Understanding the effect size and its measures

Cristiano Ialongo*[1,2]

[1]Laboratory Medicine Department, "Tor Vergata" University Hospital, Rome, Italy
[2]Department of Human Physiology and Pharmacology, University of Rome Sapienza, Rome, Italy

*Corresponding author: cristiano.ialongo@gmail.com

Abstract

The evidence based medicine paradigm demands scientific reliability, but modern research seems to overlook it sometimes. The power analysis represents a way to show the meaningfulness of findings, regardless to the emphasized aspect of statistical significance. Within this statistical framework, the estimation of the effect size represents a means to show the relevance of the evidences produced through research. In this regard, this paper presents and discusses the main procedures to estimate the size of an effect with respect to the specific statistical test used for hypothesis testing. Thus, this work can be seen as an introduction and a guide for the reader interested in the use of effect size estimation for its scientific endeavour.

Key words: biostatistics; statistical data analysis; statistical data interpretation

Received: February 05, 2016 *Accepted: April 26, 2016*

Introduction

In recent times there seems to be a tendency to report ever fewer negative findings in scientific research (1). To see the glass "half full", we might say that our capability to make findings has increased over the years, with every researcher having a high average probability of showing at least something through its own work. However, and unfortunately, it is not so. As long as we are accustomed to think in terms of "significance", we tend to perceive the negative findings (i.e. absence of significance) as something negligible, which is not worth reporting or mentioning at all. Indeed, as we often feel insecure about our means, we tend to hide them fearing of putting at stake our scientific reputation.

Actually, such an extreme interpretation of significance does not correspond to what formerly meant by those who devised the hypothesis testing framework as a tool for supporting the researcher (2). In this paper, we aim to introduce the reader to the concept of estimation of the size of an effect that is the magnitude of a hypothesis which is observed through its experimental investigation. Hereby we will provide means to understand how to use it properly, as well as the reason why it helps in giving appropriate interpretation to the significance of a finding. Furthermore, through a comprehensive set of examples with comments it is possible to better understand the actual application of what is explained in the text.

Technical framework

Stated simply, the "significance" is the magnitude of the evidence which the scientific observation produces regarding to a certain postulated hypothesis. Such a framework basically relies on two assumptions: 1) the observation is intimately affected by some degree of randomness (a heritage of theory of error from which statistics derives), and 2) it is always possible to figure out the way the observation would look like when the phe-

nomenon is completely absent (a derivation of the "goodness of fit" approach of Karl Pearson, the "common ancestor" of modern statisticians). Practically, the evidence can be quantified through the hypothesis testing procedure, which we owe to Ronald Fisher on one hand, and Jerzy Neyman and Egon Pearson (son of Karl) on the other hand (2). The result of hypothesis testing is the probability (or P-value) for which it is likely to consider the observation shaped by chance (the so-called "null-hypothesis") rather than by the phenomenon (the so-called "alternative hypothesis"). The size at which the P-value is considered small enough for excluding the effect of chance corresponds to the statistical significance. Thus, what is the sense of a non-significant result? There are two possibilities:

- there is actually no phenomenon and we observe just the effect of chance, and
- a phenomenon does exist but its small effect is overwhelmed by the effect of chance.

The second possibility poses the question of whether the experimental setting actually makes possible to show a phenomenon when there is really one. In order to achieve this, we need to quantify how large (or small) is the expected effect produced by the phenomenon with respect to the observation through which we aim to detect it. This is the so-called effect size (ES).

P-value limitations

A pitfall in hypothesis testing framework is that it assumes the null hypothesis is always determinable, which means it is exactly equal to a certain quantity (usually zero). Under a practical standpoint, to achieve such a precision with observation would mean to get results which are virtually identical to each other, since any minimal variability would produce a deviation from the null hypothesis prediction. Therefore, with a large number of trials, such a dramatic precision would cause the testing procedure of getting too sensible to trivial differences, making them looking like significant, even when they are not (3). To an intuitive level, let's imagine that our reference value is 1 and we set precision level at 10%. By the precision range of 0.9–1.1 it would result, a 0.1% difference in any actual measure would be shown not significant as $1 + 0.1\% = 1.001 < 1.1$. Contrarily, increasing precision up to 0.01% would give a range of 0.9999–1.0001, thus showing a 0.1% difference as significant since $1.001 > 1.0001$. With respect to experimental designs, we can assume that each observation taken on a case of the study population corresponds to a single trial. Therefore enlarging the sample would increase the probability of getting small P-value even with a very faint effect. As a drawback, especially with biological data, we would risk to misrecognize the natural variability or even to measure error as a significant effect.

