Paper 335-2012

A unified approach to measuring the effect size between two groups using SAS®

Dongsheng Yang and Jarrod E. Dalton

Departments of Quantitative Health Sciences and Outcomes Research
Cleveland Clinic
Cleveland, OH, USA

ABSTRACT

Standardized difference scores are intuitive indexes which measure the effect size between two groups. Compared to a t-test or Wilcoxon rank-sum test, they are independent of sample size. Thus, their use can be recommended for comparing baseline covariates in clinical trials as well as propensity-score matched studies. In this paper, we show how to calculate sample standardized differences for continuous and categorical variables and how to interpret results. We also provide a SAS macro which performs the calculation without using the IML procedure.

INTRODUCTION

In randomized studies, chance imbalance between groups on baseline variables can potentially confound the relationships of interest. Since experimental subjects are randomly assigned to the two groups, it is not appropriate to perform inferential tests regarding the equivalence of population parameters are equivalent. Along these lines, journals often request authors to show baseline summary statistics in a table without hypothesis test *P*-values.

Inferential tests on baseline variables in non-randomized studies can also be troublesome, albeit for a different reason. While it is theoretically justifiable to test for differences in population parameters within a non-randomized sample, the results from these tests are largely dependent on sample size and can be difficult to interpret (e.g., propensity-score matched studies). Nonetheless, researchers and readers still want to assess the comparability of the two groups on baseline characteristics. A unified approach to quantifying the magnitude of difference between groups on baseline variables can thus be helpful for this goal.

Cohen (1962) proposed an effect size index *(*Cohen's d) for the comparison of two sample means [1]. This quantity can be interpreted as a sample-based estimate of the strength of the relationship between two variables in a statistical population; more specifically, it can be interpreted as "a measure of the average difference between means expressed in standard deviation units." [2]. Cohen's *d* is appropriate for assessing effect size based on two symmetrically-distributed samples [3].

However, problems to calculate Cohen's *d* can arise with skewed samples [4]. Yuen (1974) [5] proposed to use robust estimates of means and variances (e.g., trimmed means and Winsorized variances) for the effect size calculation. Cliff (1993, 1996) also proposed a delta statistic to calculate the effect size for ordinal data [6, 7]. Recently, a simple non-parametric effect size statistic was proposed for skewed variables and ordinal variables [8]. This statistic is interpreted as the difference in mean rankings divided by a pooled estimate of the within-group standard deviation of rankings.

In this paper, we show how to calculate sample standardized differences for continuous and categorical variables and how to interpret results. We also provide a SAS macro which performs the calculation without using the IML procedure.

STANDARDIZED DIFFERENCE

Below we describe how the standardized difference is calculated for both continuous and categorical baseline variables:

1. *Continuous baseline variable*

For continuous variables, the standardized difference is

$$d = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{s_1^2 + s_2^2}{2}}} \qquad (1)$$

where $\overline{X_1}$ and $\overline{X_2}$ denote the sample mean of a baseline variable in each group, and s_1^2 and s_2^2 denote the sample variances, respectively. For skewed variables, equation (1) can be modified using rank statistics.

2. *Categorical baseline variables*

For a binary categorical baseline variable, the standardized difference is

$$d = \frac{(\widehat{P_1} - \widehat{P_2})}{\sqrt{\frac{[\widehat{P_1}(1-\widehat{P_1}) + \widehat{P_2}(1-\widehat{P_2})]}{2}}}$$

where $\widehat{P_1}$ and $\widehat{P_2}$ denote the proportion or mean of a binary baseline variable in the treatment and control group, respectively.

For categorical baseline variables with K levels, Dalton (2008) proposed to use a multivariate Mahalanobis distance method to generalize the standardized difference metric to handle a multinomial sample [9]:

Let T = $(\widehat{P_{12}}, \widehat{P_{13}}, \ldots \widehat{P_{1K}})'$

C = $(\widehat{P_{22}}, \widehat{P_{23}}, \ldots \widehat{P_{2K}})'$

where $\widehat{P_{jk}} = \Pr(\text{category } k \mid \text{treatment group } j)$, $j \in \{1,2\}$, and $k \in \{2, 3, \ldots, K\}$.

