CHAPTER 4

Standardized Differences

If one would like to compare groups with respect to some continuous characteristics (e.g., Academic achievement, IQ, cholesterol level, etc.), the obvious method is to compare the arithmetic means of the groups with one another. Consider the following example:

Example 4.1:

Two random samples of size 100 and 200, drawn from populations A and B, have mean IQs of 110 and 107 respectively and standard deviations 10 and 12 respectively. The usual *z*-test (*t*-test for large samples) produces:

$$z = \frac{|110 - 107|}{\sqrt{\frac{10^2}{100} + \frac{12^2}{200}}} = 2{,}29\,(\,p < 0{,}05\,).$$

This means that the null hypothesis $H_0 : \mu_A - \mu_B = 0$ will reject at a 5% significance level, where μ_A and μ_B are the population means. Therefore, if we say that μ_A and μ_B differ, then the probability that this statement is incorrect is less than 0,05. It might be found that a difference in the means of 3 units is significant; the interpretation of this is that in only 5% of the cases where one repeatedly draws samples of the same sizes from populations A and B, the statement (of rejecting H_o) would not be supported. This does not necessarily indicate that the difference is an important difference. Psychologists who have experience with the IQ scale might use their own judgement and say that a difference of 3 is actually too small to indicate any sort of *practical significance.* The difference in the means provides an effective measure to characterize the practical significance or importance of the difference between the populations. □

In another example we find that the mean attitude of a group of first year students concerning certain aspects of their subject is 6,8 (measured on a stanine scale, i.e., a normalized scale with scale values between 1 and 9, a mean of 5 and standard deviation of approximately 2). The value of 6,8 is almost one standard deviation greater than the mean scale value, and many people might be tempted to say that this value is practically significantly different from 5. Another group, on the other hand, might have a mean attitude score of 4,6 and one might feel that this value is very close to 5, meaning that there is not a practically significant difference.

In both of the above examples the scale is "known" in the sense that the standard deviation is known. The difference of 3 units between the groups (in the first example) when compared to the standard deviations of 10 and 12 is thus small when compared to the difference of 1,8 (in the second example) above the standardized mean, because in this case the standard deviation is only 2. It thus seems sensible to divide the differences in the means by a standard deviation such that

$$\delta = \frac{\mu_1 - \mu_2}{\sigma^*} \tag{4.1}$$

is an effect size that is not *scale dependent*. Here the quantities μ_1 and μ_2 are the two population means and $\sigma*$ is any one of four possibilities for the standard deviation. If σ_1 and σ_2 are the two population SDs then the four choices of $\sigma*$ are:

(a) $\sigma* = \sigma$ the common SD of the two populations (i.e., the assumption here is that the SD of both populations is equal to σ);

(b) $\sigma* = \sigma_1$ or σ_2, depending on one's point of view (for example, if population 1 is the control group or the standard treatment group, then $\sigma^* = \sigma_1$);

(c) $\sigma^* = \max(\sigma_1, \sigma_2)$, the maximum of the two SDs;

(d) $\sigma^* = \sqrt{w_1\sigma_1^2 + w_2\sigma_2^2}$, a weighted SD, where w_1 and w_2 are the weights so that $w_1 + w_2 = 1$.

In the following sections we will focus on each of the abovementioned possibilities.

4.1 Cohen's *d*

Cohen (1969, 1977, 1988) assumes *homogeneity of population variances* and uses the common SD σ in the denominator:

$$d \equiv \delta = \frac{\mu_1 - \mu_2}{\sigma} \qquad (4.2)$$

Note:
In the text that follows we will denote *population* effect size indices using Greek letters (like δ), therefore we rather refer to the symbol δ instead of the letter *d* (which Cohen uses).

4.1.1 Estimation of δ:

Whenever random samples are drawn from two populations, the value δ cannot be calculated directly, but must be estimated from the sample.

Example 4.2:
Similar to Example 1 discussed earlier, consider the IQs of students, except that now we have two complete populations. Let $\mu_1 = 111$ and $\mu_2 = 105$ with $\sigma_1 = \sigma_2 = 10$, then

$$\delta = \frac{111 - 105}{10} = 0,6\,.$$

This result shows that the difference between the two population means is 0,6 standard deviation units. Note that if the example used a different variable, for example, the mean diastolic blood pressure, and had means of 75 and 66 for the two groups and common SD of 15, then

$$\delta = \frac{75 - 66}{15} = 0,6.$$

This result can be interpreted in the same way as the previous one (i.e., the difference is 0,6 standard deviation units, now measured in mmHg). □□

□

A natural estimator (Hedges *g*) is given by

$$g \equiv \hat{\delta} = \frac{\overline{x}_1 - \overline{x}_2}{s_p}, \qquad (4.3)$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1}}, \qquad (4.4)$$

Is the common SD and

$\overline{x}_1, \overline{x}_2$: sample means

n_1, n_2: sample sizes

s_1, s_2: sample standard deviations.

Note:

We will indicate that a quantity is an estimator for the population index based on sample information by placing a “hat” on the symbol, e.g., $\hat{\delta}$ is an estimator for δ.

An alternative formula in terms of the Student t-statistic (or z-statistic if n_1 and n_2 are large) is:

$$\hat{\delta} = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}} \tag{4.5}$$

The estimator $\hat{\delta}$ is actually *positively biased* meaning that, on average, $\hat{\delta}$ over estimates δ. To correct this, especially for smaller values of n_1 and n_2, the following modified effect size index can be used (see Hedges and Olkin, 1985:81):

$$\hat{\delta}_a = \left(1 - \frac{3}{4n - 9}\right)\hat{\delta}, \tag{4.6}$$

where $n = n_1 + n_2$.

Example 4.3:

(Kline, 2004a: 104 -105)

Suppose that the sample means of the aptitude scores of men and women are 13,0 and 11,0 with variances of 7,5 and 5,0 respectively. Table 4.1 displays the results of the *t*-tests and effect size indices for different sample sizes $\left(n_1 = n_2\right)$.