Development of ES measures

The issue of achieving meaningful results is measuring, or rather estimating, the size of the effect. A concept which could seem puzzling is that the effect size needs to be dimensionless, as it should deliver the same information regardless of the system used to take the observations. Indeed, changing the system should not influence the size of effect and in turn its measure, as this would disagree with the objectiveness of scientific research.

Said so, it is noteworthy that much of the work regarding ES measuring was pioneered by statistician and psychologist Jacob Cohen, as a part of the paradigm of meta-analysis he developed (4,5). However, Cohen did not create anything which was not already in statistics, but rather gave a means to spread the concept of statistical power and size of an effect among non-statisticians. It should be noticed that some of the ES measures he described were already known to statisticians, as it was regarding to Pearson's product-moment correlation coefficient (formally known as r, eq. 2.1 in Table 1) or Fisher's variance ratio (known as eta-squared, eq. 3.4 in Table 1). Conversely, he derived some other measures directly from certain already known test statistic, as it was with his "d" measure (eq. 1.1 in Table 1) which can be considered stemming strictly from the z-statistic and the Student's t-statistic (6).

A relevant aspect of ES measures is that they can be recognized according to the way they capture the nature of the effect they measure (5):

Table 1. Effect size measures

Measure	Test	Equation	
		Formula	Number
Cohen's d	t-test with equal samples size and variance	$d = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{(s_1^2 + s_2^2)/2}}$	1.1
Hedge's g	t-test on small samples / unequal size	$g = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$	1.2
Glass's Δ	t-test with unequal variances / control group	$\Delta = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{s_{control}^2}}$	1.3
Glass's Δ*	t-test with small control group	$\Delta^* = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{\frac{1}{N-1}\sum(x - \bar{x})^2}}$	1.4
Steiger's ψ (*psi*)	omnibus effect (ANOVA)	$\psi = \sqrt{\frac{\sum(\bar{y} - GM)^2}{(k-1)MSE}}$	1.5
Pearson's r	linear correlation	$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(x - \bar{x})^2}}$	2.1
Spearman's ρ (*rho*)	rank correlation	$\rho = \frac{6\sum(u - v)^2}{N(N^2 - 1)}$	2.2
Cramer's V	nominal association (2 x 2 table)	$V = \sqrt{\chi^2 / N(m-1)}$	2.3
ϕ (*phi*)	Chi-square (2 x 2 table)	$\phi = \sqrt{\chi^2 / N}$	2.4
r^2	simple linear regression	$r^2 = \frac{(\sum(x - \bar{x})(y - \bar{y}))^2}{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}$	3.1
adjusted r^2	multiple linear regression	$r_{adj}^2 = r^2 - \frac{(1 - r^2)p}{N - p - 1}$	3.2
Cohen's f^2	multiple linear regression	$f^2 = \frac{r^2}{1 - r^2}$	3.3a
	n-way ANOVA	$f^2 = \frac{\eta^2}{1 - \eta^2}$	3.3b
η^2 (*eta – squared*)	1-way ANOVA	$\eta^2 = \frac{SS_{factor}}{SS_{total}}$	3.4
partial η^2	n-way ANOVA	$\eta_{partial}^2 = \frac{SS_{factor}}{SS_{factor} + SS_{error}}$	3.5
ω^2 (*omega – squared*)	1-way / n-way ANOVA	$\omega^2 = \frac{SS_{factor} - (k-1)MSE}{SS_{total} + MSE}$	3.6
Odds ratio (OR)	2 x 2 table	$OR = \frac{(x_1 y_1)(x_0 y_0)}{(x_1 y_0)(x_0 y_1)} = \frac{ad}{bc}$	4.1a
	logistic regression	$OR = e^{\beta}$	4.1b

Effect size (ES) measures and their equations are represented with the corresponding statistical test and appropriate condition of application to the sample; the size of the effect (small, medium, large) is reported as a guidance for their appropriate interpretation, while the enumeration (Number) addresses to their discussion within the text.

