computed if all data from two or more groups were known*. To use the exact solution, only common descriptive statistics (i.e., sample size, standard deviation, and mean) are needed for each group. As questions continue to arise regarding the use of null hypothesis statistical testing (e.g., Cumming, 2014), efforts that improve effect size calculations and meta-analysis will increase in importance.

Following a meta-analytic purpose, a researcher may be interested in answering the following question: given *k* groups each with sample size *n*, standard deviation *s*, and mean *M*, what is the exact solution for the pooled standard deviation? The exact solution represents the standard deviation that would be calculated if all of the raw data from the *k* groups were combined into a single database and analyzed.

2. Discussion

Let:
k = number of groups
n_i = sample size of i^{th} group
M_i = mean of i^{th} group
$N = \sum_{i=1}^{k} n_i$ = pooled sample size.

The sample statistic variance for the variable *x* is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - M)^2$$

where the standard deviation is $s = \sqrt{s^2}$.
The mean square statistic is given by

$$MS = \frac{n-1}{n}s^2 . \qquad (1)$$

For the pooled sample, the pooled mean is the weighted mean given by

$$M_P = \frac{1}{N}\sum_{i=1}^{k} n_i M_i \ . \qquad (2)$$

The exact solution for the pooled mean square is then given by

$$MS_P = \frac{1}{N}\sum_{i=1}^{k} n_i\left[MS_i + (M_i - M_P)^2\right] \qquad (3)$$

where the pooled variance and standard deviation, respectively, are given by

$$s_P^2 = \frac{N}{N-1}MS_P \qquad (4)$$

$$s_P = \sqrt{s_P^2} \ . \qquad (5)$$

Thus, the procedures outlined from equations (1) – (5) provide the exact solution for the pooled standard deviation. Novel to this procedure is equation (3) that accounts for the change in deviation scores due to the difference between group and pooled means (cf. Hertzog, 1986, that addresses the pooling of covariance matrices). The proof of equation (3) is presented in the Appendix.

2.2 Example

Suppose a researcher is interested in the pooled standard deviation for two groups with descriptive statistics given in Table 1. (Note that actual data were analyzed using SPSS to generate the descriptive statistics presented.)

Thus, for this example
$k = 2$
$N = \sum_{i=1}^{k} n_i$ = 700 + 1500 = 2200.

Using equation (1), the mean square for Group 1 is

$$MS = \frac{n-1}{n}s^2 = \frac{699}{700}2.0847^2 = 4.3398;$$

similarly for Group 2, *MS* = 3.6698.

Using equation (2), the pooled mean is

$$M_P = \frac{1}{N}\sum_{i=1}^{k} n_i M_i = \frac{1}{2200}\left[(700)(6.7279)+(1500)(7.2123)\right] = 7.0582.$$

Using equation (3), the exact solution for the pooled mean square is

$$MS_P = \frac{1}{N}\sum_{i=1}^{k} n_i\left[MS_i + (M_i - M_P)^2\right]$$

$$= \frac{1}{2200}\left\{(700)\left[4.3398+(6.7279-7.0582)^2\right]+(1500)\left[3.6698+(7.2123-7.0582)^2\right]\right\}$$

= 3.9339 .

Using equations (4) and (5), the pooled variance and standard deviation, respectively, are given by

$$s_P^2 = \frac{N}{N-1}MS_P = \frac{2200}{2199}(3.9339) = 3.9357$$

$$s_P = \sqrt{s_P^2} = \sqrt{3.9357} = 1.9839.$$

Compiling the original two groups into a single data file and analyzing the data using SPSS, the descriptive statistics are presented in Table 2. Note that the standard deviation using SPSS is identical to the calculated pooled standard deviation above.

2.2 Comparison to the Pooled Estimate

Without an exact solution, previous analysis relied upon the pooled estimate of the population variance, which represents a weighted average of multiple sample variances. This pooled estimate is given by the following (cf. Hinkle, Wiersma, & Jurs, 1998, p. 357):

$$\hat{s}_P^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)} \quad . \qquad (6)$$

The error produced by using equation (6) instead of the exact procedure is a function of (a) the differences between the means of the original groups and the pooled mean and (b) sample sizes. The effect of (a) is created because if all of the data were combined into a single file, deviation scores would be calculated between each random variable value and the pooled mean rather than each (unpooled) group mean. Following the example in the previous section, the standard deviation calculated for the Group 1 data is based on the deviation scores between individual random variable values and the group mean (i.e., $M_1 = 6.7279$, see Table 1); however, if these data were combined with the Group 2 data then the deviation scores would be calculated with respect to the pooled mean (i.e., $M_P = 7.0582$, see Table 2). Obviously, if there were no differences between group means then the term $(M_i - M_P)^2$ in equation (3) would be zero, and the pooled estimate using equation (6) would be very close to the exact solution for large values of $n_1, n_2, \ldots, n_k$.

Of argument is the notion of why one would ever combine two separate data sets of disparate means into a single data file in the first place. In the field of meta-analysis, results are combined based upon arguments that the studies are separate observations of the same phenomenon; therefore, the premise of the argument is that apples and oranges are not being added in the first place. Even with replication studies using well-controlled procedures, for finite values of $n_1, n_2, \ldots, n_k$ groups means are random variables; therefore, differences between group means is to be expected at least to some degree. Certainly with large values of $n_1, n_2, \ldots, n_k$ one would not assume huge variations in the means produced by different groups of data; however, with small values of $n_1, n_2, \ldots, n_k$ larger variations between means is more likely. As the increase in computational effort for the exact solution is trivial when compared to the pooled estimate, incorporating the exact solution for both large and small sample sizes may be warranted—provided it is consistent with the research purpose.

