Research article

Effect sizes of the differences between means without assuming variance equality and between a mean and a constant

Satoshi Aoki

Graduate school of Science, the University of Tokyo, Bunkyo, Tokyo, Japan

A R T I C L E   I N F O

Keywords:
Psychology
Mathematics
Effect size
Standardized mean difference
Cohen’s d
Hedges’ d
Heteroscedasticity

A B S T R A C T

Effect sizes of the difference, or standardized mean differences, are widely used for meta-analysis or power-analysis. However, common effect sizes of the difference such as Cohen’s d or Hedges’ d assume variance equality that is fragile and is often violated in practical applications. Based on Welch’s t tests, we defined a new effect size of the difference between means, which did not assume variance equality, thereby providing a more accurate value for data with unequal variance. In addition, we presented the unbiased estimator of an effect size of the difference between a mean and a known constant. An R package is also provided to compute these effect sizes with their variance and confidence interval.

1. Introduction

Effect sizes of the difference or, more precisely, standardized mean differences between two groups, are widely used to estimate the magnitude of effect independent of the sample size [1], to conduct meta-analysis [2], or to conduct power-analysis [3]. The American Educational Research Association (AERA) or the American Psychological Association (APA) strongly recommend effect sizes are reported in the corresponding fields [4, 5]. Furthermore, the misuse and misunderstanding of p-value have become public [6], and use of effect sizes is spreading beyond pedagogy and psychology, where effect sizes have developed, into areas such as in biology [1]. In spite of such importance, the classical effect sizes of the difference assume variance equality (homoscedasticity), which is hard to assume practically or is even expected to be violated a priori in clinical data [7]. While Bonett [8] defined a confidence interval of an effect size estimator which did not assume homoscedasticity, its parameter was not defined. This problem of variance inequality (heteroscedasticity) has been long debated [9, 10]. In addition, the unbiased estimator of an effect size of the difference between a mean and a constant was undefined. To solve these problems, based on Welch’s t test [11, 12], we defined an effect size of the difference between means that does not assume homoscedasticity and calculated the unbiased estimator of an effect size of the difference between a mean and a constant.

Effect size of the difference was developed by Cohen [13], who studied in the field of psychology. Cohen [3, 13] defined the effect size as a parameter for two independently and normally distributed populations, $Y^1 \sim N(\mu_1, \sigma^2)$ and $Y^2 \sim N(\mu_2, \sigma^2)$:

$$\delta = (\mu_1 - \mu_2)/\sigma,$$

which is expressed as $d$ in the original articles [3, 13]. Note that both populations share the common variance $\sigma^2$. The estimator of this parameter was represented as $d_i$ in [3]. However, we refer to this estimating statistic as $g$ to distinguish it from the other $d$ we introduce later. The statistic $g$ is defined as

$$g = (Y^1 - Y^2)/\text{S}^{\text{pooled}},$$

where

$$\text{S}^{\text{pooled}} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}},$$

and, for $i = 1, 2$,

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (Y^i_j - \bar{Y}^i)^2}{n_i - 1}.$$

(3)

Here, $Y^1$, $Y^1_j$, and $n_1$ are the mean of the sample, the sample (random variable), and the sample size of group 1, respectively, while $Y^2$, $Y^2_j$, and $n_2$ are those of group 2. For the denominator, this effect size uses the pooled standard deviation, which suggests the most precise population variance under the assumption of equal variance [14].

* Corresponding author.

E-mail addresses: aoki171@g.ecc.u-tokyo.ac.jp, aokis1ll1@gmail.com.

https://doi.org/10.1016/j.heliyon.2020.e03306

Received 20 September 2019; Received in revised form 17 October 2019; Accepted 23 January 2020

2405-8440/© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
In the field of pedagogy, Glass [2] suggested another effect size of the difference, independently of Cohen’s works. He defined it as “the mean difference on the outcome variable between treated and untreated subjects divided by the within group standard deviation,” where “the within group standard deviation” corresponds to the standard deviation of the untreated group. He clearly distinguished the treated (experimental) group from the untreated (control) group, and there was no assumption regarding the two groups. His effect size was subsequently formulated and named Glass’ Δ by Hedges [14], which is

\[ \Delta = \frac{(Y^E - Y^C)}{S^C}, \]  

(4)

where \( Y^E \) is the mean of the variable in the experimental group, \( Y^C \) is that in the control group, and \( S^C \) is the unbiased standard deviation of the control group.

Hedges [14] also defined the \( d \) (1) and the \( g \) (2) independently of Cohen. Furthermore, Hedges [14] indicated that \( g \) (2) is biased from \( d \) (1), making it unsuitable for analyses that do not treat the entire population. The unbiased estimator of \( d \) (1) is defined as \( g^E \) in [14] and \( d \) in [15]. In this study, we call it \( d \), which is

\[ d = J(n_1 + n_2 - 2) \gamma. \]  

(5)

Using the gamma function, the correction coefficient \( J \) is defined as

\[ J(m) = \frac{\Gamma(m/2)}{\sqrt{m/2} \Gamma((m - 1)/2)}. \]  

(6)

The effect sizes \( g \) (2) and \( d \) (5) are used in various fields of science, but they assume homoscedasticity just like Student’s t-test [16, 17]. When this assumption of homoscedasticity is violated, Grissom [9] recommended the use of Glass’s \( \Delta \) (4) instead of \( d \) (5). However, Glass’s \( \Delta \) (4) and \( d \) (5) have different meanings because of the difference in denominator. Therefore, Glass’s \( \Delta \) (4) cannot substitute for \( d \) (5) in a strict sense. Behavior of \( g \) (2), \( \Delta \) (4), and \( d \) (5) under heteroscedasticity was studied in [10], although the justification for using effect size parameter \( a \), that they defined, to measure the statistic bias under heteroscedasticity was not shown.