For $n_1 = n_2 = 5$, then $\hat{\delta}$ is modified to become $\hat{\delta}_a$ as follows:

$$s_p = \sqrt{\frac{(5-1)7{,}5 + (5-1)5}{5 + 5 - 2}} = \sqrt{\frac{30 + 20}{8}} = \sqrt{6{,}25} = 2{,}5$$

$$\hat{\delta} = \frac{13 - 11}{2{,}5} = 0{,}8$$

$$\hat{\delta}_a = \left(1 - \frac{3}{4\ x\ 10 - 9}\right)\hat{\delta}$$

$$= \left(1 - \frac{3}{31}\right)0{,}8$$

$$= 0{,}903\ \text{x}\ 0{,}8\ =\ 0{,}722.$$

Table 4.1

	Sample sizes		
Statistic	**5**	**15**	**30**
t-test:			
t	1,26	2,19	3,1
df*	8	28	58
p	0,243	0,037	0,003
Standardized difference:			
$\hat{\delta}$	0,80	0,80	0,80
$\hat{\delta}_a$	0,72	0,78	0,79

df* : degrees of freedom $n_1 + n_2 - 2$

From this table it is clear that the t-statistic values become larger (and the p-values become smaller) as the sample size increases. The estimated effect size index $\hat{\delta}$, however, remains the same because it does not rely on either n_1 or n_2. Further, the values of the modified estimator $\hat{\delta}_a$ are all smaller than $\hat{\delta}$, but increase to $\hat{\delta}$ as n_1 and n_2 become larger. □□□

4.1.2 Confidence intervals for δ

If one assumes that the populations are normally distributed with equal variances, then the statistic

$$t_{v,ncp} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}} \tag{4.7}$$

follows a *non-central t-distribution* with $v = n_1 + n_2 - 2$ degrees of freedom (df) and *non-centrality parameter*

$$ncp = \frac{(\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} = \delta\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{4.8}$$

Cumming & Finch (2001) discuss the non-central *t*-distribution at length and they make use of their ESCI-software and, in particular, the NonCentral tNET-program (which can be downloaded from this manual's website) to further clarify how this distribution is used.

One can easily construct a $100(1-\alpha)\%$ confidence interval (CI) for ncp by making use of the appropriate computer software (see, for example, Kline, 2004a: Table 4.6). The program should only require you to input the values *t, df* and α. From this output and equation (4.8), a CI for δ can be constructed. The SAS-program **VI_delta** (available on the website of this manual) makes use of this method. Zou (2007) also provides this program.

More details concerning the methods used to construct CI are made available in Appendix A.

An approximate CI for δ is given by Hedges & Olkin (1985:86):

Lower bound: $\delta_L = \hat{\delta}_a - z_{\alpha/2}\hat{\sigma}_\delta$

Upper bound: $\delta_U = \hat{\delta}_a + z_{\alpha/2}\hat{\sigma}_\delta$, (4.9)

where $z_{\alpha/2}$ is the $100\left(1-\alpha/2\right)$-th percentile of the standard normal distribution and where

$$\hat{\sigma}_{\delta}^{2}=\frac{n_1+n_2}{n_1 n_2}+\frac{\hat{\delta}_a^2}{2\left(n_1+n_2\right)}, \tag{4.10}$$

is the estimated asymptotic variance of $\hat{\delta}_a$. In a simulation study conducted by Hedges & Olkin(1985: 86-88), it would appear that this approximate interval achieves the correct coverage probability when equal sample sizes greater than 10 are used.

Other approximate CI's for δ are given in Wu et.al. (2006). According to various simulation studies conducted by Wu et al., the following method of CI approximation produces the correct coverage probability for sample sizes as small as 5:

Let $a=\sqrt{4+2n_1/n_2+2n_2/n_1}$, then a variance stabilizing transformation of *d* is: $h(d)=\sqrt{2}\ln\left[d/a+\sqrt{d^2/a^2+1}\right]$, so that $\sqrt{n}\left(h(d)-h(\delta)\right)$ follows an approximate standard normal distribution. The $100\left(1-\alpha\right)\%$ CI for $h(\delta)$ is then $\left(L_{h(\delta)},U_{h(\delta)}\right)=h(d)\pm z_{\alpha/2}/\sqrt{n}$.

The inverse transformation is:

$\delta=\dfrac{a\left(e^{\sqrt{2}h(\delta)}-1\right)}{2e^{h(\delta)/\sqrt{2}}}$, which means that the $100\left(1-\alpha\right)\%$ CI for δ is:

$$\left[\frac{a\left(e^{\sqrt{2}L_{h(\delta)}}-1\right)}{2e^{L_{h(\delta)}/\sqrt{2}}},\frac{a\left(e^{\sqrt{2}U_{h(\delta)}}-1\right)}{2e^{U_{h(\delta)}/\sqrt{2}}}\right].$$

Example 4.4:

In Example 4.1, if one assumes that the population variances are equal, then the estimator $\hat{\delta}$ is calculated as follows:

$$s = \sqrt{\frac{(100-1)10^2 + (200-1)12^2}{100+200-2}} = \sqrt{\frac{38556}{298}} = 11.37$$

$$\hat{\delta} = (110-107)/11.37 = 0.26.$$

From (4.5) it follows that $t = \hat{\delta}\sqrt{\frac{n_1 n_2}{n_1+n_2}} = 0,26 \text{ x } \sqrt{\frac{100 \times 200}{100 + 200}} = 2,12.$

Note that the *t*-value is smaller than *z*=2,29 (in Example 4.1) where no assumptions of homogeneity of variances are made. As a result of the large values of n_1 and n_2, $\hat{\delta}_a$ is also 0,26. Let $\alpha = 0,05$ and note that the degrees of freedom are *v* = 100 + 200 – 2 = 298, then the exact *95% CI* for δ calculated using computer software (see Appendix A) is: (0,018; 0,500). This means that the population effect size index can be as small as 0,018 and as large as 0,5 with probability 0,95. In other words, if samples of the same size are randomly and repeatedly drawn from populations A and B then 95% of the CI's calculated from these samples will contain the unknown value δ.

The approximate 95% *CI* for δ can also be calculated. Here $z_{0,025} = 1,96$ and

$$\hat{\sigma}^2_{\delta} = \frac{100+200}{100 \times 200} + \frac{0,26^2}{2(100+200)} = 0,01500 + 0,000113$$

$$= 0,015113$$

$$\delta_L = 0,26 - 1,96\sqrt{0,0151} = 0,26 - 0,241 = 0,019$$

$$\delta_U = 0,26 + 0,241 = 0,501.$$

The interval (0,019; 0,501) is very close to the exact interval calculated above.