MSE – mean squared error = SS_{error} / (N – k). Bessel's correction – n / (n-1)[$\chi^2 = \sum \frac{(x_{observed} - x_{expected})^2}{x_{expected}}$].

$\bar{x}$; $\bar{y}$ – average of group / sample. x, y – variable (value). GM – grand mean (ANOVA). s^2 – sample variance. n – sample cases. N – total cases. $\sum_{(...)}$ – summation. χ^2 – chi-square (statistic). u, v – ranks. m – minimum number of rows / columns. p – number of predictors (regression). k – number of groups (ANOVA). SS_{factor} – factor sum of squares (variance between groups). SS_{error} – error sum of square (variance within groups). SS_{total} – total sum of squares (total variance). $x_m y_n$ – cell count (2 x 2 table odds ratio). e – constant (Euler's number). β – exponent term (logistic function).

- through a difference, change or offset between two quantities, similarly to what assessed by the t-statistic
- through an association or variation between two (or more) variates, as is in the correlation coefficient r.

The choice of the appropriate kind of ES measure to use is dictated by the test statistic the hypothesis testing procedure relies on. Indeed, it determines the experimental design adopted and in turn the way the effect of the phenomenon is observed (7). For instance in Table 1, which provides the most relevant ES measures, each of them is given alongside the test statistic framework it relates to. In some situations it is possible to choose between several alternatives, in that almost all ES measures are related each other.

Difference-based family

In the difference-based family the effect is measured as the size of difference between two series of values of the same variable, taken with respect to the same or different samples. As we saw in the previous section, this family relies on the concept formerly expressed by the t-statistic of standardized difference. The prototype of this family was provided by Cohen through the uncorrected standardized mean difference or Cohen's d, whose equation is reported in Table 1 (eq. 1.1; and Example 1).

Cohen's d relies on the pooled standard deviation (the denominator of equation) to standardize the measure of the ES; it assumes the groups having (roughly) equal size and variance. When deviation from this assumption is not negligible (e.g. one group doubles the other) it is possible to account for it using the Bessel's correction (Table 1) for the biased estimation of sample standard deviation. This gives rise to the Hedge's g (eq. 1.2 in Table 1 and Example 1), which is a standardized mean difference corrected by the pooled weighted standard deviation (8).

A particular case of ES estimation involves experiments in which one of the two groups acts as a control. In that we presume that any measure on control is untainted by the effect, we can use its standard deviation to standardize the difference between averages in order to minimize the bias, as it is done in the Glass's delta (Δ) (eq. 1.3 in Table 1 and Example 1) (9). A slight modification of Glass's Δ (termed Glass's Δ*) (eq. 1.4 in Table 1), which embodies Bessel's correction, is useful when the control sample size is small (e.g. less than 20 cases) and this sensibly affects the estimate of control's standard deviation.

It is possible to extend the framework of difference family also to more than two groups, correcting the overall difference (difference of each observation from the average of all observations) by the number of groups considered. Under a formal point of view this corresponds to the omnibus effect of a 1 factor analysis of variance design with fixed effect (1-way ANOVA). Such an ES measure is known as Steiger's psi (ψ) (eq. 1.5 in Table 1 and Example 2) or root mean square standardized effect (RMSSE) (10,11).

As a concluding remark of this section we would mention that it is possible to compute Cohen's d

Example 1

Two groups of subjects, 30 people each, is enrolled to test the serum blood glucose after the administration of an oral hypoglycemic drug. The study aims to assess whether a race-factor might have an effect over the drug. Laboratory analyses show a blood glucose concentration of 7.8 ± 1.3 mmol/L and 7.1 ± 1.1 mmol/L, respectively. According to eq. 1.1 in Table 1, the ES measure is:

$$d = \frac{|7.8 - 7.1|}{\sqrt{\frac{(1.3)^2 + (1.1)^2}{2}}} = 0.581$$

For instance, the power analysis shows that such a cohort ($n_1 + n_2 = 60$) would give 60% of probability to detect an effect of a size as large as 0.581 (that is the statistical power). Therefore we shall question whether the study was potentially inconclusive with respect to its objective.

In another experimental design on the same study groups, the first one is treated with a placebo instead of the hypoglycemic drug. Moreover this group's size is doubled (n = 60) in order to increase the statistical power of the study.