Bosett [8] in psychology proposed a confidence interval (CI) of effect size which does not assume homoscedasticity. First, he defined a general effect size estimator

\[ \hat{d} = \sum_{j=1}^{k} c_j \bar{Y}_j / s, \]  

where \( \sum_{j=1}^{k} c_j = 0 \), \( \bar{Y}_j \) is a sample mean, and \( s^2 = \sqrt{\sum_{j=1}^{k} \sum_{j=1}^{k} S_j^2} \). Concerning effect size of the difference between two means, substituting \( k = 2 \), \( c_1 = 1 \) and \( c_2 = -1 \) gives

\[ \hat{d} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{(S_1^2 + S_2^2)/2}}. \]  

(8)

Then, he assumed its corresponding parameter and its CI. The CI was calculated using approximation of CI [18] and variance of the estimator which was approximately calculated without assuming homoscedasticity. The parameters estimated by \( \hat{d} \) (7) or (8) were not formulated. Namely, he defined the CI for heteroscedasticity without defining a parameter, and this can be a problem. When the estimator does not always correspond to a single parameter, the CI of an undefined parameter loses its consistency in what to estimate, and heteroscedasticity or difference of sample sizes can change the correspondence between an estimator and a parameter (see section 5.2). Although his CI was effective relative to the other CIs in his simulation experiment where the parameter was given a value, what the value meant could change depending on the variance and sample size, and the change could not be expected since the parameter was not formulated.

It should be noted, Cohen [3] also defined a parameter of an effect size of the difference between a mean and a constant for a normally distributed population \( N_i(\mu_i, \sigma_i^2) \) and a known constant \( C \) as

\[ \gamma = (\mu - C) / \sigma_i. \]  

(9)

Cohen [3] originally referred to this as \( d_i^* \), but we refer to this as \( \gamma \) (9) to clearly distinguish it from \( d \) (5). Cohen [3] also defined a biased estimator of an effect size for a normally distributed population with the sample value \( Y_i^j \) (\( i = 1, \ldots, n_i \)), the sample mean \( \bar{Y}_i \), and a known constant \( C \) as

\[ e_{\text{biased}} = (\bar{Y}_i - C) / \sigma_i. \]  

(10)

The \( s_i \) is the square root of (3). Cohen [3] originally referred to this as \( d_i^* \), but we refer to this as \( e_{\text{biased}} \) for the reason described above. To the best of my knowledge, the unbiased estimator of \( \gamma \) (9) has not been shown.

When there are other effect sizes of the difference that do not assume normality or independence. Since their assumption is different from that of effect size we focus on, we do not treat them in detail and briefly introduce them. Dunlap et al. [19] invented effect size of the difference between two correlated paired groups. Algina et al. [20] proposed robust effect size of the difference, which is based on \( g \) (2) using 20% trimmed mean and 20% Winsorized variance assuming that samples are taken from an observing population and another contaminating population.

2. Theory

2.1. An effect size of the difference between means without assuming homoscedasticity

First, we define the parameter of an effect size of the difference between means for two independently and normally distributed populations \( N_1(\mu_1, \sigma_1^2) \) and \( N_2(\mu_2, \sigma_2^2) \)

\[ e = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + r \sigma_2^2) / (r + 1)}} \]  

(11)

where \( r \) is a non-negative real number. This parameter is not generalization of \( d \) (1) and is different from it. Then, suppose two independently and normally distributed populations with the samples \( Y_i^1 \) (\( i = 1, \ldots, n_1 \)) and \( Y_i^2 \) (\( i = 1, \ldots, n_2 \)), and the sample mean \( \bar{Y}_1 \) and \( \bar{Y}_2 \). Based on the statistic \( t_w \), the so-called Welch’s \( t \) [11, 12], a biased estimator of \( e \) (11) is defined as

\[ e_{\text{biased}} = t_w / \sqrt{n}. \]  

(12)

Finally, \( e \), the unbiased estimator of \( e_{\text{biased}} \) (11), is

\[ e = e_{\text{biased}} J(f). \]  

(15)

Therefore,

\[ E(e) = e. \]

Here, \( r \) corresponds to the ratio \( n_1 \) to \( n_2 \). \( J \) is the correction coefficient that is defined in equation (6). The degree of freedom \( f \) is approximately calculated using the Welch-Satterthwaite equation [11, 21] as

\[ f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{s_1^2/n_1^2(n_1 - 1) + s_2^2/n_2^2(n_2 - 1)}. \]  

(16)

The variance of \( e \) (15) is

\[ \text{var}(e) = \frac{f}{f - 2} J^2(f)(1/\hat{\theta} + c_i^2) - c_i^2. \]
Although this effect size is derived from the difference, we refer to it as \( \epsilon \) not \( d \). This is because Cohen’s \( d \) (2) and Hedges’ \( d \) (5) already exist, and more \( d \) would cause further confusion. The proof of the bias correction and variance derivation does not assume homoscedasticity (see the Appendix). In addition, \( e \) (15) is a consistent estimator of \( \epsilon \), (11) at the same time. See the Appendix for the proof of the consistency.