For example, Table 1 shows the notations for T and C for a hypothetical comparison of blood types between treated and control patients.

Table 1. Notation for estimated conditional probabilities

Baseline variables	Treatment	Control
Blood type		
A	$\widehat{P_{11}}$	$\widehat{P_{21}}$
B	$\widehat{P_{12}}$	$\widehat{P_{22}}$
AB	$\widehat{P_{13}}$	$\widehat{P_{23}}$
O	$\widehat{P_{14}}$	$\widehat{P_{24}}$
Total	1	1

The standardized difference is then defined as

$$d = \sqrt{(T - C)' S^{-1} (T - C)} \qquad (2)$$

Where S is a $(k-1) \times (k-1)$ covariance matrix defined as:

$$S = [S_{kl}] = \begin{cases} \frac{[\hat{P}_{1k}(1-\hat{P}_{1k})+\hat{P}_{2k}(1-\hat{P}_{2k})]}{2}, & k = l \\ \frac{[\hat{P}_{1k}\hat{P}_{1l}+\hat{P}_{2k}\hat{P}_{2l}]}{2}, & k \neq l \end{cases}$$

For a binary baseline variable($K = 2$), we can verify that it is special case of (2).

Here T = $\widehat{P_{12}}$, and C = $\widehat{P_{22}}$

S = [S_{22}] = $\frac{[\widehat{P_{12}}(1-\widehat{P_{12}})+\widehat{P_{22}}(1-\widehat{P_{22}})]}{2}$, $k = l = 2$

From formula (2), we get

$$d = \sqrt{\frac{(\widehat{P_{12}}-\widehat{P_{22}}) \times (\widehat{P_{12}}-\widehat{P_{22}})}{\frac{[\widehat{P_{12}}(1-\widehat{P_{12}})+\widehat{P_{22}}(1-\widehat{P_{22}})]}{2}}} = \frac{|(\widehat{P_{12}}-\widehat{P_{22}})|}{\sqrt{\frac{[\widehat{P_{12}}(1-\widehat{P_{12}})+\widehat{P_{22}}(1-\widehat{P_{22}})]}{2}}}$$

3. Confidence interval for standardized difference

Hedges and Olkin (1985) provided a formula to calculate the confidence interval for standardized difference [10]. A 95% confidence interval for d: d ± 1.96× σ[d].

Where σ[d] = $\sqrt{\frac{n_1+n_2}{n_1 \times n_2} + \frac{d^2}{2(n_1+n_2)}}$, n_1 and n_2 denote sample sizes in each group, respectively.

INTERPRETATION

An effect size (d) can be treated as equivalent to a Z-score of a standard normal distribution. Cohen (1988) related d to three different measures of non-overlap between two populations (**Table 2)**: the percentage of non-overlap of the two distributions (U_1), the percentage in the second population that exceeds the same percentage in the first population (U_2), and the percentage of the first population which the upper half of the second population exceeds (U_3) [3].

Table 2. Interpretations of effect sizes

Effect Size	Percent of Non-Overlap of two populations (U_1)	Percentage in the second population that exceeds the same percentage in the first population (U_2)	Percentage of the first population which the upper half of the second population exceeds (U_3)	Common Language Effect Size (CLES)
0.0	0.0	50.0	50.0	0.50
0.1	7.7	52.0	54.0	0.53
0.2	14.8	54.0	57.9	0.56
0.3	21.3	56.0	61.8	0.58
0.4	27.4	57.9	65.5	0.61
0.5	33.0	59.9	69.1	0.64
0.6	38.2	61.8	72.6	0.66
0.7	43.0	63.7	75.8	0.69
0.8	47.4	65.5	78.8	0.71
0.9	51.6	67.4	81.6	0.74
1.0	55.4	69.1	84.1	0.76
F(d) = the cumulative normal distribution function of d; U_3 = F(d); U_2 = F($\frac{d}{2}$) ; U_1 = $\frac{2U_2-1}{U_2}$; CLES = F($\frac{d}{\sqrt{2}}$)				

For example, a standardized difference of 0.2 indicates that there is 15% of non-overlap in the two distributions (U_1), that 54% of control group observations with values greater than 54% of treatment group observations (U_2), and that the mean of the treated group is at the 58th percentile of the control group. (U_3) [11].These can be visualized in **Figure 1**.

