Cohen's *d* for two independent samples

Paul White,

Applied Statistics Group,

Faculty of Environment and Technology,

Univeristy of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul.white@uwe.ac.uk

Paul Redford,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul2.redford@uwe.ac.uk

James Macdonald,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

james.macdonald@uwe.ac.uk

***Abstract*—** An exposition of standardized effect for two independent samples (under an assumption of normality) is given along with an insight into interpretation and reporting.

***Keywords*—** *standardized effect size, independent samples*

I. Introduction

A common misunderstanding in null hypothesis significance testing is the incorrect reasoning that a *p*-value [1] quantifies the strength of a relationship or the strength of a difference. This is not true. When a relationship or difference exists, the *p*-value is a joint function of *both* the strength of the relationship (i.e. effect size) and sample size.

In a correlation study, the effect size is quantified by the (usually unknown) true correlation coefficient and may be estimated by the sample derived correlation coefficient, *r*.

In a difference study, the effect size is quantified by the (usually unknown) true difference between the means, and may be estimated by the sample derived difference, $\bar{x}_1 - \bar{x}_2$. For instance, in a weight loss study, suppose a particular intervention showed a statistically significant weight loss of 5kg. The estimated, or sample derived, effect size would be 5kg.

In certain studies, the outcome of interest might not be measured on a ratio scale with meaningful units of measurement (unlike height, or weight) but may be measured on interval-like scales of measurement (such as anxiety, or depression, or body image). In these latter situations, it might not be meaningful to talk about (say) a change of 5 without any units of measurement. For this reason, it is preferable to consider a scaled measurement of effect size known as a standardized effect size (δ). For two populations, with an assumed common variance (σ^2), the population standardised effect size may be defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

where μ_1 and μ_2 are the respective population means.

To motivate matters this brief note will consider the standardised effect size for two idealised normal distributions with a common variance (σ^2). Section III will then consider the two-sample situation where the standardised effect size is quantified using Cohen's *d*.

II. Standardised effect for idealised Normal Distributions

Consider two normal distributions (Distribution A with mean μ_A and Distribution B with mean μ_B) and with a common variance σ^2.

To simplify matters, and without any loss of generality, let's further assume the common variance is equal to 1. Figure 1 gives an example in which Normal Distribution A has a mean of 0 and standard deviation 1, and Normal Distribution B has a mean of 0.8 and a standard deviation of 1. In this example situation it may be verified that the effect size and the standardized effect size is δ = 0.8.

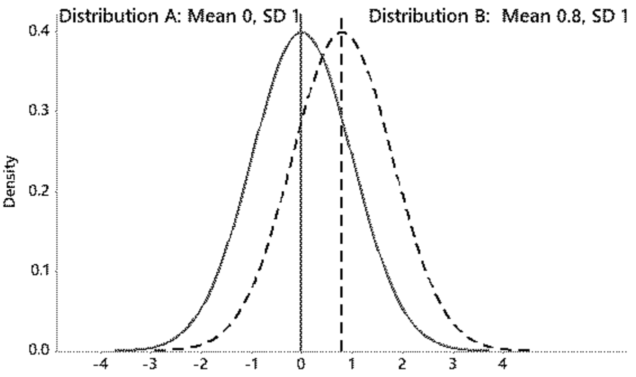

Figure 1 Standardized effect size = 0.8.

Now suppose we were to take a single random sample from Distribution B (as shown in Figure 1) and a single random sample from Distribution A (as shown in Figure 1). What is the probability that the observed sample value from Distribution B would have a value greater than the observed sample value Distribution A? The answer to this question is 0.66.

Note that with Distribution A held fixed (mean zero, standard deviation 1) we could consider the above question for different mean values for Distribution B. These type of calculations are summarized in Table 1 for standardized effect size 0(0.1)2 (i.e. 0 to 2 in steps of 0.1). Table 1 gives the mapping between the standarised effect size, δ, and the probability that a randomly selected observation from Distribution B will be greater in value than a randomly selected observation from Distribution A and this mapping holds for any two normal distributions with constant variance.

