The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine

eJIFCC

Letter to the Editor

The effect size: beyond statistical significance

Ruth Cano-Corres, Javier Sánchez-Álvarez, Xavier Fuentes-Arderiu
Laboratori Clínic, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat, Catalonia, Spain

Corresponding Author:

Ruth Cano-Corres
C/ Breda 13, ático 1º.
08029 Barcelona Spain
Tel.: 0034 686 29 38 34
e-mail: rcano@bellvitgehospital.cat

Key Words

Effect size; statistics; statistical tests; statistical significance.

The branch of statistics known as *inferential statistics* tries to find out information about a population by studying a representative sample of that population. The main tools of inferential statistics are the statistical significance tests.
Every statistical significance test includes two hypotheses; the null hypothesis (H0) and the alternative hypothesis (H1). The null hypothesis assumes that there are no differences between the estimators under study. On the other hand, the alternative hypothesis supposes the existence of differences between them.
These tests are conceived to demonstrate that the alternative hypothesis is the true one; these tests never demonstrate anything about the null hypothesis.
The significance level (α) is the probability of being wrong when the alternative hypothesis is accepted. Since statistical significance tests try to demonstrate that the alternative hypothesis is the true one, the significance level is always defined at the beginning. For most experimental and observational studies it is equal to 0.05.
For a statistical test, when the alternative hypothesis is accepted, the probability of taking the correct decision must be indicated. This probability is called *P* and must be lower than the significance level (α). Nevertheless, assuming that the conclusion is that a significant difference does exist, the *P* value does not inform about the magnitude of this difference.
For example, the statistic *t* (*t* -Student test) and its associated *P* value, only allows to affirm that the difference between two means is significant, but it does not inform about the importance of the difference.
There are other weakness to taking into account only the value of the statistic under study (*t* for the *t*-Student, for example), because this value depends on the mean value and especially on the sample size (and also on the variance in the case of the *t*-Student test).
Some years ago, in other scientific areas, especially in psychology the importance of the magnitude of the difference was described and designed as the *effect size* (1). Currently, the *Publication Manual* of the American Psychological Association, in its sixth edition, provides guidance on reporting effect sizes (2), emphasizing the need of reporting it systematically as a complement of the *P* value. The effect size is considered an essential complement of the statistical significance test when a significant difference is found, because it allows to know the relevance of the difference and discerning between the statistical significance of a test and its practical importance.
This is especially important when the groups under study are formed by a great number of items, because in these cases it is more frequent to find statistical significant differences (See example 1).

The effect size also allows to make comparisons between the statistical significant differences from groups with a very different number of items, and studying groups from different scientific works, as in meta-analysis. This is in fact, the most important application of the effect size.
When a significant difference between the means of two groups is found, there are two main ways to calculate the effect size: by standard scores and by correlation coefficients.

Standard scores method

Frequently, when the term effect size appears in the bibliography, it is referred to the standard score way of calculating it. When two groups are compared by the *t*-Student test and a significant difference is found between them, the obtained *t* value does not inform about the importance of the difference. For this reason, it has been proposed to calculate the effect size. There are different ways to calculate it by standard scores, but the most common one was proposed by Cohen and it is known as the Cohen`s *d*. It is calculated as follows:

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

Where *d* is the effect size, $\bar{x}_1$ and $\bar{x}_2$ are the means of each group, and s is the combined standard deviation from the samples. Having a look to the equation, it is obvious that it is in fact a z score value. It tells about the number of standard deviations that the difference found between the means is equivalent to.

There are some special cases where another way of calculating the effect size and different standard deviations are employed (3).
(a) When the variances of the groups are different, Cohen's *d* overestimates the effect size, so in these cases it should be calculated by the Hedge´s g, which employs the combined standard deviation obtained from the populations instead of the one obtained from the samples (Annex I).
(b) When one of the two groups under comparison is a reference group (not experimental) and the other is the experimental group, the effect size is calculated by the Glass' Δ and the standard deviation employed is the one from the reference group (Annex I).
(c) When the groups under study are not independent groups, but their data proceeded from the same sample "before" and "after" a modification or test. The effect size is calculated by Cohen's *d* and the standard deviation employed is the one obtained from the group "after" (Annex I).