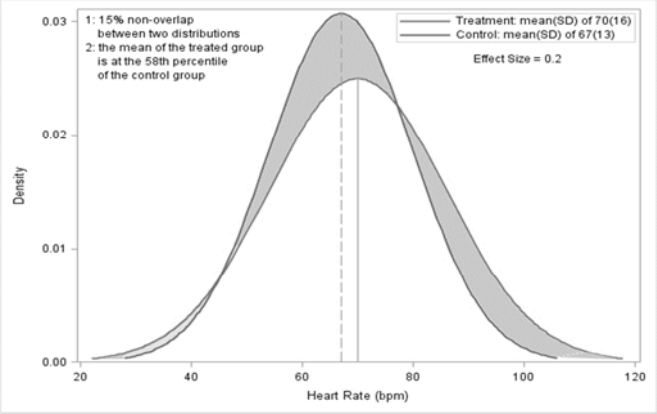

Fig 1. The distribution of heart rate by study groups.

McGraw and Wong (1992) proposed a 'Common Language Effect Size' (CLES) statistic to interpret the effect size, indicating the probability that a randomly selected score from the population of the treatment group will be greater than a randomly sampled score from the distribution of the control group [12]. For example, if we assume that the distributions are normal, then an effect size of 1.0 indicates the probability that a randomly selected participant in the treatment group will be higher than a randomly selected participant in the control group is 76%.

Cohen (1988) suggested that Effect Size Indices of 0.2, 0.5, and 0.8 can be used to represent small, medium, and large effect sizes, respectively. According to Cohen, “a medium effect of .5 is visible to the naked eye of a careful observer. A small effect of .2 is noticeably smaller than medium but not so small as to be trivial. A large effect of .8 is the same distance above the medium as small is below it.” [3]

Austin (2009) proposed methods to “estimate the empirical sampling distribution of the standardized difference of the mean under the assumption that the mean (or prevalence) of a covariate is equal between two groups” [13]. He suggested that one can use $1.96 \times \sqrt{\frac{2}{n}}$ to decide a cut-off value of an effect size, assuming the two groups have equal number of subjects ($n_1 = n_2 = n$).

SAS MACRO

It is easy to calculate a Mahalanobis distance for a categorical variables using PROC IML. However, some users may not have a license to use PROC IML. So we developed a SAS macro to calculate the effect size only using SAS/BASE. In there, we use PROC ORTHOREG to determine if an inverse matrix is a singular. A sample call would look somehow like the following:

```
%stddiff(
        inds = temp,                    /*  input SAS data */
        groupvar = treatment,           /*  a group variable: must be coded as 0 and 1 */
        numvars = age bmi/r asa/r,      /*   a list of continuous variables.  */
                                        /*  “/r” indicates a ranked-based method
                                            used to calculate  standardized difference */
        charvars = female Race,         /*  a list of categorical variables */
        stdfmt = 5.2,                   /*   a format of standardized difference */
        ouds =                          /*   output data set */
```

);

After calling the macro, standardized differences for each baseline variable will be reported. **Table 3** is recommended for clinical trials or propensity-score matched studies to summarize baseline characteristics by study groups.