□

Note:

As the sample sizes n_1 and n_2 become smaller, the difference between the approximate interval and the exact interval becomes more pronounced. Table 4.2 shows both of these intervals for three sets of sample sizes used in Table 4.1:

Table 4.2

VI	$n_1 = n_2 = 5$	$n_1 = n_2 = 15$	$n_1 = n_2 = 30$
Exact	(-0,523 ; 2,072)	(0,048 ; 1,539)	(0,271 ; 1,324)
Approximate	(-0,559 ; 1,999)	(0,039 ; 1,521)	(0,264 ; 1,316)

The approximate interval clearly improves as n_1 and n_2 become larger. In Example 4.3 with $n_1 = n_2 = 30$**,** Kline (2004a: Table 4.6) states that the 95% *CI* for δ is (0,271 ; 1,324).

The EXCEL spreadsheet EffectSizeCalculator.xls enables one to calculate the estimates $\hat{\delta}$ and $\hat{\delta}_a$, and the approximate *CI's* by only having to specify the means, SDs and group sample sizes. This spreadsheet is available on the website for this manual. Another handy EXCEL spreadsheet, also available on this manual's website, is Effect_Sizes_Spreadsheet.xls which calculates Cohen's *d* for different combinations of *t*, n_1 and n_2. The 90% $(\alpha = 0{,}10)$ and 99% $(\alpha = 0{,}01)$ *CI's* are respectively (0,355; 1,239) and (0,016; 1,489). Note that the interval becomes "narrower" for large values of α, and becomes wider for small values of α.

4.1.3 Counternull effect sizes

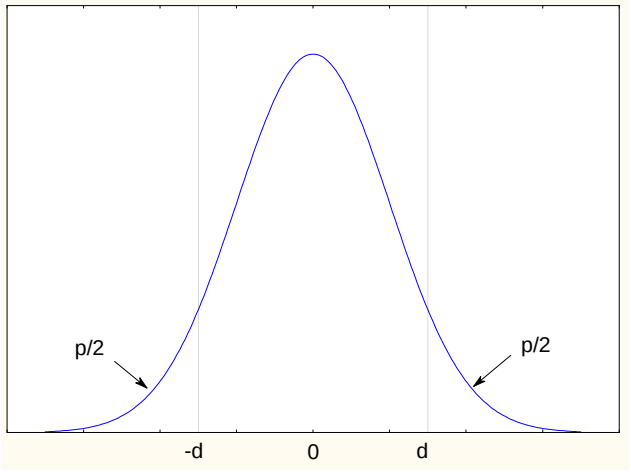
Suppose for the effect size δ given in (4.2) that σ is known and is therefore estimated by

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sigma}.$$

The null hypothesis states that $H_o : \delta = 0$, therefore, under H_o:

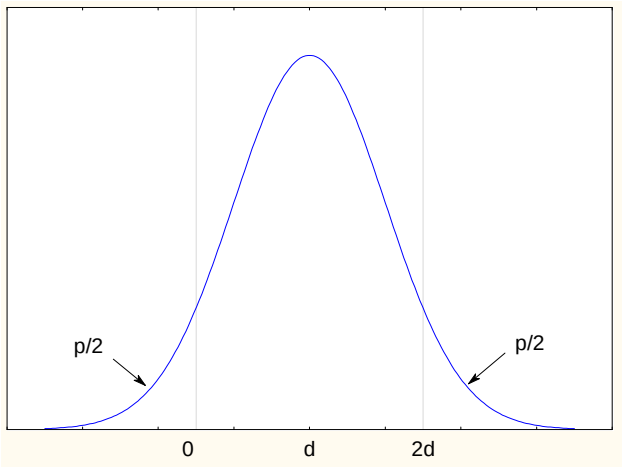
$P\left(\left|\frac{\overline{X}_1 - \overline{X}_2}{\sigma}\right| > d\right) = p$, which is the two-sided p-value of, for example, a z-test (under the assumption of normally distributed populations or large samples from

any population). The following figure illustrates this, using the sampling distribution of $(\bar{X}_1 - \bar{X}_2)/\sigma$:



If the hypothesis $H_o : \delta = d$ is to be tested, then under H_o we find that:

$$P_{\delta=d}\left(\left|\frac{\bar{X}_1 - \bar{X}_2}{\sigma}\right| > d\right) = P_{\delta=d}\left(0 < \frac{\bar{X}_1 - \bar{X}_2}{\sigma} < 2d\right) = p.$$



We see in the in the figure above that the distribution of $\frac{\bar{X}_1 - \bar{X}_2}{\sigma}$ was shifted d units to the right but the shape of the distribution remained the same. This only ever happens if the distribution's standard deviation does not depend on δ . In these cases we say that the effect size d is translation invariant.

This means that the probability that the effect size is smaller than or equal to 0 is equal to the probability that it will be greater than or equal to 2d, if $\delta = d$. This interval (0,2d) is called the **"null-counternull" interval** for δ and 2d is known as the "**counternull" value** of the effect size.

The $(1-\alpha)100\%$ confidence interval (CI) for δ is:

$$d \pm z_{\alpha/2}\sqrt{\frac{n_1+n_2}{n_1 n_2}},$$

which covers δ with probability $(1-\alpha)100\%$. It is clear that there is no direct relationship between the CI and the null-counternull interval, except that they are equivalent if we choose

$$d = z_{\alpha/2}\sqrt{\frac{n_1+n_2}{n_1 n_2}}\ .$$

If $H_o : \delta = 0$ is rejected at a significance level of α, then d is typically large and the interval (0,2d) will be wide. The interval can also be interpreted as a CI with confidence coefficient greater than 1 - α.

Example 4.4 (continued):

The null-counternull interval is (0, 2d) = (0; 0,52), meaning that the probability that the effect size δ is equal to zero is the same as the probability that it is equal to 0,52. Note that this closely corresponds to the 95%CI of (0,018; 0,500) which was obtained earlier.

Remarks:

(a) The interval (0,2d) is obtained from the symmetry of the sampling distribution of d and because d is translation invariant.