For recalculating the effect size, the Glass's Δ is used instead, as the first group here clearly acts as control. Knowing that its average glucose concentration is 7.9 ± 1.2 mmol/L, according to eq. 1.3 it is:

$$\Delta = \frac{|7.9 - 7.1|}{1.2} = 0.667$$

The ES calculated falls close to the Cohen's d. However when the statistical power is computed based on new sample size (N = 90) and ES estimate, the experimental design shows a power of 83.9% which is fairly adequate. It is noteworthy that the ES calculated through eq. 1.2 gave the following estimate:

$$g = \frac{|7.9 - 7.1|}{\sqrt{\frac{(60-1)\,x\,(1.2)^2 + (30-1)\,x\,(1.1)^2}{60 + 30 - 2}}} = 0.685$$

Example 2

A cohort of 45 subjects is randomized into three groups (k = 3) of 15 subjects each in order to investigate the effect of different hypoglycemic drugs. Particularly, the blood glucose concentration is 8.6 ± 0.2 mmol/L for placebo group, 7.8 ± 0.2 mmol/L for drug 1 group and 6.8 ± 0.2 mmol/L for drug 2 group. In order to calculate the Steiger's ψ, data available through the ANOVA summary and table were obtained using MS Excel's add-in ToolPak (it can be found under Data→Data Anaysis→ANOVA: single factor):

ANOVA SUMMARY				
Groups	Count	Sum	Average	Variance
Drug 1	15	116.3	7.8	0.06
Drug 2	15	102.3	6.8	0.03
Placebo	15	128.3	8.6	0.02

ANOVA TABLE							
Variance component			**DF**	**MS**	**F**	**P**	**F crit**
Between Groups	SS_{factor}	22.5	2	11.24	288	< 0.01	3.2
Within Group	SS_{error}	1.6	42	0.04			
Total	SS_{total}	24.1	44				

ss – sum of squares, DF – degrees of freedom, MS – mean squares.

Notice that the ANOVA summary displays descriptive statistics for the groups in the design, while the ANOVA table gives information regarding the results of ANOVA calculations and statistical analysis. Particularly with respect to power analysis calculations (see later on in Example 4), it shows the value of the components which are the between groups (corresponding to the factor's sum of squares, SS_{factor}), the within groups (corresponding to the error's sum of squares, SS_{error}) and the total variance (that is given by the summation of factor's and error's sum of squares).

Considering that the grand mean (average of the all the data taken as a single group) is 7.7 mmol/L, the formula becomes:

$$\psi = \sqrt{\frac{(7.8-7.7)^2 + (6.8-7.7)^2 + (8.6-7.7)^2}{(3-1)\,\mathrm{x}\,0.04}} = 4.51$$

From the ANOVA table we notice that this design had a very large F-statistic (F = 288) which resulted in a P-value far below 0.01, which agrees with an effect size as large as 4.51.

also for non-Student's family test as the F-test, as well as for non-parametric tests like Chi-square or the Mann-Whitney U-test (12-14).

Association – based family

In the association-based family the effect is measured as the size of variation between two (or more) variables observed in the same or in several different samples. Within this family it is possible to do a further distinction, based on the way the variability is described.