2.2. An effect size of the difference between a mean and a known constant

Using \( e^{biased} \) (10), the unbiased estimator of the effect size parameter \( \gamma \) (9) is defined for a normally distributed population with the sample value \( Y^1 \) (\( i = 1, \ldots, n_i \)), the sample mean \( \bar{Y}^1 \), and a known constant \( C \) as

\[
e = e^{biased} J(n_1 - 1).
\]

Therefore,

\[
E(e) = \gamma.
\]

The correction coefficient \( J(6) \) is the same as the one used above. The variance of \( e \) is

\[
var(e) = \frac{n_i}{n_i - 3} J^2(n_1 - 1)(\frac{1}{n_i - 1} + \gamma^2) - \gamma^2.
\]

See the Appendix for proofs of the bias correction and the derivation of the variance. In addition, \( e \) (17) is a consistent estimator of \( \gamma \) (9) (see the Appendix for the proof). When interested in constants rather than variables, \( e \) defined as

\[
e' = (C - \bar{Y}^1) J(n_1 - 1)/s_i
\]

can be used instead of \( e \).

2.3. Confidence intervals of effect sizes

In terms of the effect sizes of the difference, the CI based on a noncentral \( t \) variate is not directly given by a formula [22]. The CI is derived from that of noncentral parameters of noncentral \( t \)-distribution, which is in turn obtained by some searching method. The CI based on the biased effect sizes are given as:

\[ [ncpl/\sqrt{n_1}, ncpp/\sqrt{n}] : CI \text{ based on } g \text{ and } e^{biased}. \]

and

\[ [ncpl/\sqrt{n_i - 1}, ncpp/\sqrt{n_i - 1}] : CI \text{ based on } e^{biased}. \]

where \( ncpl \) is the noncentral parameter that gives the upper limit of cumulative probability (e.g., 0.975 cumulative probability for 95\% CI) for noncentral \( t \)-distribution with the corresponding \( v \) value (see the discussion section) and the degree of freedom, and \( ncpp \) is that which gives the lower limit (e.g., 0.025 cumulative probability for 95\% CI), and \( n_1 \) and \( n_i \) are the same as \( (14) \) and \( (10) \). The CIs based on the unbiased estimator of the effect sizes are given by multiplying the corresponding correction coefficient \( J(6) \) of the corresponding degree of freedom to the above intervals.

The CI by Bonett [8] is calculated using variance of the estimator which is approximately calculated without assuming homoscedasticity and approximate assumption of CI [18]. Therefore, it is not necessary to apply Bonett’s CI to \( e \) (15) or \( e \) (17), because the derivation of their CIs does not assume homoscedasticity, and their exact CIs can be calculated without approximation.

3. Calculation method

I developed a new package esdif for R [23]. It enables the statistics \( d \) (5), \( e \) (15), \( e \) (17), their biased statistics, variance, and CI based on the two samples or their mean, variance, and sample size to be computed. In this package, approximation of \( J(6) \) [14] is not employed unless its degree of freedom exceeds 342, when the gamma function returns values that are too large to be treated in \( R \). The CI is obtained by binary search.

The remainder of this section presents some examples of the package. First, the following script calculates \( d \) (5), \( e \) (15), their variances and 95\% CIs for data 1 \{(0,1,2,3,4)\} and data 2 \{(0,0,1,2,2)\}.

\[
> \text{library(esdif)}
> \text{data1<-c(0,1,2,3,4)}
> \text{data2<-c(0,0,1,2,2)}
> \text{es.d(data1,data2)}
\]

[1] [2] "Hedges’ d: " "0.682379579393545" [2] "variance: " "0.484026380702367" [3] "CI: " "[ -0.5035272167375147 , 1.82938058482178 ]"

Using options of the function, you can change the type I error rate for the CI, calculate biased effect sizes, and output results in the vector style. For example, \( e^{biased} \) (10) with 99\% CI in the vector style is calculated by this script.

\[
> \text{library(esdif)}
> \text{data1<-c(0,0,1,2,2)}
> \text{data2<-c(2)}
> \text{es.e(data1,data2,alpha=0.01, unbiased=FALSE, vector_out=TRUE)}
\]

[1] -1.0000000 0.59290237 -2.5390625 0.5778885

In the vector-style output, the four values in the vector show the effect size, its variance, and lower and higher limits of the CI. In addition, this package includes functions that can output effect sizes from the (estimated) parameters and the sample sizes. The following scripts compute \( d \) (5) and \( e \) (15) for two populations, \( N(1,2) \) and \( N(0,1) \) with the sample size 5 and 10, respectively.

\[
> \text{library(esdif)}
> \text{mean1<-1}
> \text{mean2<-0}
> \text{var1<-2}
> \text{var2<-1}
> \text{nl<-5}
> \text{n2<-10}
> \text{es.para.d(mean1,mean2,var1,var2,n1,n2)}
\]

[1] [2] "Hedges’ d: " "0.822865297143979" [2] "variance: " "0.349443397657368" [3] "CI: " "[ -0.2488827678382689 , 1.86661883367494 ]"

\[
> \text{es.para.e(mean1,mean2,var1,var2,n1,n2)}
\]

[1] [2] "Unbiased e: " "0.674259756444758" [2] "variance: " "0.41613476136966" [3] "CI: " "[ -0.354146439977423 , 1.65626025590509 ]"

These types of functions also have the options for the type I error rate, the biased effect size, and the vector-style output.
4. Application & Simulation

While the situation to use $c$ (17) is clearly different, the $e$ (15) and $d$ (5) have a similar application range in practice. Therefore, we prepared an example of the applications in which the sample variances are not equal. Table 1 shows well-known data of three Iris species by Fisher [24], which can also be checked in R [23] using a command “iris”. Note that only the petal width of $I. setosa$ has fewer significant digits. For this data, we calculated $d$ (5), $e$ (15), the ratio of $d$ (5) to $e$ (15), and the ratio of the standard deviations of the two comparing data. Theoretically, $e$ (15) is a more precise estimator of its own parameter than $d$ (5) under this heteroscedasticity. The calculated result is shown in Table 2. When considering their significant digits, the comparing pair of the sepal length of $I. setosa$ and $I. virginica$ showed the different effect size of $d$ (5) and $e$ (15) (in bold in Table 2). Even though most pairs showed identical values of $d$ (5) and $e$ (15), the result revealed that violation of the assumption of homoscedasticity in $d$ (5) can affect the result even in two significant digits.