(b) If σ is unknown and is estimated from the sample, then we obtain, for example, the estimator $\hat{\delta}$ (in 4.3), which follows a non-central t-distribution

under the assumptions of normality of the population. This estimator for δ is no longer translation invariant and is also positively biased. This means that if $\delta = d$ then the probability that that the effect size is 0 or less is approximately equal to the probability that it is 2d or greater. This approximation improves with larger samples and smaller values of d.

Grissom & Kim (2005: 65-67) discuss the counternull effect sizes in greater depth. We will return to this topic in later chapters.

4.1.4 Interpretation of the counternull effect sizes

Rosenthal et.al (2000: 14) provides the following two examples to illustrate the usefulness of counternull effect sizes:

1. Suppose that $\hat{\delta} = 0,3$ and that the test for $H_0 : \mu_1 - \mu_2 = 0$ produced p = 0,15. The approximate counternull value is then $2\hat{\delta} = 0,6$ indicating that δ could just as easily be as large as 0,6 as it could be as small as 0, notwithstanding the rather large p-value of 0,15. We are thus cautioned against not necessarily interpreting a statistically non-significant result as a zero value of the effect (i.e., $\delta = 0$). This results in the problem described in the bottom-left cell of Table 1 in Chapter 1.
2. In a very large clinical trial a new, expensive medication used to lower body temperature is tested against aspirin.

   A clear statistically significant effect in favour of the new medication is obtained with p = 0,013 (one-sided) whereas $\hat{\delta} = 0,03$. This results in a null-counternull interval of (0,00; 0,06), and consequently produces a 1 - 2 (0,013) = 0,974, i.e., 97,4% confidence interval. Since δ can only be as high as 0,06 with high probability, it is clear that one cannot justify the use of the more expensive medication in practice. This shows that a statistically significant non-zero effect size does not necessarily indicate a

scientifically important effect – this is the situation discussed in the top-right cell of Table 1 of Chapter 1.

The null-counternull interval produces something similar to a confidence interval for the effect sizes and it provides more information when judging the actual effect size values.

4.2 Glass' Δ

If one has an *experimental population* and a *control population,* then Glass (1976) suggests the following effect size index

$$\Delta = \frac{(\mu_E - \mu_K)}{\sigma_K}, \tag{4.11}$$

where μ_E and μ_K are the experimental and control population means respectively and σ_K is the control population's standard deviation.

Example 4.5:

For Example 4.2, suppose that μ_E=111 with a SD of 10 and μ_K =105 with σ_K =15, then

$$\Delta = (111-105)/15 = 0{,}4.$$

This index value is smaller than 0,6 found in Example 4.2 because the difference is divided by a different SD, namely σ_K instead of σ, the common SD. □

4.2.1 Estimation and confidence intervals for Δ.

In the case where a random sample is drawn from a population, one can use the following estimator:

$$\hat{\Delta} = \frac{\bar{x}_E - \bar{x}_K}{s_K}. \tag{4.12}$$

This estimator uses the sample means of both groups and the sample standard deviation of the control group instead of using the population parameters.

According to Kline (2004a: 108) the asymptotic standard error of $\hat{\Delta}$ is given by:

$$\hat{\sigma}_{\Delta} = \sqrt{\frac{n_E + n_K}{n_E n_K} + \frac{\hat{\Delta}^2}{2(n_K - 1)}}, \tag{4.13}$$

so that the approximate $100(1-\alpha)\%$ *CI* for Δ is given by the bounds:

$$\begin{aligned} \Delta_L &= \hat{\Delta} - Z_{\alpha/2}\hat{\sigma}_{\Delta} \\ \Delta_U &= \Delta + Z_{\alpha/2}\hat{\sigma}_{\Delta} \end{aligned} \tag{4.14}$$

Example 4.6:

In Example 4.3, assume that the women are the control group with sample size 20 and the sample size of the men's group is 30.

$$\hat{\sigma}_{\Delta} = \sqrt{\frac{30+20}{30x20} + \frac{0{,}894^2}{2(20-1)}} = \sqrt{0{,}0833 + 0{,}0210}$$

$$= 0{,}323$$

The bounds of a 90% *CI* are thus:

$$\Delta_L = 0{,}894 - 1{,}645x0{,}323 = 0{,}894 - 0{,}531$$

$$= 0{,}363$$

$$\Delta_U = 0{,}894 + 0{,}531 = 1{,}425$$

This means that, with 90% probability, the value of Δ can be as low as 0,363 and as high as 1,425. □

For this index one does not necessarily have to assume homogeneity of variances and the control population is used as a reference point. The mean of an experimental group will sometimes be raised (or lowered) by a treatment or

intervention and, occasionally, the variance is altered at the same time. In these cases Glass's Δ would be the recommended index to use.

4.3 Effect Size Indices for Heterogeneous variances

If one can*not* assume that $\sigma_1 = \sigma_2$, then there are various options for choosing the denominator of the index.

4.3.1 Choice of any population SD

The estimator

$$\hat{\Delta}_1 = (\bar{x}_1 - \bar{x}_2)/s_1, \tag{4.15}$$

is an estimator for

$$\Delta_1 = (\mu_1 - \mu_2)/\sigma_1 \ . \tag{4.16}$$

Similarly

$$\hat{\Delta}_2 = (\bar{x}_1 - \bar{x}_2)/s_2 \tag{4.17}$$

is an estimator for

$$\Delta_2 = (\mu_1 - \mu_2)/\sigma_2 . \tag{4.18}$$

These effect size index values can differ greatly if the values σ_1 and σ_2 differ greatly.

The CI's for Δ_1 and Δ_2 can be determined in a similar manner as was done for Δ by using the following quantity:

$$\hat{\sigma}_{\Delta_1} = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{\hat{\Delta}_1^2}{2(n_1 - 1)}} \ .$$

Example 4.7

(a) Suppose that $\mu_1 - \mu_2 = 5$ and $\sigma_1 = 20$, $\sigma_2 = 5$, then $\Delta_1 = 5/20 = 0{,}25$ and $\Delta_2 = 5/5 = 1{,}0$ which means that Δ_1 is four times smaller than Δ_2.

(b) In Example B of Chapter 3, the J/P preference score would be Δ_1 = (91,08 – 70,07)/28,6 = 0,73.

The value of 0,73 is slightly smaller than the value of Δ_2 = 0,81 obtained if the lecturers were used as the reference point.

