Journal of Industrial Psychology, 2000, 26(3), 1-3
Tydskrif vir Bedryfsielkunde, 2000, 26(3), 1-3

PRACTICAL SIGNIFICANCE OF THE DIFFERENCE IN MEANS

HS STEYN (JR.)
Statistical Consultation Service
Potchefstroom University for CHE

ABSTRACT

It is shown how the standardised difference (the effect size) between two population means can be used to establish significance when the populations are observed in totality. When dealing with two samples methods are given to determine the practical importance of a statistically significant difference. The usual effect size formula is adapted to deal with cases where populations have different standard deviations.

OPSOMMING

Dit word aangetoon hoe die gestandaardiseerde verskil (die effekgrootte) tussen twee populasiegemiddeldes gebruik kan word om beduidenheid t.o.v. volledig waargenome populasies te bepaal. In die geval van twee steekproewe word metodes gegee om die praktiese belangrikheid van 'n statistiese beduidende verskil vas te stel. Die gewone effekgrootte formule word aangepas ten einde gevalle waar populasies verskillende standaardafwykings het te hanteer.

When comparing the means of two groups in respect of a specified variable the standard statistical method is to test the (null) hypothesis of no difference in means. Recently the use of statistical significance testing in a routine manner has been criticised (Cohen, 1990; Cohen, 1994; Falk & Greenbaum, 1995 and Huysamen, 1991). Also from the editorial side of some periodicals there has been an appeal to authors to place more emphasis on confidence intervals and effect sizes (cf. Bartlett, 1997; Thompson, 1994; Thompson, 1996).

Another topic which has not received much attention is the determination of practical (or psychological) significance. Such significance is particularly important when dealing with complete populations. A relevant example is a study where rapists and armed robbers in a population consisting of three prisons were compared (Verwey, 1986). Both groups i.e. rapists and armed robbers were small enough for psychological testing to be conducted in respect of all the available persons belonging to the groups at the given time. In academic circles it is a common practice to use a complete group of students say first year psychology students and compare their scores to some existing norm (e.g. in an aptitude test). Samples were not drawn in the example above due to the availability of complete groups. In some cases researchers are forced to deal with data coming from a survey which was planned as a probability sample but turned out due to non-response to consist of only a portion (say 20%) of the original sample. Here the respondents can no longer be viewed as a probability sample and therefore they form a subpopulation of the original population which has to be studied as such.

Significance is usually understood in a statistical sense. This means that the null-hypothesis (Ho) that two population means are equal can be rejected in such a way that the probability to conclude that Ho is rejected is controlled to be mostly the significance level (say 5%).

Since statistical significance testing is only appropriate for probability samples (usually assumed to be random) from a population it is not relevant when dealing with complete populations (see Steyn, 1992). To determine significance between e.g. the means of two populations an effect size can be used.

To distinguish between this significance and that of the significance in a statistical sense (i.e. when testing a null-hypothesis) the significance based on effect sizes will be called practical significance (or even psychological- clinical- etc. significance). This is the topic of the next section.

In cases where statistical significance have been established (i.e. where null-hypotheses were rejected or the p-values were small) it might only be due to the very large samples.

Requests for reprints should be addressed to: Steyn H. S., Statistical Consultation Service PU for CHE, Private Bag X6001, Potchefstroom 2520

Example 1
Let the means of the IQ's of two samples of sizes 100 and 200 from populations A and B be 110 and 107 respectively while the standard deviations are 10 and 12. Then the usual two-sided z-test results in:

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

While the difference between the mean IQ's is statistically significant (at the 5%-level) the difference of 3 units on the IQ-scale cannot be considered an important difference. Therefore by using an effect size (which does not depend on the sample sizes) as a measure of significance it is possible to make a judgement concerning the importance of the difference between the means. Methods to do this will be the topic of Section 3.

The effect size in respect of the difference between two population means
In the above example of the class of first year students the researcher can decide whether the mean aptitude of the class deviates enough from the norm to be significant (i.e. of practical value). This is possible when aptitude is measured on a normalised scale such as a stanine scale. If the class mean is 6,8 in comparison with a norm of say 5, the difference is almost 2 scale-points, which is clearly a practical difference. In contrast, a difference of say 0,4, on this scale can be regarded as not significant. In the same way, when dealing with scales which the researcher can interpret (e.g. an IQ-scale, a percentile rank, percentages, etc.) it would be possible to compare two population means by looking at the difference.