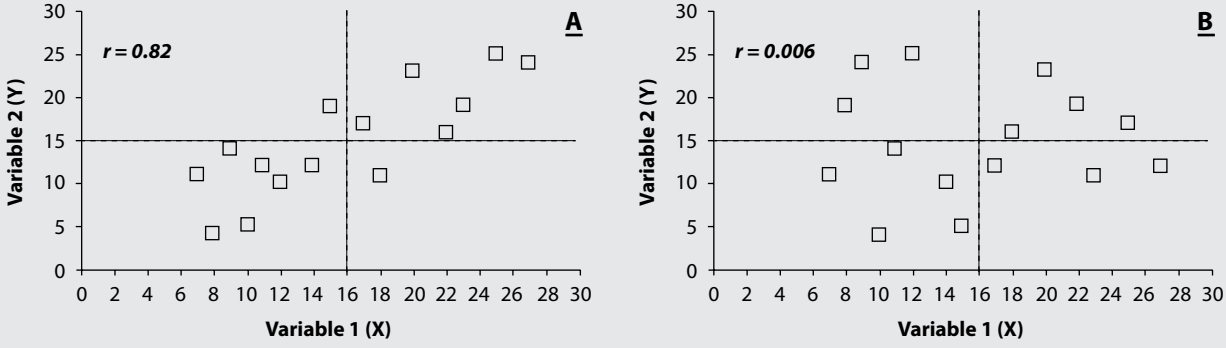
Associated variability: correlation

In the first sub-family, variability is shown as a joint variation of the variables considered. Under a formal point of view it is nothing but the concept which resides in the Pearson's product moment correlation coefficient, which is indeed the progenitor of this group (eq. 2.1 in Table 1 and Example 3). In this regard it should be reminded that by definition the correlation coefficient is nothing but the joint variability of two quantities around a common focal point, divided by the product of the variability of each quantity around its own barycentre or average value (15). Therefore, if the two variables are tightly associated to each other, their joint variability equals the product of their individual variabilities (which is the reason why r can range only between 1 and -1), and the effect can be seen as what forces the two variables to behave so.

When a non-linear association is thought to be present, or the continuous variable were discretized into ranks, it is possible to use the Spear-

Example 3

The easiest way to understand how the ES measured through r works is to look at scattered data:



In both panels the dashed lines represent the average value of X (vertical) and of Y (horizontal). In panel A the correlation coefficient was close to 1 and the data gave the visual impression of lying on a straight line. In panel B, the data of Y were just randomly reordered with respect to X, resulting in a coefficient r very close to zero although the average value of Y was unchanged. Indeed the data appeared to be randomly scattered with no pattern. Therefore the effect which made X and Y to behave similarly in A was vanished by the random sorting of Y, as randomness is by definition the absence of any effect.

man's rho (ρ) instead (eq. 2.2 in Table 1) (6). Alternatively, for those variable naturally nominal, if a two-by-two (2 x 2) table is used, it is possible to calculate the ES through the coefficient phi (ϕ) (eq. 2.4 in Table 1). In case of unequal number of rows and columns, instead of eq. 2.4, the Cramer's V can be used (eq. 2.3 in Table 1), in which a correction factor for the unequal ranks is used, similarly to what is done with the difference family.

Explained variability: general linear models

In the second sub-family the variability is shown through a relationship between two or more variables. Particularly, it is achieved considering a dependence of one on another, assuming that the change in the first is dictated by the other. Under a formal standpoint, the relationship is a function between the two (in simplest case) variables, of which one is dependent (Y) and the other is independent (X). The easiest way to give so is through a linear function of the well-known form $Y = bX + e$, which suits the so-called general linear models (GLM), to which ANOVA, linear regression, and any kind of statistical model which can be considered stemming from that linear function belong. Particularly, in GLM the X is termed the design (one or a set of independent variables), b weight and e the random normal error. In general, such models aim to describe the way Y varies according to the way X changes, using the association between variables to predict how this happens with respect to their own average value (15). In linear regression, the variables of the design are all continuous, so that estimation is made point-to-point between X and Y. Conversely, in ANOVA, the independent variables are discrete/nominal, and thus estimation is rather made level-to-point. Therefore, the ways we assess the effect for these two models slighlty differ, although the conceptual frame is similar.

With respect to linear regression with one independent variable (predictor) and the intercept term (which corresponds to the average value of

Y), the ES measure is given through the coefficient of determination or r^2 (eq. 3.1 in Table 1). Noteworthy, in this simplest form of the model, r^2 is nothing but the squared value of r (6). This should be not surprising because if a relationship is present between the variables, then it can be used to achieve prediction, so that the stronger the relationship the better is the prediction. For mutiple linear regression, where we have more than one predictor, we can use the Cohen's f^2 instead (eq. 3.3a in Table 1) in which the r^2 is corrected by the amount of variation that predictors leave unexplained (4). Sometimes the adjusted r^2 (eq. 3.2 in Table 1) is usually presented alongside to r^2 in multiple regression, in which the correction is made for the number of predictors and the cases. It should be noticed that such a quantity is not a measure of effect, but rather it shows how suitable the actual set of predictors is with respect to the model's predictivity.