Fig. 1 shows the ratio of $d$ (5) to $e$ (15) plotted against the ratio of standard deviations of the comparing data. This figure shows that the similar two standard deviations give similar $d$ (5) and $e$ (15). In other words, the more different two standard deviations encourage the use of $e$ (15) over $d$ (5) more.

To examine the nature of $d$ (5) and $e$ (15), we also conducted a simulation study. In addition to $d$ (5) and $e$ (15), Bonett’s statistic $\hat{\delta}$ (8) was also included as a reference, although its accuracy cannot be discussed because of the lack of the parameter definition. The above effect sizes and their width of 95% CI were calculated for 100,000 Monte Carlo replications from $N(1, \sigma_1^2)$ and $N(0, \sigma_2^2)$ for each condition, and
they were represented by their average values. The population means were fixed to 1 and 0. The sample sizes were changed from 10 to 30 by 10. The population standard deviation \( \sigma_1 \) was fixed to 1 and \( \sigma_2 \) was changed to 10 by 1. However, some redundant data were omitted from the result. The calculation was conducted using es.diff R package shown above and metafor R package [25]. The R source code used for the simulation was shown in the Appendix.

Table 3 shows the result of the simulation. When the sample size ratio was conserved under \( \sigma_1 \neq \sigma_2 \), e (15) gave more similar and concordant values than \( d \) (5). For example, e (15) for \( n_1 = n_2 = 10, 20, 30 \) under \( \sigma_2 = 10 \) were 0.142, 0.140 and 0.141, whereas the corresponding \( d \) (5) were 0.148, 0.143 and 0.143. This is the nature and advantage of e (15) which is designed to estimate the same parameter under heteroscedasticity and the same sample size ratio. The width of CI was narrowest for \( d \) (5) under \( \sigma_1 = \sigma_2 \), and e (15) had the second narrowest. Under \( \sigma_1 \neq \sigma_2 \), e (15) and \( \delta \) (8) had the narrowest CI under \( n_1 = n_2 \), \( n_1 > n_2 \), and \( n_1 < n_2 \), respectively. The narrowest CIs of e (15) were followed by \( d \) (5), whereas what followed the narrowest CIs of \( \delta \) (8) was not fixed. It was shown that e (15) had wider situation under which it had the narrowest or second narrowest CI than \( d \) (5) or \( \delta \) (8). Bonett’s statistic \( \delta \) (8) equaled to \( d \) (5) under \( n_1 = n_2 \) as their definition. Under \( n_1 \neq n_2 \) and \( \sigma_1 \neq \sigma_2 \), e (15) was closer to \( \delta \) (8) than \( d \) (5). This might imply relative accuracy of \( \delta \) (8) over \( d \) (5) under heteroscedasticity.

5. Discussion

5.1. Correspondence of effect sizes and t tests

Comparison of t tests and the effect sizes of the difference except \( \delta \) (8) shows the clear correspondence between them (Table 4). Statistic d (5) corresponds to the unpaired two-sample t test [16, 17], whose statistic is the basis of g (2). Statistic \( \delta \) biased (12) uses the statistic (13) of Welch’s t test [12], which aims to test two means with unequal variances, and \( \delta \) un biased (10) uses the same statistic as the one-sample t test [17]. Considering this, it is natural that power analyses should be conducted, using the corresponding pair of the effect size and t test. In other words, power analyses of Student’s one-sample t test, Student’s unpaired two-sample t test, and Welch’s t test should be conducted based on the c statistic (17), d (5), and the e statistic (15), respectively. Co-use of non-similar-t-test and effect size causes inconsistency of the assumption about the population(s).

5.2. Influence of sample size on effect size

In this subsection, the relationship between the effect sizes of the difference and sample sizes is described. The value of g (2), a biased estimator of the effect size of the difference under homoscedasticity, is independent of the sample sizes when the assumption of homoscedasticity (\( s_1 = s_2 \)) is fulfilled. When \( s_1 \neq s_2 \), it depends on the ratio \( q = (n_1 - 1)/(n_2 - 1) \) as implied in [9]. This is because g (2) is no longer an estimator of \( \delta \) (1) under \( s_1 \neq s_2 \), and it will be a biased estimator of the other parameter \( \delta \), which is

\[
\delta' = \frac{\mu_1 - \mu_2}{\sqrt{(q \sigma_1^2 + \sigma_2^2)/(1 + q)}},
\]

Note that even d (5) cannot be the unbiased estimator of \( \delta' \) when \( s_1 \neq s_2 \), because g (2) is not distributed as non-central t variate in this situation. Even if \( n_1 \) and \( n_2 \) vary, g (2) roughly estimates the same parameter, given the ratio q is fixed.

Next, the \( \delta \) biased (12) is a biased estimator of e (11), but e (11) equals to the other parameters in the particular situation. When \( s_1 = s_2 \), \( e = \delta \), and \( \delta \) biased (12) equals to g (2), and is independent of the sample sizes. When \( s_1 \neq s_2 \) and \( n_1 = n_2 \), e (15) is its unbiased estimator. Therefore, usage of e (15) is always preferable to d (5) in this situation. When \( s_1 \neq s_2 \) and \( n_1 \neq n_2 \), \( \delta \) biased (12) depends on the rate \( r = n_1/n_2 \). Therefore, strictly speaking, multiple \( \delta \) biaseds can be comparable only when the sample size ratio r is identical.