In addition, there could be research applications where groups of disparate means are intentionally combined. As an example, a researcher may wish to estimate the standard deviation of a random dependent variable (DV) for a population stratified with respect to various levels of a given independent variable (IV). Descriptive statistics from studies focusing on individual levels of this IV may exist with means that vary due to the relationship between the DV and IV; thus, combining these studies would simulate creating a population representative of all levels of the IV. The best estimate of the population standard deviation would be the exact solution rather than the pooled estimate due to errors exacerbated by the DV and IV relationship and its effect on the variation of means between groups. Of course, the researcher would have to assure that the fidelity of population proportions across levels is assured because nonrepresentative proportions would distort the quality of the estimate provided by the exact solution.

However, it is also entirely possible to use the exact solution in a manner that provides an estimate for known group proportions in a population. The prevalent usage of nonprobability samples often produces sample proportions that do not accurately reflect the population. The proposed procedure would be to use known/estimated/desired population proportions for each i^{th} level of the IV in equations (2) and (3) instead of the sample factor n_i / N. For example, suppose acquired data yielded gender percentages of 30% males and 70% females. If a researcher were interested in an estimate of the pooled standard deviation for a population with equal proportions of males and females, .50 would be used for each gender group in equations (2) and (3) instead of the actual measured proportions.

3. Conclusion

Combining statistics across studies in an attempt to provide better estimates of population parameters and effect sizes is critically dependent upon the research purpose. On some occasions, a researcher may argue that the pooled estimate for standard deviation is the appropriate metric for analysis. However, in those situations where an exact solution is argued as appropriate, the procedures outlined in this paper provide such a solution with minimal additional effort required above the pooled estimate computation.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3-8.

Hertzog, C. (1986). On pooling covariance matrices for multivariate analysis. *Educational and Psychological Measurement, 46*, 349-352.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston, MA: Houghton Mifflin.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw Hill.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1,* 331-340.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches.* Cambridge, MA: Hogrefe & Huber.

Appendix

Proof of Exact Solution

If the data for variable *x* associated with *k* groups each with sample size *n* were aggregated into a single database of resultant size *N* and pooled mean M_P, the mean square statistic would be given by the following:

$$MS_P = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ji} - M_P\right)^2 \ .$$

Equation (3) was previously given by

$$MS_P = \frac{1}{N}\sum_{i=1}^{k} n_i\left[MS_i + \left(M_i - M_P\right)^2\right] . \qquad (3)$$

Thus, proof of equation (3) is supported by showing the following equivalence to be true:

$$\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ji} - M_P\right)^2 = \frac{1}{N}\sum_{i=1}^{k} n_i\left[MS_i + \left(M_i - M_P\right)^2\right] .$$

We begin with the following:

$$\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ji} - M_P\right)^2 = \frac{1}{N}\sum_{i=1}^{k} n_i\left[MS_i + \left(M_i - M_P\right)^2\right] .$$

Canceling the 1/*N* term and removing the common outer summation, the problem reduces to showing the following:

$$\sum_{j=1}^{n_i}\left(x_{ji} - M_P\right)^2 = n_i\left[MS_i + \left(M_i - M_P\right)^2\right] .$$

Expanding the terms,

$$\sum_{j=1}^{n_i}\left(x_{ji}{}^2 + M_P^2 - 2x_{ji}M_P\right) = n_i\left(MS_i + M_i^2 + M_P^2 - 2M_iM_P\right) \ .$$

Noting that

$$MS = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - M\right)^2$$

we see that the first term on the RHS can be rewritten as follows:

$$\sum_{j=1}^{n_i}\left(x_{ji}{}^2 + M_P^2 - 2x_{ji}M_P\right) = \sum_{j=1}^{n_i}\left(x_{ji} - M_i\right)^2 + n_i\left(M_i^2 + M_P^2 - 2M_iM_P\right) \ .$$

Expanding the summation on the RHS,

$$\sum_{j=1}^{n_i}\left(x_{ji}{}^2 + M_P^2 - 2x_{ji}M_P\right) = \sum_{j=1}^{n_i}\left(x_{ji}^2 + M_i^2 - 2x_{ji}M_i\right) + n_i\left(M_i^2 + M_P^2 - 2M_iM_P\right) \ .$$

Noting that

$$\sum_{i=1}^{n} x_i = nM$$

And

$$\sum_{i=1}^{n} M = nM \ ,$$

the LHS second and third terms and the RHS summation can be rewritten as follows:

$$\sum_{j=1}^{n_i} x_{ji}^2 + n_iM_P^2 - 2n_iM_iM_P = \sum_{j=1}^{n_i} x_{ji}^2 + n_iM_i^2 - 2n_iM_i^2 + n_iM_i^2 + n_iM_P^2 - 2n_iM_iM_P .$$

Canceling terms on the RHS,

$$\sum_{j=1}^{n_i} x_{ji}^2 + n_iM_P^2 - 2n_iM_iM_P = \sum_{j=1}^{n_i} x_{ji}^2 + n_iM_P^2 - 2n_iM_iM_P .$$

Q.E.D.

Table 1. Descriptive Statistics for Unpooled Data

Group	*n*	*s*	*M*
1	700	2.0847	6.7279
2	1500	1.9163	7.2123

Table 2. Descriptive Statistics for Pooled Data

Group	*n*	*s*	*M*
1+2	2200	1.9839	7.0582