Table 1 Probability ($\boldsymbol{\pi}$) of a randomly selected observation from a normal distribution ($\boldsymbol{\mu_B}, \boldsymbol{\sigma}$) being greater in value than a randomly selected observation from a normal distribution ($\boldsymbol{\mu_A}, \boldsymbol{\sigma}$) for given standardized effect, $\boldsymbol{\delta}$

δ	π	δ	π	δ	π
0.0	0.50	0.7	0.64	1.4	0.76
0.1	0.52	0.8	0.66	1.5	0.77
0.2	0.54	0.9	0.67	1.6	0.79
0.3	0.56	1.0	0.69	1.7	0.80
0.4	0.58	1.1	0.71	1.8	0.82
0.5	0.60	1.2	0.73	1.9	0.83
0.6	0.62	1.3	0.74	2.0	0.84

In a similar way, we could ask the question, "*what is the probability that a randomly selected observation from Distribution B would be above the mean value of Distribution A?*" For δ = 0.8 the answer to this question is 0.79. These type of calculations are summarized in Table 2 for standardized effect size 0(0.1)2. Table 2 gives the mapping between the standardized effect size (δ), and the probability that a randomly selected observation from Distribution B will be greater in value than the mean value from Distribution A and this mapping holds for any two normal distributions with constant variance. Also note that these probabilities when multiplied by 100 represent the percentile of Distribution A at the position of the mean of Distribution B.

Table 2 Probability ($\boldsymbol{\pi}$) of a randomly selected observation from a normal distribution ($\boldsymbol{\mu_B}, \boldsymbol{\sigma}$) being greater in value than the mean of a normal distribution ($\boldsymbol{\mu_A}, \boldsymbol{\sigma}$) for a given standardized effect, $\boldsymbol{\delta}$

δ	π	δ	π	δ	π
0.0	0.50	0.7	0.76	1.4	0.91
0.1	0.54	0.8	0.79	1.5	0.93
0.2	0.58	0.9	0.82	1.6	0.95
0.3	0.62	1.0	0.84	1.7	0.96
0.4	0.66	1.1	0.86	1.8	0.96
0.5	0.69	1.2	0.88	1.9	0.97
0.6	0.73	1.3	0.90	2.0	0.98

Inspection of Figure 1 shows that there is distributional overlap between Distribution B and Distribution A. For the situation in Figure 1 (δ = 0.8) the degree of overlap is 0.526 and hence the degree of nonoverlap is 0.474. These type of calculations are summarized in Table 3 for standardized effect size 0(0.1)2. Table 3 gives the mapping between the standardized effect size (δ) and the degree of nonoverlap or *separation* and this mapping holds for any two normal distributions with constant variance.

Table 3 Degree of non-overlap (separation) between two normal distributions with common variance for given standardized effect, $\boldsymbol{\delta}$

δ	Separation	δ	Separation	δ	Separation
0.0	0.000	0.7	0.430	1.4	0.681
0.1	0.077	0.8	0.474	1.5	0.707
0.2	0.147	0.9	0.516	1.6	0.731
0.3	0.213	1.0	0.554	1.7	0.754
0.4	0.274	1.1	0.589	1.8	0.774
0.5	0.333	1.2	0.622	1.9	0.794
0.6	0.382	1.3	0.653	2.0	0.811

Relatedly, suppose there is an observed value of -1 and based on Figure 1 we had to guess whether the observation was from Distribution A or Distribution B. In this case we hedge our bets and guess an observation of -1 would have come from Distribution A as an observation of -1 is far removed from Distribution B and relatively closer to the mean of Distribution A. In general, given an observed value we would "guess" that the observation would come from the distribution with the closer mean. If we used this rule, then we could ask "what is the probability of getting the decision correct?" Table 4 summarises these probabilities for delta 0(0.1)2.