□

There is no correct or incorrect choice between Δ_1 and Δ_2, because it only depends on which population's variance you use as your reference point. It thus always best to report both values.

4.3.2 Choice of the largest SD

To be more consistent, one can decide to *always* divide by the largest variance (see also Steyn, 1999, 2000 and Ellis & Steyn, 2003), so that

$$\Delta_m = (\mu_1 - \mu_2)/\sigma_{max} = min(\Delta_1, \Delta_2) \text{ where } \sigma_{max} = max(\sigma_1, \sigma_2). \quad (4.19)$$

The following could then be used as an estimator:

$$\hat{\Delta}_m = (\bar{x}_1 - \bar{x}_2)/s_{max} \quad (4.20)$$

In Example 4.7(a) the value is $\Delta_m = \Delta_1 = 0,25$, because σ_1 is the largest of the two SDs, while in Example 4.7(b) the value is $\Delta_m = \Delta_1 = 0,73$ because in that example σ_1=28,6 is greater than σ_2=25,93.

4.3.3 Weighted SD

When there are noteworthy differences between the population standard deviations, the square root of the weighted mean of the two variances can be used, i.e.,

$$\delta_g = \frac{\mu_1 - \mu_2}{\sqrt{W_1\sigma_1^2 + W_2\sigma_2^2}}, \qquad (4.21)$$

where $W_1 + W_2 = 1$.

The population sizes can be used to determine the weights, i.e., $W_1 = N_1/(N_1 + N_2)$ and $W_2 = 1 - W_1$, where N_1 and N_2 are the population sizes. If the weights are chosen to be equal $(W_1 = W_2 = 1/2)$, then Cohen (1977: 44) provides a special case of (4.21) as the index, namely

$$\delta_c = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}} \quad . \qquad (4.22)$$

Example 4.8:

From Example B in Chapter 3, the results from Table B.1 are, W_1 = 254/(254+28) = 0,9 and W_2 =0,1. The effect size indices for J/P are:

$$\delta_g = \frac{91{,}08 - 70{,}07}{\sqrt{0{,}9 \times 28{,}6^2 + 0{,}1 \times 25{,}93^2}} = \frac{21{,}01}{\sqrt{736{,}16 + 67{,}24}} = 21{,}01 / 28{,}34 = 0{,}74.$$

This produces a slightly larger value than $\Delta_m = \Delta_1$.

If we use Cohen's equal weighted indices, then $\delta_c = \frac{91{,}08 - 70{,}07}{\sqrt{(28{,}6^2 + 25{,}93^2)/2}}$

= 21,01 / 27,3 = 0,77, which is closer to value Δ_2 = 0,8, where the lecturers are used as the reference point. □

When δ_g is estimated from a sample, then we have

$$\hat{\delta}_g = (\bar{x}_1 - \bar{x}_2)/\sqrt{W_1 s_1^2 + W_2 s_2^2}. \qquad (4.23)$$

If population sizes are unknown, or if it is difficult to obtain weights, the following estimator can be used

$$\hat{\delta}_c = (\bar{x}_1 - \bar{x}_2)/\sqrt{\tfrac{1}{2}(s_1^2 + s_2^2)}. \qquad (4.24)$$

Alternatively, one can also use $\hat{\Delta}_m$ found in equation (4.20). Both of these estimators are biased for δ_g. Steyn (1999) found, through the use of simulation studies, that $\hat{\Delta}_m$ underestimates δ_g, but by no more than 0,08 if $\delta_g = 0{,}5$ and $n_2 > 10$, $n_1 \leq 1{,}5n_2$ and $\sigma_1 < 2\sigma_2$. With the same restrictions on n_1, n_2, σ_1 and σ_2, the estimator $\hat{\Delta}_m$ will underestimate δ_g by no more than 0,13 if $\delta_g = 0{,}8$.

Example 4.9:

Suppose that, in Example 4.1, the sample sizes are proportional to the population sizes, so that $W_1 = 100/300 = \frac{1}{3}$ and $W_2 = \frac{2}{3}$. If one can assume $\sigma_1 \neq \sigma_2$, then the following index can be used as an estimator for δ_g:

$$\hat{\delta}_g = (111-105)/\sqrt{\tfrac{1}{3}\times 10^2 + \tfrac{2}{3}\times 12^2}$$
$$= 6/\sqrt{33{,}33+96} = 0{,}527$$

With no knowledge of W_1 and W_2, the indices

$\hat{\delta}_c = 6/\sqrt{\frac{1}{2}\left(10^2+12^2\right)} = 6/\sqrt{122} = 0{,}543$ or $\hat{\Delta}_m = 6/12 = 0{,}5$ serve as conservative estimators.

□

4.4 Effect size indices for dependent groups

Suppose that similar measurements are obtained from the same individuals at different points in time. These groups of measurements are then dependent and are usually obtained when some or other treatment or intervention is applied between measurement times. It is thus important to be able to measure the effect of the intervention.

Example 4.10:

Using the results from Example A in Chapter 3, the following table shows the descriptive statistics of POMS depressions before and after therapy of the *n*=25 heart patients in the experimental group:

Before		After		Difference			
$\bar{x}_1$	s_1	$\bar{x}_2$	s_2	$\bar{x}_D$	s_D	t	p
18,00	12,26	8,08	9,84	9,92	13,38	3,71	0,001

The difference in the mean of the difference between the scores after and before a test, $x_1 - x_2$, namely, $\bar{x}_D$, for the sample data, as in Example 4.10, should be able to reveal the effect of the therapy. However, to obtain an effect size index from this quantity, we need to divide it by some standard deviation.

The population index is thus

$$\delta_D = \mu_D / \sigma^*, \tag{4.25}$$

where μ_D is the mean of the population distribution of the differences and σ^* is defined at the beginning of this chapter. Choosing $\sigma^* = \sigma_1 = \sigma_2$ makes the assumption that the standard deviations are equal at each measurement opportunity. Usually, if a baseline initial measurement is made and then a follow-up measurement is made after the intervention is applied, then this assumption is not very realistic. A more appropriate choice would be to simply make use of σ_1 because it represents a sort of reference variation (i.e., the base line or before test variation) (see Kline, 2004a: 105). This produces the following index

$$\Delta_D = \mu_D / \sigma_1, \tag{4.26}$$

and associated estimator

$$\hat{\Delta}_D = \bar{x}_D / s_1, \tag{4.27}$$

which is known as “Becker’s *g*”.