The effect size estimator \( \delta \) (8) did not have a defined parameter, but when \( n_1 = n_2 \) and \( s_1 = s_2 \), \( \delta \) (8) equals to g (2) and \( \delta \) unbiased (12), and is independent of sample size. Under \( n_1 = n_2 \) and \( s_1 \neq s_2 \), \( \delta \) (8) also equals to g (2) and suffers from the same problem as it. Under \( n_1 \neq n_2 \) and \( s_1 \neq s_2 \), the value of \( \delta \) (8) is no longer the same as g (2), and precise discussion on its behavior is hindered by the lack of its parameter definition. When trying to consider \( \delta \) (8) as a noncentral t- variate like the other effect sizes, its degree of freedom should be about \( n_1 + n_2 - 2 \), and \( n_1 \) and \( n_2 \) should affect the degree of freedom under \( s_1 \neq s_2 \).

Unlike g (2) or e \( \delta \) biased (12), e (15) is always independent of the sample size.

The behavior of the unbiased estimator of the effect sizes (d (5), e (15), and e (17)) are almost identical to those that are biased, but they slightly increase as the sample sizes become large. This is because of the correction coefficient J (6), and its behavior is illustrated in detail in [14].

In summary, in terms of the effect size of the difference between two means, usage of e (15) is preferable to d (5) or \( \delta \) (8), and e (15) can be the remedy for application of effect size of the difference under
Table 3

Comparison of effect sizes in simulation.

| n   | d.ES | e.ES | B.ES | d.Par. | e.Par. | B.Par. | d.CI | e.CI | B.CI |
|-----|------|------|------|--------|--------|--------|------|------|------|
| 10  | 1.000| 1.000| 0.995| 1.000  | 1.000  | U.D.   | 1.823| 1.828| 1.911|
| 10  | 0.998| 1.000| 0.997| 1.000  | 1.002  | U.D.   | 1.579| 1.664| 1.646|
| 10  | 0.999| 1.000| 0.998| 1.000  | 1.000  | U.D.   | 1.579| 1.664| 1.646|
| 10  | 0.999| 1.000| 0.997| 1.000  | 1.002  | U.D.   | 1.579| 1.664| 1.646|
| 10  | 0.999| 1.000| 0.998| 1.000  | 1.000  | U.D.   | 1.579| 1.664| 1.646|
| 10  | 0.999| 1.000| 0.998| 1.000  | 1.000  | U.D.   | 1.579| 1.664| 1.646|
| 10  | 0.999| 1.000| 0.998| 1.000  | 1.000  | U.D.   | 1.579| 1.664| 1.646|

Note: \(d\) = effect size \(d\) (5); \(e\) = effect size \(e\) (15); \(B\) = effect size \(\delta\) (8); \(Par\.) = parameter of effect size; \(CI\) = width of confidence interval; N.C. = not calculable; U.D. = undefined. The narrowest CI in each row is underlined.

Table 4

Correspondence of assumptions, \(t\) values, and effect sizes of the difference.

| As. | One sample & a constant | Two samples under homoscedasticity | Two samples under heteroscedasticity |
|-----|--------------------------|----------------------------------|-------------------------------------|
|     | Normality                | Normality, Independence, & Homoscedasticity | Normality & Independence |
| \(t\) | \(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S^2 (n_1 + n_2)}}}\) | \(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S^2 (n_1 + n_2)}}}\) | \(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S^2 (n_1 + n_2)}}}\) |
| \(ES\) | \(\frac{\bar{Y}_1 - \bar{Y}_2}{S_1 + S_2}\) | \(\frac{\bar{Y}_1 - \bar{Y}_2}{S_1 + S_2}\) | \(\frac{\bar{Y}_1 - \bar{Y}_2}{S_1 + S_2}\) |

Note: \(As.\) = assumption; \(t\) = \(t\) value; \(ES\) = effect size. The degree of freedom of \(J\) is omitted for the space and must be calculated corresponding degree of freedom.

heteroscedasticity. However, when the ratio of the two sample sizes cannot be set as uniform under heteroscedasticity, neither \(d\) (5) nor \(e\) (15) can be precisely compared. This is a form of the Behrens-Fisher problem, which cannot be solved strictly.

5.3. Potential applications of the new effect sizes

The effect size \(e\) (15) has a vast applicable range covering all kinds of natural and social sciences. This is because \(e\) (15) corresponds to Welch's \(t\) test, whose use is nowadays encouraged over Student's \(t\) test (e.g., [26]). The effect size \(e\) (15) is the best option, especially when the ratio of the sample sizes of two groups can be fixed. The effect size \(c\) (17) has a relatively narrower range regarding the application. In comparison of paired two groups (the difference in pairs vs. 0) and in some simulation studies (result of simulation vs. the optimal value) or physics (result of experiment vs. physical constant), an effect size of the constant may be needed.

Declarations

Author contribution statement

Satoshi Aoki: Conceived and designed the analysis; Analyzed and interpreted the data; Contributed analysis tools or data; Wrote the paper.

Funding statement

This study was partly supported by National Bioresource Project from AMED, grant number 16km0210053j0005.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://CRAN.R-project.org/package= esdif.