Table 4 Probability ($\boldsymbol{\pi}$) of a correct guess in allocating an observation to one of two normal curves with equal variance for given standardized effect, $\boldsymbol{\delta}$

δ	π	δ	π	δ	π
0.0	0.50	0.7	0.64	1.4	0.76
0.1	0.52	0.8	0.66	1.5	0.77
0.2	0.54	0.9	0.67	1.6	0.79
0.3	0.56	1.0	0.69	1.7	0.80
0.4	0.58	1.1	0.71	1.8	0.82
0.5	0.60	1.2	0.73	1.9	0.83
0.6	0.62	1.3	0.74	2.0	0.84

In summary, for two normal distributions with constant variance, any value of δ can be interpreted in terms of the probability that a randomly selected observation from Distribution B has a larger value than a randomly selected observation from Distribution A; that a randomly selected observation from Distribution B will exceed the mean of Distribution A; mapped to the degree of non-overlap (separation) or overlap between the two distributions; mapped to the probability of correctly guessing group membership. So, for instance, if δ = 1.2, then the probability that a randomly selected observation from Distribution B being greater in value than a randomly selected observation from Distribution B is 0.73; if δ = 1.2, then there is an 88% chance a randomly selected observation will exceed the mean of Distribution A; if δ = 1.2 then the degree of distributional separation is 0.622; if delta = 1.2 then there is a 73% chance of correctly guessing group membership.

III. Cohen's d

The foregoing has considered

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

as a measure of effect size for two normal distributions. Cohen [2] proposed a sample estimate effect size for two independent groups to be

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

where $\bar{x}_1$, and $\bar{x}_2$ are the sample means, and *s* is the sample estimate of the population standard deviation, σ . In experimental research there is an argument to use the standard deviation from the control group (because the intervention might affect both the mean and standard deviation in the experimental arm). However, in practice, the standard deviation in the denominator of

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

is typically obtained by pooling the sample standard deviations i.e.

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and then by taking the square root.

Table 1, Table 2, Table 3 and Table 4 can be used to help interpret δ and therefore give some insight into interpreting d. Table 5 is an attempt to give a less abstract way of visualising the magnitude of Cohen's *d*. Table 5 gives estimated *d* derived for comparing the heights of females for different age groups (data sourced from the World Health Organisation). For instance, height would be an excellent discriminator between girls aged 5 and girls aged 8 (*d* = 3.20), but height would not be a good discriminator between girls age 15 and girls aged 16 (*d* = 0.05).

For social science research, Cohen tentatively suggested that values of *d* > 0.8 be considered a large effect (i.e. one in which an effect should be reasonably obvious to a reasonable person just by viewing the data); that values of at least *d* = 0.5 be considered to be a medium sized effect (i.e. one which other knowledgeable researchers would consider to be important but not necessarily obvious); that *d* between 0.2 and 0.5 be classed as small (i.e. with diminished practical utility); and non-zero values below 0.2 to be potentially of theoretical but not necessarily of practical interest.

Table 5 Cohen's *d* for height by age groups

Age groups	*d*	Age groups	*d*
16 and 17	0.05	9 and 10	0.98
15 and 16	0.12	8 and 9	1.00
14 and 15	0.27	7 and 8	1.00
13 and 14	0.50	6 and 7	1.10
12 and 13	0.75	5 and 6	1.11
11 and 12	0.93	5 and 7	2.19
10 and 11	0.97	5 and 8	3.20

The thresholds given by Cohen are at best a guide and not viewed as hard and fast ways of interpreting the magnitude of *d*. The values for what might be considered good thresholds for *d* will vary from one subject discipline to another and are very much context dependent. Despite this, others have proffered a more granular interpretation for *d* such as

d = 0 indicates the absence of an effect

and for statistically significant effects,

0 < *d* < 0.1 indicates a trivial effect,

0.1 < *d* < 0.2 indicates a small effect,

0.2 < *d* < 0.5 indicates a moderate effect,

0.5 < *d* < 0.8 indicates a medium size effect,

0.8 < *d* < 1.3 indicates a large effect,

1.3 < d < 2.0 indicates a very large effect

d > 2.0 huge!!

Of course the same caveats apply; specifically interpretation of Cohen's *d* is context dependent and the above thresholds are simply meant to be a guide while acknowledging they do not hold in all circumstances.