A second possibility is to select the SD as $\sigma^* = \sigma_D$, i.e., the SD of the population distribution of differences. Usually σ_D is a great deal smaller than either σ_1 or

σ_2, especially if there is a high correlation between the two measurements. This produces the effect size index

$$\delta_D = \mu_D / \sigma_D \tag{4.28}$$

with associated estimator

$$\hat{\delta}_D = \overline{x}_D / s_D \,. \tag{4.29}$$

This estimator is, similarly to the independent group case with $\hat{\delta}$, a biased estimator for δ_D. According to Hedges & Olkin (1985: 79) the estimator

$$\hat{\delta}_{D,a} = \left(1 - \frac{3}{4n-5}\right)\hat{\delta}_D \,, \tag{4.30}$$

is a modified estimator which should be used, particularly if *n* is small.

The problem with these indices is that σ_D (and s_D) no longer express the variation of the scale of the original measurements made at each measurement opportunity, but rather the variation of the differences between the measurement opportunities. Cumming & Finch, (2001: 569 – 570) discuss the following example, where δ_D is compared to the index

$$\delta'_D = \mu_D / \sigma \,, \tag{4.31}$$

(where it is assumed that $\sigma = \sigma_1 = \sigma_2$), and with associated estimator

$$\hat{\delta}'_D = \overline{x}_D / s_P \,, \tag{4.32}$$

where

$$s_P = \sqrt{\tfrac{1}{2}\left(s_1^2 + s_2^2\right)} \tag{4.33}$$

is used as an estimator for σ.

Example 4.11 (Cohen ,1988: 49):

In a verbal ability test, it is known that $\mu = 100$ and $\sigma = 15$. Suppose that the effect of a healthy breakfast is determined by comparing the verbal capabilities of a group of children before given a healthy breakfast and then, later, given a healthy breakfast. The mean difference between the two opportunities is 4,1 in

favour of the test after the healthy breakfast, while the standard deviation of the difference of the scores was 7,7. The estimated effect sizes are:

$\hat{\delta}_D = 4{,}1/7{,}7 = 0{,}53$ while $\delta'_D = 4{,}1/15 = 0{,}27.$

(where we use $\sigma = 15$ because it is known).

In terms of the original scale on which the verbal ability is measurement, the index is approximately half of what it is when we consider it in terms of the variation of the difference measurements.

From Example 4.10 we find that $\hat{\delta}_D = 9{,}92/13{,}38 = 0{,}74$ and

$\hat{\delta}_{D,a} = (1 - \frac{3}{4\times 25 - 5})0{,}74 = 0{,}72.$ Further, we find that $\hat{\delta}_D' = \frac{9{,}92}{\sqrt{(12{,}26^2 + 9{,}84^2)/2}}$ =

9,92 / 11.12 = 0,89. Here the opposite to (a) above is true: the variation of differences between the differences of the before and after measurements is larger that the variation on the original scale, i.e., $\hat{\delta}_D' > \hat{\delta}_D$. This occurrence of larger variation with differences can be attributed to a weak correlation between the before and after measurements (in this case it was 0,28). In this case it appears as though the before and after measurement's SDs differ, and so it is reasonable to say that Δ_D is the proper index to be used. This index can then be estimated by $\hat{\Delta}_D$=9,92 / 12,26 = 0,81. If the original scale is used as a basis, then Δ_D would be the recommended index, rather than δ_D'.

Discussion:

The following is a list of recommendations:

- If there is an indication that the variation is larger (or smaller) for the two measurements (caused by the intervention): use Δ_D and $\hat{\Delta}_D$.
- If homogeneity of variances can be assumed and if you want to use the variation of the of the original measurements as a basis: use δ'_D and $\hat{\delta}_D'$.
- In all other cases, use δ_D and $\hat{\delta}_{D,a}$.

The estimator $\hat{\delta}_{D,a}$ is the *only unbiased estimator***.** If *n* is small, one should keep in mind that the estimators $\hat{\Delta}_D$ and $\hat{\delta}_D{}'$ can overestimate or underestimate the parameters Δ_D and $\delta_D{}'$.

Note:

The effect size indices Δ_D and $\delta_D{}'$ are used and calculated in precisely the same way as Δ and δ for the independent groups, except that they are based on difference measurements. If one wants to calculate $\hat{\delta}_D{}'$ from the *t*-statistic, then formula (4.5) is not the correct one to use, but one should rather make use of (Kline, 2004a: 107):

$$\hat{\delta}_D{}' = t_D \sqrt{\frac{2s_D^2}{n\left(s_1^2 + s_2^2\right)}}, \tag{4.34}$$

where t_D is the dependent groups *t*-statistic .

4.4.1 Confidence intervals for δ_D and $\delta_D^{'}$

As per Johnson et. al. (1995: 513), the asymptotic variance of $\hat{\delta}_{D,a}$ is given by:

$$\hat{\sigma}^2_{\delta_{D,a}} = \frac{1}{n}(1 + ½\hat{\delta}^2_{D,a}), \tag{4.35}$$

so that the $100(1-\alpha)\%$ *CI* for δ_D can be expressed using the following lower and upper bounds:

$$\delta_{D,L} = \hat{\delta}_{D,a} - z_{\alpha/z}\hat{\sigma}_{\delta_{D,a}}$$

$$\delta_{D,U} = \hat{\delta}_{D,a} + z_{\alpha/z}\hat{\sigma}_{\delta_{D,a}} \ . \tag{4.36}$$

For δ'_D Kline (2004a) provides the asymptotic variance of $\hat{\delta}'_D$ as :

$$\hat{\sigma}_{\delta'_D}{}^2 = \frac{2(1-r_{12})}{n} + \frac{\hat{\delta}'^2_D}{2(n-1)} \tag{4.37}$$

where r_{12} is the correlation coefficient between x_1 and x_2. The $100(1-\alpha)\%$ *CI*

for δ'_D can then be expressed using the following lower and upper bounds:

$$\delta'_{D,L} = \hat{\delta}'_D - z_{\alpha/2}\hat{\sigma}_{\delta'_D}$$
$$\hat{\delta}'_{D,U} = \hat{\delta}'_D + z_{\alpha/2}\hat{\sigma}_{\delta'_D} \qquad (4.38)$$

Note that the *CI*'s in (4.36) and (4.38) are only approximations and only hold when *n* is reasonably large (say, $n > 30$).