With respect to ANOVA, the linear model is rather used in order to describe how Y varies when the changes in X are discrete. Thus, the effect can be thought as a change in clustering of Y with respect to the value of X, termed the factor. In order to assess the magnitude of the effect, it is necessary to show how much the clustering explains the variability (where the observations of Y locate at the change of X) with respect to the overall variability observed (the scatter of all the observations of Y). Therefore, we can write the general form of any ES measure of this kind:

$$\mathrm{ES}_{variance} = \frac{Variation_{explained}}{Variation_{total}}$$

Recalling the law of variance decomposition, for a 1-way ANOVA the quantity above can be achieved through the eta-squared (η^2), in which the variation between clusters or groups accounts for the variability explained by the factor within the design (eq. 3.4 in Table 1 and Example 4) (4,6). The careful reader will recognize at this point the analogies between r^2 and η^2 with no need for any further explanation.

It must be emphasized that η^2 tends to inflate the explained variability giving quite larger ES estimates than it should be (16). Moreover, in models with more than one factor it tends to underestimate ES as the number of factors increases (17). Thus, for designs with more than one factor it is advisable to use the partial-η^2 instead (eq. 3.5), remarking that the equation given herein is just a general form and the precise form of its terms depends on the design (18). Noteworthy, η^2 and partial-η^2 coincide in case of 1-way ANOVA (19,20). A most regarded ES for ANOVA, which is advisable

Example 4

Recalling the ANOVA table seen in Example 2, we can compute η^2 accordingly:

$$\eta^2 = \frac{22.5}{24.1} = 0.934$$

Thereafter for ω^2 we got instead:

$$\omega^2 = \frac{22.5 - ((3-1)\ \mathrm{x}\ 0.04)}{24.1 + 0.04} = 0.929$$

If we recall the value we got previously for ψ (4.51) we notice a considerable difference between these two. Actually, ψ can be influenced by a single large deviating average within the groups, therefore omnibus effect should be regarded as merely indicative of the phenomenon under investigation. Noteworthy, it should be possible to assess the contrast ES (e.g. largest average *vs* others) properly rearranging the Hedge's g.

to use in place of any other ES measure in that it is virtually unbiased, is the omega-squared (ω^2) (eq. 3.6 in Table 1 and Example 4) (16,18,21). Lastly, it should be noticed that Cohen's f^2 can also suit n-way ANOVA (eq. 3.3b) (4). It should be emphasized that in general it holds $\eta^2 > \text{partial-}\eta^2 > \omega^2$.

Odds ratio

The odds ratio (OR) can be regarded as a peculiar kind of ES measure because is suits both 2 x 2 contingency tables as well as non-linear regression models like logistic regression. In general, OR can be tought as a special kind of association family ES for dicothomous (binary) variables. In plain words, the OR represents the likelihood that an event occurs due to a certain factor against the probability that it arises just by chance (that is when the factor is absent). If there is an association then the effect changes the rate of outcomes between groups. For 2 x 2 tables (like Table 2) the OR can be easily calculated using the cross product of cells frequency (eq. 4.1a in Table 1 and Example 5A) (22).

Table 2. 2 x 2 nominal table for odds ratio calculation

Factor (X)	Outcome (Y)	
	1	**0**
1	x_1y_1 ($P_{present}$) or a	x_1y_0 ($1 - P_{present}$) or b
0	x_0y_1 (P_{absent}) or c	x_0y_0 ($1 - P_{absent}$) or d

1 – presence; 0 – absence. The terms presence and absence refer to the factor as well as to the outcome.
a,b,c,d – common coding of cell frequencies used for the cross product calculation.

However, OR can be also estimated by means of logistic regression, which can be considered similar to a linear model in which the dependent variable (termed the outcome in this model) is binary. Indeed, a logistic function is used instead of a linear model in that outcome abruptly changes between two separate statuses (present/absent), so that prediction has to be modelled level-to-level (23). In such a model, finding the weight of the design (that is b in the GLM) is tricky, but using a logarithmic transformation, it is still possible to esti-

Example 5A

Getting OR from 2 x 2 tables is trivial and can be easily achieved by hand calculation as it is possible by the table below:

Factor	Outcome	
	present	**absent**
present	44	23
absent	19	31

Therefore using eq. 4.1a in Table 1 it can be calculated:

$$\mathrm{OR} = \frac{(44)\,\mathrm{x}\,(31)}{(19)\,\mathrm{x}\,(23)} = 3.12$$

It is noteworthy that in this case the Cramer's V gave also an intermediate ES (0.275). Nonetheless they represent quite distant concepts in that Cramer's V is aimed to show wheter variability within the cross-tab frame is due to the factor, while OR shows how factor changes the rate of outcomes in a non-additive way.

mate it through a linear function. It is possible to show that b (usually regarded as beta in this framework) is the exponent of a base (the Euler's number or *e*) which gives the OR (23). Noteworthy, each time there is a unit increase in the predictor, the effect changes according to a multiplicative rather than additive effect, differently than what seen in GLM. A major advantage of logistic regression relies in its flexibility with respect to cross tables, in that it is possible to estimate ES accounting for covariates and factors more than binary (multinomial logistic regression). Moreover, through logistic regression it is also possible to achieve OR for each factor in a multifactor analysis similarly to what is done through GLM.