Another special case is the analysis of the variance. This analysis studies the differences between more than two groups. When a significant difference between all the groups is found, for calculating the effect size the groups are taken and compared by pairs applying the Cohen's *d*.
The effect size does not depend on the original data so it is very useful for comparing data from different studies, like in metha-analysis, pooling data from different instruments and many different situations.

There is a problem in how to interpret the effect size and to simplify it, Cohen proposed a transformation of the rational scale into an ordinal scale (4):
$d = 0{,}20 \rightarrow$ small difference
$d = 0{,}50 \rightarrow$ medium difference
$d = 0{,}80 \rightarrow$ big difference
This scale, although arbitrary, is defined frequently found in the bibliography. This transformation scale is ambiguous and to clarify it, a modification is proposed in this letter:
$d < 0{,}20 \rightarrow$ very small difference
$d = [0{,}20 - 0{,}49] \rightarrow$ small difference
$d = [0{,}50 - 0{,}79] \rightarrow$ medium difference
$d > 0{,}80 \rightarrow$ big difference
Hopkins proposed another classification of the effect size, taking into account that the Cohen's *d* can be transformed into a point-biserial correlation coefficient (5) by the following equation:

$$r_{bp} = d \ / \ [d^2 + (1/pq)\ ]^{0,5}$$

Where p and q are the proportions of each group from the general population, so p+q =1.
The classification proposed by Hopkins is:
$d = 0.20 \rightarrow r = 0.10 \rightarrow$ small difference
$d = 0.63 \rightarrow r = 0.30 \rightarrow$ medium difference

$d = 1.15 \rightarrow r = 0.50 \rightarrow$ big difference
This classification is also arbitrary, and in this letter a more practical classification is proposed:
$r < 0{,}10 \rightarrow$ very small difference
$r = [0{,}10 - 0{,}29] \rightarrow$ small difference
$r = [0{,}30 – 0{,}49] \rightarrow$ medium difference
$r > 0{,}50 \rightarrow$ big difference

Correlation coefficients method
When a significant difference is found between two groups applying a *t*-Student test, it is possible to calculate a correlation coefficient from the obtained *t* value to estimate the magnitude of the difference. In this case the correlation coefficient is also a point-biserial correlation coefficient (3), because reflects the correlation between a dichotomous variable (pertaining or not to a group) and a continuous dependent variable. It is similar to classical Pearson´s correlation coefficient (*r*).

The point-biserial correlation coefficient (r_{pb}) is calculated as follows:
$r_{pb} = [t^2 / t^2 + (n_1 + n_2 - 2)]^{0,50}$

where n_1 and n_2 are the sample sizes of the two groups under comparison.
As it has been before described, the effect size calculated by Cohen's *d* can be transformed into an r_{pb}.
The squared of *r*, in this case r_{pb}, is the coefficient of determination (r^2). In this case r^2 expresses the percentage of the variance of the continuous variable that is explained by the dichotomous one (pertaining or not to a group).

The aim of this letter is to emphasize the importance of calculating the effect size when significant differences are found, because it allows to understand better the conclusions of statistical tests. It is really important to know that a significant difference is not the same than important differences. Expressing the conclusions of a study by affirming that there are statistically significant differences is a mistake and a poor scientific strategy.
Also, it must be considered that when an alternative hypothesis can not be rejected, it is usually because of the sample size is small, and that with a long sample size almost any alternative hypothesis can be accepted.
The calculation of effect size is already applied in some other areas of knowledge and its use should be extended to other scientific areas.

Here two examples are exposed of the use of the calculation of the effect size:

Example 1: Means comparison with the t Student test
All the measured values of substance concentration of cholesterol in plasma obtained in a clinical laboratory during one day were selected, and divided in two groups (men and women). The means were compared.
After that, the measured values of substance concentration of cholesterol in plasma obtained during one month in the same clinical laboratory were collected and also divided in two groups (men and women) and the means were compared.
The results are shown in Table 1. There was no statistical difference between the means obtained from both groups for one day of work. However, in the case of data from one month of work, there was a statistical difference between the two groups, but is this difference relevant?