IV. Relationship with and Confidence Intervals

Often in the case of two independent groups, the independent samples t-test is used to test $H_0\, \mu_1 = \mu_2$. The t-test statistic, $t_{independent}$, is functionally related to *d* by

$$d = t_{independent} \frac{(n_1 + n_2)}{\sqrt{n_1 n_2}\ \sqrt{(n_1 + n_2 - 2)}}$$

Hence, *d* can be derived directly from *t* and vice versa [3].

The statistics *d* is just that; a statistic. Hence, *d* is subject to sampling error and an approximate formula for the standard error of *d* is given by

Cohen's *d*

$$s.e.(d) = \sqrt{\left(\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)}\right)}$$

and hence an *approximate* 95% CI for δ would be

$$d \ \pm 1.96\, s.e.(d)$$

Note that this interval is an *approximate* interval. Precise intervals (under a precise assumption of normality) can be calculated using software which uses the non-central t-distribution (and closed formulae do not exist for these procedures).

A 95% confidence interval for δ, which excludes 0, indicates a value for *d* significantly different from zero. The approximate nature of the above formula means that the result of a t-test (significant or not) might not always perfectly align with the conclusion drawn from the confidence interval for *d* when using the approximation. Other calculators which use the non-central t-distribution would produce exact confidence intervals (under an assumption of perfect normality) which would not give logical inconsistencies between the results of a t-test and confidence intervals for .

It turns out that the formula

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

provides a biased estimate for δ. For instance suppose two independent samples of sizes n_1 and n_2 are taken from two Normal distributions (e.g. Distribution A with mean μ_A and Distribution B with mean μ_B) and with a common variance σ^2, and *d* is calculated. Conceptually, this process of sampling using sample sizes n_1 and n_2 each time can be repeated indefinitely and the sampling distribution for *d* obtained under these idealized conditions. The average value of *d* under this procedure would not perfectly reproduce the value of δ; if it did then *d* would be an unbiased estimator of δ; however it does not and *d* is a biased estimator of δ.

For an unbiased estimator for δ, Hedges and Olkin [4] have proposed the statistic

$$d_{unbiased} = d\left[1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right]$$

For large sample sizes, the numeric value of *d* and $d_{unbiased}$ will be close to one another, but there could be substantial differences if sample sizes are small. $d_{unbiased}$ is also known as Hedges' *d*.

V. Summary

This note has considered one way of quantifying effect size for two independent samples using a standardised effect. It should be noted that this is only one way of quantifying effect size and many other indices exist for other situations (e.g. correlation coefficients, odds ratios and so on). The development and understanding of *d* has been predicated on an assumption of normality. It turns out that many of the properties discussed will not hold if the samples are severely non-normal.

When the results of a t-test are reported, and when normality has not been grossly violated, then *d* should be routinely reported preferably with its supporting 95% confidence derived using exact methods if possible. It should be specified whether Hedges' or Cohen's *d* is being reported.

Acknowledgments

The authors gratefully acknowledge the support of the Learning and Teaching Initiative in the Faculty of Health and Applied Sciences, University of the West of England, Bristol in supporting the wider Qualitative and Quantitative Methods teaching programme.

Series Bibliography

White P, Redford PC, and Macdonald J (2019) A primer on validity and design terminology in comparative designs, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the chi-squared test of association (2 by 2), *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the independent samples t-test and the Welch test, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –6

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the paired samples t-test, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –5

White P, Redford PC, and Macdonald J (2019) A primer on statistical hypotheses and statistical errors, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) A reverse look at p-values, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –5.

White P, Redford PC, and Macdonald J (2019) That assumption of normality, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –11

White P, Redford PC, and Macdonald J (2019) Cohen's *d* for two independent sample, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –4

References

[1] White P, Redford PC, and Macdonald J (2019) A reverse look at p-values, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –5.

[2] Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences, *Routledge.*

[3] Nakagawa S and Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists, *Biological reviews,* Vol. 82, p. 591-605.

[4] Hedges L and Olkin I (1985) Statistical methods for Meta Analyses, *Academic Press*., New York, NY.