Confidence interval

Considering that they are estimates, it is possible to give confidence interval (CI) for ES measures as well, with their general rules holding also in this case, so that the narrower the interval the more precise the estimate is (24). However, this one is not a simple task to achieve because ES has non-central distribution as it represents a non-null hypothesis (25). The methods devised to overcome such a pitfall should deserve a broader discussion which would take us far beyond the scope of this paper (10,11,26).

Nonetheless quite easy methods based on estimation of ES variance can be found and have been shown to work properly up to mild sized effects as is for Cohen's d (Example 6) (25). For instance, CI estimation method regarding OR and can be easily achieved by the cells frequency of the 2 x 2 table (Example 5B) (6).

We would remark that although CI of ES might exquisitely concern meta-analysis, actually they represent the most reliable proof of the ES reliability. An aspect which deserves attention in this regard is that CI of ES reminds us that any ES actually measured is just an estimate taken on a sample, and as such it depends on the sample size and variability. It is sometimes easy to misunderstand or forget this, and often the ES obtained through an experiment is erroneously confused with the one hypothesized for the population (27). In this regard, running power analysis after the fact would be helpful. Indeed, supposing the population ES being greater or at least equal with the one actually measured, it would show the adequacy of our

Example 5B

In order to calculate the CI of OR from Example 5A it is necessary to compute the standard error (SE) as follows:

$$SE = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{44}+\frac{1}{31}+\frac{1}{19}+\frac{1}{23}} = 0.39$$

First, it is necessary to transform the OR taking its natural logarithm (ln) for using the normal distribution to get the confidence coefficient (that one which corresponds to the α level). Therefore we got ln (3.12) = 1.14, so that:

$$95\%\ \mathrm{CI}_{\mathrm{OR}} = 1.14 \pm (1.96 \times 0.39) = 0.38\ ;\ 1.90$$

A back transformation through the exponential function makes possible to get this result in its original scale. Hence, if $e^{0.38} = 1.46$ and $e^{1.90} = 6.72$, the 95% CI is 1.46 to 6.72. Noteworthy, if the interval doesn't comprise the value 1 (recalling that ln (1) = 0), the OR and in turn the ES estimate can be considered significant. However, we shall object that the range of CI is quite wide, so that the researcher should pay attention when commenting the point estimation of 3.12.

Example 6

Using the data from Example 1, we can calculate the Cohen's d variance estimate with the following equation:

$$s_d = \sqrt{\frac{n_1 + n_2}{n_1 \times n_2} + \frac{(d)^2}{2 \times (n_1 + n_2)}} = \sqrt{\frac{15 + 15}{15 \times 15} + \frac{(0.581)^2}{2 \times (15 + 15)}} = 0.373$$

Then, we can use this value to compute the 95% CI accordingly:

$$95\%\ CI_d = d \pm 1.96 \times s_d = 0.581 \pm (1.96 \times 0.373) = -0.150\ ;\ 1.312$$

Therefore the estimate falls within the interval ranging -0.150 and 1.312. Interestingly, this shows that the value of the ES estimated through that design was unreliable, because the confidence interval comprises the zero value. Indeed the experimental design aforementioned gave a non-statistically significant result when testing the average difference between the two groups by means of unpaired t-test. This is in accordance with the finding of an underpowered design, which is unable to show a difference if there is one, as well as to give for it any valid measure.

experimental setting with respect to a hypothesis as large as the actual ES (28). Such a proof will surely guide our judgment regarding the proper interpretation of the P-value obtained whereby the same experiment.