Acknowledgements

I would like to thank Prof. Motomi Ito and Dr. Fumio Tajima for their contribution to the manuscript and Dr. Masakazu Shimada for highlighting a miscalculation in the manuscript.
Appendix A

A.1. Proofs of unbiasedness and variation of $e$

In summary, this proof is an application of the proof in [14] to the statistic $c$ in [11]. Suppose two independently and normally distributed populations $N_1(\mu_1, \sigma_1^2)$ and $N_2(\mu_2, \sigma_2^2)$. Their sample means are $\bar{Y}_1$ and $\bar{Y}_2$, and their samples are $Y_i^1$ ($i = 1, \ldots, n_1$) and $Y_i^2$ ($i = 1, \ldots, n_2$). The statistic $c_{\text{biased}}$ (12) between them can be converted into

$$\sqrt{n_{1-1}} c_{\text{biased}} = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

where

$$w = \frac{s_1^2/n_1 + s_2^2/n_2}{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}.$$

Here, since $N_1$ and $N_2$ are independently and normally distributed, the numerator of (18) has the normal distribution of $N(\theta, 1)$, where

$$\theta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

and the $s_i^2$ is the same as (3). In the denominator, $w/f$ is approximately distributed as $\chi^2(f)$ [11]. Therefore, $\sqrt{n_{1-1}} c_{\text{biased}}$ is distributed as a non-central $t$ variate with the non-centrality parameter $\theta$ and approximate degree of freedom $f$ (16). From the nature of the non-central $t$ distribution [27], the expected value of $c_{\text{biased}}$ (12) is

$$E(\sqrt{n_{1-1}} c_{\text{biased}}) = \theta \sqrt{n}/J(f).$$

$$E(c_{\text{biased}}) = \theta \sqrt{n}/J(f).$$

Now, assuming $r = n_1/n_2$, then $\theta/\sqrt{n} = c_r$. In this case, the expected value of $e$ (15) is

$$E(e) = E(c_{\text{biased}} J(f)) = (\theta/\sqrt{n}) J(f) = c_r.$$

Thus, $e$ (15) is an unbiased estimator of $c_r$ (11). The variation of $e_{\text{biased}}$ (12) is

$$\text{var}(\sqrt{n_{1-1}} c_{\text{biased}}) = \frac{f}{f-2} (1 + \theta^2 - \theta^2/2) J^2(f)/J(f).$$

$$\text{var}(c_{\text{biased}}) = \frac{f}{f-2} (1/\sqrt{n} + \theta^2/\sqrt{n}) - (\theta/\sqrt{n})^2.$$

Therefore, the variation of $e$ (15) is

$$\text{var}(e) = \text{var}(c_{\text{biased}} J(f)) = \frac{f}{f-2} J^2(f)(1/\sqrt{n} + (\theta/\sqrt{n})^2) - (\theta/\sqrt{n})^2.$$

$$= \frac{f}{f-2} J^2(f) (1/\sqrt{n} + c_r)^2 - c_r^2.$$

A.2. Proofs of unbiasedness and variation of $c$

The bias correction and derivation of the variance can be proved in the same way as that of $d$ (5). The statistic $c_{\text{biased}}$ (10) can be converted into

$$\sqrt{n-1} c_{\text{biased}} = \frac{(\bar{Y}_1 - C)}{\sqrt{n_1 - 1}}.$$

and this (19) is distributed as a non-central $t$ variate with non-centrality parameter $\mu - C/\sqrt{n_1 - 1}$ and degree of freedom $n_1 - 1$. Therefore, the expected value $c_{\text{biased}}$ (10) is

$$E(\sqrt{n_{1-1}} c_{\text{biased}}) = \frac{\mu - C}{\sigma/\sqrt{n_1 - 1}} \frac{(n_1 - 2)/2}{\Gamma((n_1 - 1)/2)}.$$

Therefore, $E(c_{\text{biased}}) = (\mu - C)/(\sigma/\sqrt{n_1 - 1}) = \gamma$. Thus, $c$ is an unbiased estimator of the effect size parameter $\gamma$ (9). The variation of $c_{\text{biased}}$ (10) is

$$\text{var}(\sqrt{n_{1-1}} c_{\text{biased}}) = \frac{n_1 - 1}{n_1 - 1 + 1}(1 + (\mu - C)^2/\sigma^2/\sqrt{n_1 - 1}).$$

$$\text{var}(c_{\text{biased}}) = \frac{n_1 - 1}{n_1 - 3}(\frac{1}{n_1 - 1} + (\mu - C)^2/\sigma^2).$$

Therefore, the variation of $c$ (17) is

$$\text{var}(c) = \text{var}(c_{\text{biased}} J(n_1 - 1)) = \frac{n_1 - 1}{n_1 - 3} J^2(n_1 - 1) + (\gamma^2 - \frac{1}{J^2(n_1 - 1)}).$$

A.3. Proof of consistency of $c$

First, we treat the proof of $c$, which is simpler than that of $e$. For the proof, we introduce a lemma.

**Lemma 1.** Assume random samples $Y_1^1, \ldots, Y_1^n$ from the population with the population mean $\mu$, and the population variance $\sigma^2$, and consider a parameter $\beta$ and its statistic $b = b(Y_1^1, \ldots, Y_1^n)$. Then, $\lim_{n \to \infty} \text{var}(b) = 0 \Rightarrow b$ is a consistent estimator of $\beta$.