Sometimes scales can be somewhat arbitrary and it is difficult to get a "feeling" for the scale. An example is when measuring attitudes using a questionnaire with questions on a 4-point Likert-scale. Will the difference in mean attitude for men (1,9) versus that of women (1,5) be practically significant? This decision would be easier if one knew how the responses varied on this 4-point scale. Looking at the difference of 0,4 in context with a standard deviation (SD, which measures the variability of responses) of say 0,3, the difference would seem more significant than if the SD was 0,8.

Therefore, taking μ_1 and μ_2 to be the means of the two populations and σ and as their (common) standard deviation, a sensible measure of practical significance would be the standardised difference $\Delta = (\mu_1 - \mu_2)/\sigma$, (1)

also called an effect size.

Cohen (1988), denoted this effect size by d and gave some guideline-values as an aid to the interpretation of the extent of practical significance:

a) $|\triangle| = 0,2$: small effect, which means that if this occurs in new research, the experiment or survey ought to be replicated to determine whether there is an effect or whether the result is practically non-significant.
b) $|\triangle| = 0,5$: medium effect, which is detectable and might point towards practical significance. A better planned experiment or survey might result in more significant results.
c) $|\triangle| = 0,8$: large effect, the results are practically significant and therefore of practical importance.

Note that $\triangle$ can be negative when $\mu_1 < \mu_2$, therefore the guidelines are given in terms of the absolute $\triangle$-value. Other measures of effect sizes also exist, of which the best known one is omega-squared ω^2, (see du Toit, 1984; Hays, 1973 and Thomas & Nelson, 1990), which is the proportion of the variance accounted for by population membership. It relates to $\triangle$ as follows:

$$\omega^2 = \triangle^2/(\triangle^2 + (pq)^{-1}), \tag{2}$$

where p is the proportion of elements belonging to the first population and $q = 1 - p$, the proportion of the remainder.

Example 2
Consider the IQ's (as in example 1). Let us assume that the population means were known as $\mu_1 = 111$, and $\mu_2 = 105$, while the SD of both populations was 10. The effect size is $\triangle + \dfrac{111 - 105}{10} = 0,6$ which can be considered as a medium effect. The difference between the mean IQ's of the populations is therefore not quite significant. If the population sizes were 1 000 and 2 000, then

$$p = \frac{1\,000}{1\,000 + 2\,000} = \frac{1}{3}, q = \frac{2}{3}, \text{ and}$$

$$\omega^2 = 0,6^2/\left[0,6^2 + (\tfrac{1}{3}.\tfrac{2}{3})^{-1}\right]$$
$$= 0,36/4,86 = 0,074$$

and which means that only 7,4% of the variance is accounted for by population membership.

A special case of the effect size is when the two means arise from paired observations from one population only. An example would be when pre- and post tests are administered to the members of a population and one wants to establish whether the intervention between the two tests has had an effect. The effect size is again defined by, but here is the standard deviation of the differences between the pre- and post tests.

Example 3
Let the first year class used in a previous example, be pre-tested in respect of their attitude towards their course and then post tested at the completion of the course. Let the means be 2,1 and 3,5 on a 4-point Likert scale, while the SD of the differences between each student's pre- and post-test scores is 1,2. The effect size is:

$$\triangle = \frac{3,5 - 2,1}{1,2} = \frac{1,4}{1,2} = 1,17$$

which indicates a difference of practical significance.

To cover the case where two populations have different SD's σ_1 and σ_2, the effect size in (1) can be adapted to (Steyn, 1999),

$$\triangle_a = (\mu_1 - \mu_2)\sqrt{p\sigma_1^2 + q\sigma_2^2}, \tag{3}$$

which is the same as (1) when $\sigma_1 = \sigma_2 = \sigma$.