As in paragraph 4.1.2, there is also an exact *CI* for δ_D which can be constructed under the assumption of a normally distributed population. This interval is constructed by using the fact that the statistic $\frac{\bar{x}_D - \mu_D}{s_D/\sqrt{n}}$ follows a non-central *t*-distribution with *v*=*n*-1 degrees of freedom and non-centrality parameter $\sqrt{n}\delta_D$. See Appendix A for more details concerning the method used. Computer programs (for example, the CIdeltaNET program of ESCI-software or the SAS-program named VI_delta_D) are available on the webpage to perform these calculations. Algina & Keselman (2003) showed, with the help of simulations, that these *CI*'s produce the correct coverage probability for δ_D =0,0 , 0,2 , 0,4 , … 1,6 and ρ =0,0 , 0,2 , 0,4 , … 0,8 (where ρ is the correlation between the two dependent measurements).

From (4.34) it is easy to see that $\hat{\delta}'_D$ also depends on s_1 and s_2, meaning that its distribution can not be expressed in terms of the non-central *t*-distribution, but is actually much more complicated. It is for this reason that it is sufficient to use the previous approximate *CI*'s for δ'_D.

Example 4.12:

Continuing with Example 4.11 we find

$$\hat{\sigma}_{\delta_{D,a}} = 0{,}72,$$

so that $\hat{\sigma}^2_{\delta_{D,a}} = \frac{1}{25}(1 + ½ \times 0{,}72^2) = 0{,}0504$, and

$$\delta_{D,L} = 0,72 - 1,96\sqrt{0,05} = 0,28$$

and

$$\delta_{D,U} = 0,72 + 1,96\sqrt{0,05} = 1,16\,.$$

The exact 95% *CI* is: (0,292 ; 1,180). The approximate *CI* for δ_D does not differ greatly from this exact *CI*, even for *n*= 25, which is not very big.

Further, $\hat{\sigma}_{\delta'_D}{}^2 = \frac{2(1-0,28)}{25} + \frac{0,89^2}{2\times 24} = 0,074,$

so that

$$\delta'_{D,O} = 0,89 - 1,96\times\sqrt{0,074} = 0,359$$

and $$\delta'_{D,B} = 0,89 + 1,96\times\sqrt{0,074} = 1,423.$$

4.5 Counternull values for other effect sizes used to compare two means

As already shown, $2\hat{\delta}$ is only the approximate counternull value for $\hat{\delta}$ because the standard error of $\hat{\delta}$ is a function of δ. This is also true for all the other effect sizes, viz. $\hat{\delta}_a, \hat{\Delta}, \hat{\Delta}_1, \hat{\Delta}_2, \hat{\Delta}_m, \hat{\delta}_g, \hat{\delta}_c, \hat{\Delta}_D, \hat{\delta}_D, \hat{\delta}_{D,a}$ and $\hat{\delta}_D$'.

This approximation reasonable if the sample size is large or if the value δ is small.

□

4.6 Guidelines for effect size indices based on standardized differences

The first question that comes up after an effect size index is calculated is “when can it be considered to be large and when is it considered to be small?” Cohen (1969, 1977, 1988) provides the following guidelines δ (his ‘*d*’):

- $|\delta| = 0,2$: small effect size

- $|\delta| = 0{,}5$: medium effect size
- $|\delta| = 0{,}8$: large effect size.

Notes:

Note that if δ is negative, the sign is ignored when applying these guidelines. The quantity δ is defined as $\delta = (\mu_1 - \mu_2)/\sigma$ and as such the sign of δ becomes negative if $\mu_1 < \mu_2$. Therefore, the *direction* of the effect can be determined by looking at the sign. If one is not interested in the direction, one can always calculate δ by subtracting the smallest mean from the largest one.

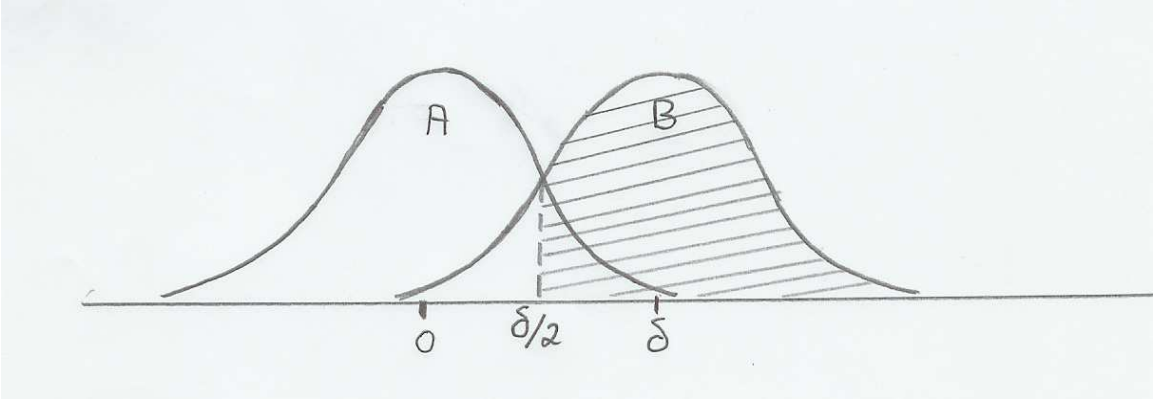
These guidelines are also applicable for the indices Δ, Δ_1, Δ_2, Δ_m, δ_g, δ_D, Δ_D and δ'_D as well as their estimators.

Cohen (1969, 1977, 1988) provide further interpretation of a standardized difference in terms of the overlap between two populations.

Consider the simple case where we have two normally distributed populations A and B with different means, μ_A and μ_B, but with common variance σ^2. The effect size used to compare μ_A and μ_B is

$$\delta = \frac{\mu_B - \mu_A}{\sigma}.$$

Cohen defines the overlapping proportion U_2 as the proportion of population B that exceeds the midpoint between the population means of the two populations (which is equivalent to the proportion of population A less than this midpoint value). The following graph illustrates this concept when, without loss of generality, the population means of populations A and B are taken to be equal to 0 and δ respectively.