Table 3. Baseline characteristics by study groups*

Factor	Treatment (N = 200)	Control (N = 180)	Standardized Difference †
Age (year)	51 ± 15	51 ± 13	-0.03
Body Mass Index (kg/m^2)	26 [21,30]	27 [22, 40]	-0.22
Female	67 %	56 %	0.22
ASA physical status			0.62
I	6 %	3 %	
II	44 %	74 %	
III	50 %	24 %	
Race (White vs. other)	100 %	94 %	0.35
* Data are presented as mean ± SD, median [inter-quartiles] or %. † Standardized difference = difference in means or proportions divided by standard error; imbalance defined as absolute value greater than 0.20 (small effect size)			

LIMITATIONS

One of the limitations of the effect size is that there is no accepted threshold to determine the significant difference between two groups. Another is that the effect size calculated from the Mahalanobis distance method does not have a direction. The third one is that population heterogeneity increases error variance and reduces the magnitude of the effect size [14, 15].

CONCLUSION

Standardized difference is an intuitive index to compare baseline characteristics in both randomized and non-randomized studies.

REFERENCES

1. Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review.Journal of Abnormal and Social Psychology, 65, 145-153.
2. Graziano, A. M., & Raulin, M. L. (2010, in press). Research Methods: A process of inquiry (7th ed.). Boston: Allyn & Bacon. Retrieved March 1, 2012. http://www.mikeraulin.org/graziano7e/supplements/effectsz.htm
3. Cohen, J. (1988). Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed). Lawrence Erlbaum Associates Publishers: Hillsdale, NJ.
4. Kraemer, H. C., & Andrews, G. A. (1982). A nonparametric technique for meta-analysis effect size calculation. Psychological Bulletin, 91, 404-412.
5. Yuen, K.K. (1974). The two-sample trimmed t for unequal population variances. Biometrika (1974) 61 (1): 165-170.
6. Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin, 114, 494-509.
7. Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. Multivariate Behavioral Research, 31, 331-350.

8. Schacht, A, Bogaerts K, Bluhmki E, Lesaffre E, A New Nonparametric Approach for Baseline Covariate Adjustment for Two-Group Comparative Studies. Volume 64, Issue 4, pages 1110–1116, December 2008
9. Dalton, J.E. (2008) *A new standardized difference metric for multinomial samples.* Unpublished work.
10. Hedges LV, Olkin I. (1985). Statistical Methods for Meta-Analysis. Academic Press: San Diego, CA .
11. Carson, C. (n.d.) The effective use of effect size indices in institutional research. Retrieved March 1, 2012 (http://www.keene.edu/ir/effect_size.pdf).
12. McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. Psychological Bulletin, 111, 361-365
13. Peter C. Austin. (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statist. Med. 2009; 28:3083–3107
14. Fern, F. E., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. Journal of Consumer Research, 23, 89–105.
15. Stephen Olejni &James Algina. (2000). Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations Contemporary Educational Psychology 25, 241–286

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Dongsheng Yang
Enterprise: Cleveland Clinic
City, State ZIP: 44195
E-mail: yangd@ccf.org

Name: Jarrod Dalton
Enterprise: Cleveland Clinic
City, State ZIP: 44195
daltonj@ccf.org

Web: http://www.lerner.ccf.org/qhs/

APPENDIX

Due to page limits, please contact authors to obtain a free copy of the SAS macro.

```
/*************************************************************************/
/* Program      : stddiff.sas
/* Purpose      : SAS macro to calculate the Standardized Difference
/* Usage        : %stddiff(inds = Studydata, groupvar = dex,
/*                         numvars = age bmi/r glucose,
/*                         charvars = female surgtype,
/*                         stdfmt = 8.5,
/*                         outds = std_result);
/*************************************************************************/
/* NOTE:             All binary variables must be coded as 0 and 1 in the dataset
/* PARAMETERS:
/*  inds:             Input dataset
/*  groupvar:         a binary variable, must be coded as 0 and 1
/*  numvars:          a list of continuous variables.
/*                    "/r" denotes to use the rank-based
/*                    mean and SD to calculate Standardized Difference
/*  charvars:         a list of categorical variables. A binary categorical
/*                    variable must be coded as 0 and 1. The level = 0 is as
/*                    a reference level.
/*  stdfmt = 8.5      the format of Standardized Difference
/*  outds             output result dataset
/*************************************************************************/
```