Example 4
Let $\sigma_1 = 10$ and $\sigma_2 = 15$ in Example 1, then

$$p\sigma_1^2 + q\sigma_2^2 = \tfrac{1}{3}.10^2 + \tfrac{2}{3}.12^2 = 33,33 + 150$$
$$= 183,33$$
$$\triangle_a = (111 - 105)/\sqrt{183,33} = 6/13,54 = 0,44$$

The value of $\triangle_a$ is somewhat smaller than that of $\triangle$ in Example 1 due to the fact that the second population had a larger SD than the common SD assumed in the former case.

Estimation of effect size from samples
When dealing with samples from populations, a statistical significance test is the standard way to compare population means. In Example 1 this testing lead to the conclusion of a significant difference between the mean IQ levels of the two population groups. The question, however, is whether this difference is of practical importance? The effect size as defined in (1) is a measure of practical significance and has to be estimated from the samples if the populations as a whole are not observed. The usual way to estimate $\triangle$ would be to substitute the sample means $\bar{x}_1$ and $\bar{x}_2$ and an estimate, s, for σ into (1):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} \tag{4}$$

When assuming a common standard deviation for both populations, the estimate for σ is given by:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2}}, \tag{5}$$

where s_1 and s_2 are the SD's for samples of sizes n_1 and n_2, with $n = n_1 + n_2$.

The estimator d, however, is a biased estimator for $\triangle$ when samples are small and tend to overestimate $\triangle$. Hedges and Olkin (1985) suggested that the unbiased estimate

$$\hat{\triangle} = \left[1 + \frac{3}{4n - 9}\right]d, \tag{6}$$

be used instead. This means that $\hat{\triangle}$ adjusts for biasedness in d by a factor of 0,571 when $n = 4$ but only by 0,984 when $n = 50$. In practice, when $n > 50$, this implies that the estimator d in (4) can be used.

Example 5
Consider Example 1.

$$s = \sqrt{\frac{(100 - 1)10^2 + (200 - 1)12^2}{100 + 200 - 2}} = \sqrt{\frac{38556}{298}} = 11,37$$

$$d = \frac{110 - 107}{11,37} = 0,26$$

$$\hat{\triangle} = \left[1 + \frac{3}{4 \times 300 - 9}\right] = \left[1 + \frac{3}{1191}\right]0,26 = 0,999 \times 0,26$$
$$= 0,26.$$

Since n is large, $\hat{\triangle}$ and d are practically equal. This effect size indicates a small effect and although the difference is statistically significant, it is practically non-significant. For the case of paired observations of a sample from one population, the effect size d in (4) can again be used as an estimator for $\triangle$, but with s now the sample SD of differences between pre- and post measurements. As above, this estimator will over-estimate $\triangle$ when the sample size is small. The proper unbiased estimate in this case is (cf. Johnson et. al. 1995):

$$\hat{\triangle} = \left[1 + \frac{3}{4n - 5}\right]$$

Example 6
From Example 3, let a sample of 20 students be used, which results in a mean difference between pre- and post tests of -1,2 and a SD of 1,3. (Note that the mean difference is the same as the difference between the mean of the pre-test and that of the post-test, but that such a relationship does not exist for the SD's).

$$d = \frac{-1,2}{1,3} = -0,92, \hat{\triangle} = \left[1 + \frac{3}{4 \times 20 - 5}\right](-0,92)$$
$$= 0,96 \times (-0,92) = -0,88.$$

The negative sign resulted from the fact that the values of the post-test were subtracted from those of the pre-test. Therefore, there was a practically significant increase of means during the course, because $\triangle = -0,88$ suggests $\triangle > 0,8$.

If the assumption of equal population SD's cannot be made, the effect size is relevant. The usual estimator here, would be:

$$d_a = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{ps_1^2 + qs_2^2}} \tag{8}$$

Again, as for d, the estimate d_a turns out to be biased for small sample sizes in the sense that it over-estimates $\hat{\triangle}_a$ (Steyn, 1999). Another problem with d_a is that the population-proportions p and q must be known, which is not always the case.

Instead, the following sample estimator is suggested (Steyn, 1999):

$$\hat{\triangle}_a = \frac{\bar{x}_1 - \bar{x}_2}{s_{max}}, \tag{9}$$

where s_{max} = maximum of s_1 and s_2.