The effect size was calculated to assess the magnitude of this difference, and a $d = 0.124$ was obtained. This means that, guided by Cohen's modified classification, the difference found is very small.

Table 1
Means of measured values of substance concentration cholesterol in plasma compared. [n = sample size; $\bar{x}$ = mean (mmol/L); s = standard deviation (mmol/L); t = statistic from t-Student test; P see the text.]

		n	$\bar{x}$	s	t	P
One day	Men	135	5.0	0.10	-1.21	0.2270
	Women	135	5.1	0.19		
One month	Men	3048	4.8	0.02	-11.72	< 0.0001
	Women	2410	5.2	0.03		

Example 2: Correlation test

Fifty patients were divided in two groups of twenty five patients, one receiving a drug to decrease plasma substance concentrations of cholesterol and another group without treatment. In this study the dichotomous independent variable is receiving or not the drug, and the continuous dependent variable is the substrate concentration of cholesterol in plasma.

The means of substance concentrations of cholesterol in plasma of both groups were compared by a *t*-Student test. The *t* value obtained was $t = 4.78$.

The rbp correlation coefficient was calculated as: $r_{pb} = [t^2/ t^2+(n_1+n_2-2)]^{0,50}$, this calculation gives $r_{bp} = 0,56$, and, consequently, $r^2 = 0,31$.

This means, that the 31 % of the variance of the substance concentration of cholesterol depends on pertaining to one group or other, or what is the same, receiving the drug or not.

References

1. Shinichi N, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev. 2007; 82: 591–605.
2. Publication manual of the American Psychological Association. 6th ed. Washington, DC: American Psychological Association; 2009.
3. Morales Vallejo *P*. El tamaño del efecto (effect size): análisis complementarios al contraste de medias Estadística aplicada a las Ciencias Sociales. Universidad Pontificia Comillas, Madrid, Facultad de Ciencias Humanas y Sociales. <http://www.upcomillas.es/personal/peter/investigacion/Tama%F1oDelEfecto.pdf> (accessed Mars 2012).
4. Cohen J. Stadistical power analisys for the behavioural sciences. 2nd ed. Hillsdate NY: Lawrence Erlbaum Associates; 1988.
5. Hopkins, WG. A new view of statistics. Sportcience. http://www.sportsci.org/resource/stats (accessed Mars 2012).

Annex I: Equations for calculating the effect size

Cohen´s *d*

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

d is the effect size, $\bar{x}_1$ and $\bar{x}_2$ are the means of each group and s is the combined standard deviation.

$s= [(n_1-1)*s^2_1+ (n_2-1)*s^2_2 \ / \ n_1+n_2-2]^{0.5}$

s_1 and s_2 are the standard deviations and n_1 and n_2 are the sample size of the two groups.

$s_1= (\Sigma(x_i - \bar{x}_1)/n_1-1)^{0,5}$ $\quad$ $s_2= (\Sigma(x_j - \bar{x}_2)/n_2-1)^{0,5}$

x_i and x_j represents the individual values from each group.

Hedge´s *g*

$g = \bar{x}_{1-} \bar{x}_2 \ / \ s$

g is the effect size, $\bar{x}_1$ and $\bar{x}_2$ are the means of each group, and s is the standard deviation.

$s= [(n_1-1)*s^2_1+ (n_2-1)*s^2_2 \ / \ n_1+n_2-2]^{0.5}$

s_1 and s_2 are the standard deviations, n_1 and n_2 are the sample size of the two groups.

$s_1= (\Sigma(x_i - \bar{x}_1)/n_1)^{0,5}$ $\quad$ $s_2= (\Sigma(x_j - \bar{x}_2)/n_2)^{0,5}$

x_i and x_j represents the individual values from each group.

Glass' Δ

$\Delta = \bar{x}_e - \bar{x}_r / s_r$

Δ is the effect size, $\bar{x}_e$ is the mean of the experimental group, $\bar{x}_r$ is the mean of the reference group, and s_r is the standard deviation of the reference group.

$s_r = (\Sigma(x_j - \bar{x}_r)/n_r - 1)^{0,5}$

x_j represents each individual value from the reference group, n_r is the sample size of the reference group and $\bar{x}_r$ is the mean value of the reference group.