**Proof.**

$$E(|b - \beta|^2) = E(b - \beta)^2 + \text{var}(b)$$

$$= (E(b) - E(\beta))^2 + \text{var}(b)$$

$$= (E(b) - \beta)^2 + \text{var}(b)$$

Given $E(b) = \beta$ and $\lim_{n \to \infty} \text{var}(b) = 0$,

$$\lim_{n \to \infty} (|E(b) - \beta|^2 + \text{var}(b)) = 0.$$

Therefore, $b$ is a mean square consistent estimator of $\beta$, namely,

$$\lim_{n \to \infty} E(|b - \beta|^2) = 0.$$


Here, for an arbitrary positive number \( \epsilon \), by applying Chebyshev’s inequality \([28]\), we get

\[
P(\|b - \beta\| \geq \epsilon) = P(\|b - \beta\|^2 \geq \epsilon^2) \\
\leq \frac{\text{Var}(b)}{\epsilon^2}.
\]

From the result shown above, we can say

\[
\lim_{\delta \to \infty} E(\|b - \beta\|^2)/\epsilon^2 = 0.
\]

Therefore, using the squeeze theorem, we get

\[
\lim_{\delta \to \infty} P(\|b - \beta\| \geq \epsilon) = 0.
\]

Thus, \( b \) is a consistent estimator of \( \beta \). \( \square \)

Now, we move on to the proof of \( c \) \((17)\). When \( n_1 \to \infty \), the variance of \( c \) will be

\[
\lim_{n_1 \to \infty} \text{Var}(c) = \frac{n_1 - 1}{n_1 - 1} \cdot f^2(n_1 - 1)(\frac{1}{n_1 - 1} + \gamma^2) - \gamma^2 \\
= 1 \cdot f^2(\infty)(\frac{1}{\infty} + \gamma^2) - \gamma^2 \\
= 0.
\]

Thus, \( \lim_{n_1 \to \infty} \text{Var}(c) = 0 \), and \( c \) is an unbiased estimator of \( \gamma \). Therefore, based on Lemma 1, \( c \) \((17)\) is a consistent estimator of \( \gamma \) \((9)\). \( \square \)

**A.4. Proof of consistency of \( e \)**

On the other hand, \( e \) \((15)\) consists of two populations. Therefore, a variation of the previous lemma is necessary.

**Lemma 2.** Assume two random samples \( Y_1, \ldots, Y_{n_1} \) and \( Y_2, \ldots, Y_{n_2} \) from the two mother populations with the population means \( \mu_1 \) and \( \mu_2 \), and the population variances \( \sigma_1^2 \) and \( \sigma_2^2 \), respectively. Consider a parameter \( \beta \) and its statistic \( b = h(Y_1, \ldots, Y_{n_1}; Y_2, \ldots, Y_{n_2}) \). Then, \( b \) is an unbiased estimator of \( \beta \), and \( \lim_{n_1, n_2 \to \infty} \text{Var}(b) = 0 \) \( \Rightarrow \) \( \{ b \) is a consistent estimator of \( \beta \} \)

This lemma can be proved in the same way as Lemma 1.

Now, consider \( n_1 = n \) and \( n_2 = n \) to assume \( \phi = \infty \), which equals to \( (n_1, n_2) \to (\infty, \infty) \). Note that \( r > 0 \) and \( \theta > 0 \), since \( n_1 \geq 1 \) and \( n_2 \geq 1 \). Using \( r \) and \( \phi \), \( f \) \((6)\) and \( \hat{\gamma} \) \((14)\) can be expressed as

\[
f = \frac{(s_1/r + s_2/r)}{(s_1^2/\sigma_1^2) + (s_2^2/\sigma_2^2)}
\]

and

\[
\hat{\gamma} = \frac{r \phi}{r + 1}.
\]

Therefore, when \( \phi \to \infty \), the variance of \( e \) \((15)\) will be

\[
\lim_{\phi \to \infty} \text{Var}(e) = \frac{\text{Var}(e)}{\epsilon^2} = \lim_{\phi \to \infty} f^2(\hat{\gamma}(\frac{1}{\hat{\gamma}} + \epsilon^2) - \epsilon^2 \\
\leq \frac{\text{Var}(\hat{\gamma})}{\epsilon^2} = \lim_{\phi \to \infty} \frac{1}{1 - 2\gamma} \cdot f^2(\infty)(\frac{1}{\infty} + \gamma^2) - \gamma^2 \\
= \lim_{\phi \to \infty} \frac{1}{1 - 0} \cdot \gamma^2 - \gamma^2 \\
= 0.
\]

The limit does not contain \( r \), meaning \( \lim_{n_1, n_2 \to \infty} \text{Var}(e) \) always gives an identical value 0. Also, \( e \) is an unbiased estimator of \( \epsilon \), \((11)\). Therefore, based on Lemma 2, \( e \) \((15)\) is a consistent estimator of \( \epsilon \), \((11)\). \( \square \)