In terms of δ-units we compare the distributions N(0;1) and N$(\delta,1)$, and it follows that $U_2 = P\left(Z \leq \frac{1}{2}\delta\right)$.

The probability of a misclassification is

$$P_{MC} = 1 - U_2 = P\left(Z > \frac{1}{2}\delta\right) \qquad (4.39)$$

Cohen attaches the following interpretations to his guidelines (with examples):

4.6.1 Small effect

This occurs in new research areas where measurements are done without any sort of proper experimental controls which can cancel out the effect of background variables. The differences in the means are therefore small relative to the error variation (0,2 or even smaller). Examples include the difference in mean IQs of twins and non-twins, the difference in mean height of girls aged 15 and 16 (12,5mm, where $\sigma = 53{,}3$mm). Here the probability of misclassification (P_{MC}) is 0,46.

4.6.2 Medium effect

This is large enough to be identifiable by inspection. Depending on the context, these differences can be classified as either small or large. Examples include

the difference in the mean IQ of clerical employees and semi-educated employees, the difference between the mean heights f 14 and 18 year old girls (25mm, with $\sigma = 51\,\text{mm}$). P_{MC} is here 0,41.

4.6.2 Large effect

It is an important difference and agrees with what is generally considered to be a definite difference. Examples include the difference in mean IQs of people with a PhD degree and first year students, or the difference between the mean height of 13 and 18 year old girls. Cohen warns that terms “small”, “medium” and “large” are relative, not only with respect to one another, but also to the field of research it is applied. Feinstein(1999:2569) – contemplates this motivation further in paragraph 5.1.2 – and Fleiss(1981) provides the boundary 0,6 as a practical cut-off point for the effect size index δ, which was also confirmed through empirical studies by Burnand et. al. (1990). Here P_{MC} = 0,34 , the misclassification probability.

It is a good idea to consider the following warnings before making use of these guidelines:

4.6.4 Warnings (Kline, 2004a: 132)

1. These guidelines were not obtained using empirical techniques. It is for this reason that Cohen states these warnings.
2. The values 0,2 for “small”, 0,5 for “medium” and 0,8 for “large” should not be applied too strictly. For example, values of 0,49 and 0,51 should not taken as “small” and “medium” by using 0,5 as a cut-off point, but rather both should be classified as “medium”, by using a cut-off in the *region* of 0,5. This is why we call the values 0,2, 0,5 and 0,8 *guidelines*.

3. As Cohen warns, the definition of the sizes does not always hold in all research areas. That which is considered to be a large effect in one research field might not be considered large in another field of research.
4. Cohen's guidelines might be more appropriate in *non-experimental studies,* while in studies which make use of experimental designs, the error variation should be smaller, so that larger guidelines might be more relevant.
5. In established research areas, meta-analysis can be used to systematically distinguish between small and large effect sizes and this information should be used to serve as guidelines. Only in newer fields, where very little is published, would Cohen's guidelines be used.
6. A major advantage of reporting of effect sizes is that they can be compared to results from previous studies and not to rely too much on the arbitrary guidelines proposed by Cohen.

4.7 Practical Significance

The question is now: "When is the effect size index's value substantial, significant or important?" We use these terms to refer to *practical significance* so that it can be distinguished from significance in a statistical sense (i.e., when the null hypothesis is rejected). This question requires proficiency in the specific research context. Kline (2004a: 134) uses an example involving the difference in the mean height of men and women. In this situation a δ-value of 2 can be obtained without being an important difference to, say, a physiologist. However, the value of 2 might be critically important in another situation involving, say, motor vehicle safety where air-bags installed in cars are investigated to see if they could introduce a greater risk for women than for men when compared to older car models.

In earlier phases of research, larger values of the indices can more easily occur than during later stages – which can also influence the determination of practical significance.

One part of the challenge inherent in any new research fields is trying to determine the standards which must be used for practical significance.

The term *clinical significance* is also applicable here. According to Kline (2004a: 135) this term refers to the case where a intervention makes a reasonable difference. For multiple groups a *criterion contrast* is one way to standardize this significance. It represents a known difference in the response variable in which we are interested like, for example, between those patients with the same disease, but exhibit the worst symptoms, versus those patients who exhibit only lesser symptoms. If a criterion contrast is, say, 0,8 before any treatments are applied and the effect size after a treatment is 0,4, then it means that the treatment's effect is half of the distance between the patients with the worst symptoms and the patients exhibiting lesser symptoms. This can be clinically significant even if it is not statistically significant, and vice versa.

According to Kirk (1996), the evaluation of practical significance is a *qualitative decision* because it relies on the researcher's knowledge of the research area without reflecting any of the researcher's personal or social values.

Consequently, the results of each of the examples in this chapter are reported with the interpretation with respect to the guidelines and practical significance.

1. In Example 4.2 the values of δ in both cases of the differences of the mean IQs and diastolic blood pressure was 0,6. According to Cohen's guidelines, this is a medium effect. The psychologist could argue that, since both population means lie above the average of 100 and only differ by 6 scale points, the difference is not important. The effect of lowering the blood pressure might, from

a biokineticist's experience, be enough to indicate that a successful exercise program.

2. In Example 4.3 where $n_1 = n_2 = 5$, the estimated effect size was $\hat{\delta}_a = 0,72$. We could judge that it would be a large effect, because it is almost 0,8 and therefore practically significant. We should rather look at the 95% *CI* which was (-0,523 ; 2,072), since it is safer to say that δ could be as high as 2,07 but also as low as -0,52 (with 95% probability). The difference could even be the opposite sign to what was originally expected, and this difference could have a medium effect!. When $n_1 = n_2 = 30$, the *CI* is (0,271 ; 1,324), which still provides some indication of a small to a large effect.

3. In Example 4.11 the estimated effect sizes are $\hat{\delta}'_D = 0,89$ and $\hat{\delta}_D = 0,74$. In terms of the original scale of the depression scale, this indicates a large effect which could in turn indicate a practically significant decrease after the intervention. We rather look at the effect size where the difference scale was used as the basis, and in this case it tends towards a medium effect. The 95% *CI* for δ'_D : (0,359 ; 1,423) indicates a medium to large effect, if Cohen's guidelines are used, while the *CI* for δ_D, namely (0,292 ; 1,180), makes us wary to even speak of a medium effect.