In a simulation study (Steyn, 1999) it was found that $\hat{\triangle}_a$ is also biased, underestimating $\hat{\triangle}_a = 0,5$ by not more than 0,08 and $\hat{\triangle}_a = 0,8$ by not more than 0,13 when $n_2 > 10$, and n_1 can be 1,5 times the size of n_1. Also σ_1 was assumed to be not more than double the value of σ_2. This under-estimation, however, is a safeguard against making the decision of practical significance too soon.

Example 7
If one is not prepared to make the assumption of equal population SD's in Example 1 (remembering the values $\sigma_1 = 10$ and $\sigma_2 = 15$ in Example 4), then one can estimate :

s_{max} = max (10;12) = 12, therefore

$$\hat{\triangle}_a = \frac{110 - 107}{12} = 0,25.$$

Although the real $\triangle_a$ could have been somewhat higher, the conclusion of a small effect can safely be made.

Discussion and conclusions
The use of the effect size $\triangle$ (or ω^2)is well established in the literature (Cohen, 1988; Du Toit, 1984; Hays, 1973; Hedges & Olkin, 1985 and Thomas & Nelson, 1990). With the exception of Hedges and Olkin (1985) and Steyn (1999), no distinction is made between population and sample cases. In this paper a distinction is made and the application of each of the methods is illustrated by some examples.

The use of practical significance need to be stressed in research. It was shown above how the calculation of an effect size can help to establish whether two completely observed populations differ significantly with respect to some variable. Also, when dealing with two samples, the effect size helps one to decide whether a statistically significant difference between the means is in fact an important difference. This second step is to the author's knowledge, not frequently taken in research.

In practice the assumption which is usually made that is based on a common SD for the two populations, is not realistic. Therefore an adapted effect size $\triangle_a$ is suggested in this paper, together with a simple estimator $\hat{\triangle}_a$, which can be used when dealing with samples.

This paper is only concerned with the difference between two population means. Many other types of effect sizes exist: for differences between proportions, for comparing more than two means, for correlations of several types, etc. (cf. Cohen, 1988 and Steyn, 1999).

REFERENCE

Bartlett, R. (1977). Editorial: The use and abuse of statistics in sport and exercise sciences. Journal of Sport Sciences, *15*, 1-2.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences.* Second Edition. Hillsdale, NJ : Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*:1304-1312.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*(12), 997-1003.

Du Toit, J.M. (1984). *Statistiese metodes; 'n Inleiding vir studente in die Sielkunde, Opvoedkunde en Sosiale Wetenskappe.* Stellenbosch : Kosmo-uitgewery.

Falk, R. & Greenbaum, C.W. (1995). Significance tests die hard. *Theory and Psychology 5*(1), 75-98.

Hays, W.L. (1973). *Statistics for the social sciences.* (2nd edition) New York : Holt, Rinehart & Winston.

Hedges, L.V. & Olkin, I. (1985) *Statistical methods for meta-analysis.* Orlando : Academic Press, Inc.

Huysamen, G.K. (1991). Steekproefgroottes in plaaslik gepubliseerde psigologiese navorsing. *South African Journal of Psychology, 21*(3), 183-189.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995). *Continuous univariate distributions,* Volume 2 (second edition), Wiley series in probability and mathematical statistics, New York : John Wiley.

Steyn, H.S. (1992). Die gebruik van effekgroottes in hipotesetoetsing by volledige populasies. *South African Journal of Psychology, 22*(2); 97.

Steyn, H.S. (jr) (1999). Praktiese beduidendheid: Die gebruik van effekgroottes. *Wetenskaplike bydraes, Reeks B: Natuurwetenskappe* nr. 117, Potchefstroom : Potchefstroomse Universiteit vir CHO.

Thomas, J.R. & Nelson, J.K. (1990). *Research methods in physical activity, human kinetics,* 2nd edition, Champaign : Human Kinetics Books.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 5*(4), 837-847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Verwey, E.T. (1986). *Die verkragter: 'n Vergelykende psigologiese ondersoek.* Ongepubliseerde proefskrif, Potchefstroom : PU vir CHO.