**A.5. Simulation source code**

The simulation study in this article was conducted using the following code in R:

```r
library(es.dif)
library(Matrix)
library(metafor)

rep<-100000
n_sd<-10
sampleSize<-matrix(0,nrow=9,ncol=2)
sampleSize[1,]<-c(10,10)
sampleSize[2,]<-c(10,20)
sampleSize[3,]<-c(10,30)
sampleSize[4,]<-c(20,10)
sampleSize[5,]<-c(20,20)
sampleSize[6,]<-c(20,30)
sampleSize[7,]<-c(30,10)
sampleSize[8,]<-c(30,20)
sampleSize[9,]<-c(30,30)

dc<-numeric(rep)
var_dc<-numeric(rep)
ci_lb_dc<-numeric(rep)
ci_ub_dc<-numeric(rep)

e<-numeric(rep)
var_e<-numeric(rep)
ci_lb_e<-numeric(rep)
ci_ub_e<-numeric(rep)

Bonett<-numeric(rep)
var_Bonett<-numeric(rep)
ci_lb_Bonett<-numeric(rep)
ci_ub_Bonett<-numeric(rep)

result_dc<-matrix(0,nrow=(n_sd)*nrow(sampleSize),ncol=4)
result_e<-matrix(0,nrow=(n_sd)*nrow(sampleSize),ncol=4)
result_Bonett<-matrix(0,nrow=(n_sd)*nrow(sampleSize),ncol=4)
counter<-1
for(k in 1:nrow(sampleSize))
{
  for(i in 1:n_sd)
  {
    for(j in 1:nrow(sampleSize[k,]))
    {
      data1<-rnorm(sampleSize[k,1],1,1)
data2<-rnorm(sampleSize[k,2],0,1)

temp_d<-es.d(data1,data2,
vector_out=T)
d[i]<-temp_d[1]
var_d[i]<-temp_d[2]
ci_lb_d[i]<temp_d[3]
ci_ub_d[i]<temp_d[4]
temp_e<-es.e(data1,data2,
vector_out=T)
e[i]<temp_e[1]
var_e[i]<temp_e[2]
ci_lb_e[i]<temp_e[3]
ci_ub_e[i]<temp_e[4]

```
temp_Bonett <- summary(escalc(measure="SMDH", m1i=mean(data1), m2i=mean(data2), sd1i=sd(data1), sd2i=sd(data2), n1i=length(data1), n2i=length(data2)))
Bonett[1] <- temp_Bonett[1,1]
var_Bonett[i] <- temp_Bonett[1,2]
ci_lb_Bonett[i] <- temp_Bonett[1,5]
ci_ub_Bonett[i] <- temp_Bonett[1,6]
result_d[counter,] <- c(mean(d), mean(var_d), mean(ci_lb_d), mean(ci_ub_d))
result_e[counter,] <- c(mean(e), mean(var_e), mean(ci_lb_e), mean(ci_ub_e))
result_Bonett[counter,] <- c(mean(Bonett), mean(var_Bonett), mean(ci_lb_Bonett), mean(ci_ub_Bonett))
counter <- counter + 1

References

[1] S. Nakagawa, I. Cuthill, Effect size, confidence interval and statistical significance: a practical guide for biologists, Biol. Rev. Camb. Philos. Soc. 82 (4) (2007) 591–605.
[2] G. Glass, Primary, secondary, and meta-analysis of research, Educ. Res. 5 (10) (1976) 3–8.
[3] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Academic Press, 1969.
[4] American Educational Research Association, Standards for reporting on empirical social science research in AERA publication, Educ. Res. 35 (6) (2006) 33–40.
[5] American Psychological Association, Publication Manual of the American Psychological Association, 6th ed., American Psychological Association, 2009.
[6] R. Wasserstein, N. Lazar, The ASA statement on p-values: context, process, and purpose, Am. Stat. 70 (2) (2016) 129–133.
[7] R. Grissom, Heterogeneity of variance in clinical data, J. Consult. Clin. Psychol. 68 (1) (2000) 155–165.
[8] D.G. Bonett, Confidence intervals for standardized linear contrasts of means, Psychol. Methods 13 (2) (2008) 99–109.
[9] R. Grissom, J. Kim, Review of assumptions and problems in the appropriate conceptualization of effect size, Psychol. Methods 6 (2) (2001) 135–146.
[10] P. Marfo, G. Okyere, The accuracy of effect-size estimates under normals and contaminated normals in meta-analysis, Heliyon 5 (6) (2019) e01838.
[11] B. Welch, The significance of the difference between two means when the population variances are unequal, Biometrika 29 (3) (1938) 330–362.
[12] B. Welch, The generalization of “Student’s” problem when several different population variances are involved, Biometrika 34 (1) (1947) 28–35.
[13] J. Cohen, The statistical power of abnormal-social psychological research: a review, J. Abnorm. Psychol. 65 (3) (1962) 145–153.
[14] L. Hedges, Distribution theory for Glass’s estimator of effect size and related estimators, J. Educ. Stat. 6 (2) (1981) 107–128.
[15] L. Hedges, I. Olkin, Statistical Methods for Meta-Analysis, Academic Press, 1985.
[16] Student, The probable error of a mean, Biometrika 6 (1) (1908) 1–25.
[17] R. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, 1925.
[18] W. Viechtbauer, Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design, J. Educ. Behav. Stat. 32 (1) (2007) 39–60.
[19] W. Dunlap, J. Cortina, J. Vaslow, M. Burke, Meta-analysis of experiments with matched groups or repeated measures designs, Psychol. Methods 1 (2) (1996) 170–177.
[20] J. Algina, H. Keselman, R. Penfield, Alternative to Cohen’s standardized mean difference effect size: robust parameter and confidence interval in the two independent group case, Psychol. Methods 10 (3) (2005) 317–328.
[21] F. Satterthwaite, Synthesis of variance, Psychometrika 6 (5) (1941) 309–316.
[22] G. Cumming, S. Finch, A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions, Educ. Psychol. Meas. 61 (4) (2001) 532–574.
[23] R core team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2019.
[24] R. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[25] W. Viechtbauer, Conducting meta-analyses in R with the metafor package, J. Stat. Softw. 36 (3) (2010) 1–48.
[26] G. Ruxton, The unequal variance t-test is an underused alternative to Student’s t-test and the Mann-Whitney U test, Behav. Ecol. 17 (4) (2006) 688–690.
[27] N. Johnson, B. Welch, Applications of the non-central t-distribution, Biometrika 31 (3–4) (1940) 362–389.
[28] P. Chebyshev, De valeurs moyennes, J. Math. Pures Appl. 12 (2) (1867) 177–184.