Conversion of ES measures

Maybe the most intriguing aspect of ES measures is that it is possible to convert one kind of measure into another (4,25). Indeed, it is obvious that an effect is as such regardless to the way it is assessed, so that changing the shape of the measure is nothing but changing the gear we use for measuring. Although it might look like appealing, this is somehow a useless trick except for meta-analysis. Moreover, it might be even misleading if one forgets what each kind of ES measure represents and is meant for. This kind of "lost-in-translation" is quite common when the conversion is made between ES measures belonging to different families (Example 7).

Contrarily, it seems to be more useful to obtain ES measure from the test statistic whenever the reported results lack of any other means to get ES (4,13,21). However, as in the case of Cohen's d from t-statistic, it is necessary to know the t score as well as the size of each sample (Example 7).

Interpreting the magnitude of ES

Cohen gave some rules of thumb to qualify the magnitude of an effect, giving also thresholds for categorization into small, medium and large size (4). Unfortunately, they were set based on the kind of phenomena which Cohen observed in his field, so that they can be hardly translatable into other domains outside behavioural sciences. Indeed there is no means to give any universal scale, and the values which we take as reference nowadays are just a heritage we owe to the way the study of ES was commenced. Interestingly, Cohen as well as other researchers have tried to interpret the different size ranges using an analogy between ES and Z-score, whereby there was a direct correspondence between the value and the probability to correctly recognize the presence of the investigated phenomenon by its single observation (29). Unfortunately, although alluring, this "percentile-like" interpretation is insidious in that it relies on the assumption that the underlying distribution is normal.

Example 7

The data which were used to generate scatterplot B of Example 3 are compared herein by means of unpaired t-test. Therefore, considering the average values of 16 ± 6 and 15 ± 6, we obtained a t-statistic of 0.453. Hence, the corresponding Cohen's d ES was:

$$d = \frac{t \times (n_1 + n_2)}{\sqrt{(n_1 + n_2 - 2) \times (n_1 \times n_2)}} = \frac{0.453 \times (15 + 15)}{\sqrt{(15 + 15 - 2) \times (15 \times 15)}} = 0.205$$

It should be noticed that panel B of Example 3 reported a correlation close to 0, that is no effect as we stated previously. By the same groups let's calculate now the Cohen's d from r:

$$d = \frac{2 \times r}{\sqrt{1 - r^2}} = \frac{2 \times 0.006}{\sqrt{1 - (0.006)^2}} = 0.012$$

Not surprisingly we obtain a negligible effect. Let's now try again with the data which produced the scatterplot of panel A. While the statistical test gives back the same result, this time the value of d obtained through r changes dramatically:

$$d = \frac{2 \times r}{\sqrt{1 - r^2}} = \frac{2 \times 0.82}{\sqrt{1 - (0.82)^2}} = 2.87$$

The explanation is utterly simple. The unpaired t-test is not affected by the order of observations within each group, so that shuffling the data makes no difference. Conversely, the correlation coefficient relies on data ordering, in that it gives a sense to each pair of observations it is computed with. Thus, computing d through r gives an ES estimate which is nothing but the difference or offset between observations that would have been produced by an effect as large as the one which produced an association as much strong.

An alternative way of figuring out ES magnitude relies on its "contextualization", that is taking its value with respect to any other known available estimation, as well as to the biological or medical context it refers to (30). For instance, in complex disease association studies, where single nucleotide polymorphisms usually have an OR ranging around 1.3, evidence of an OR of 2.5 should not be regarded as moderate (31).

Computing ES

The calculation of ES is part of the power analysis framework, thus the computation of its measures is usually provided embedded within statistical software packages or achieved through stand-alone applications (30,32). For instance, the software package Statistica (StatSoft Inc., Tulsa, USA) provides a comprehensive set of functions for power analysis, which allows computing ES as well as CI for many statistical ES measures (33). Alternatively, the freely available application G*Power (Heinrich Heine Universitat, Dusseldorf, Germany) makes possible to run in stand-alone numerous ES calculations with respect to the different statistical test families (34,35). Finally, it is possible to find online many comprehensive suites of calculators for different ES measures (36-38).

Notwithstanding, it should be noted that any ES measure showed in tables within this paper can be used for calculation with basic (not statistical) functions available through a spreadsheet like MS

Excel (Microsoft Corp., Redmond, USA). In this regard, the Analysis ToolPak embedded in MS Excel allows to get information for both ANOVA and linear regression (39).

Conclusions (Are we ready for the efect size?)

In conclusion the importance of providing an estimate of the effect alongside the P-value should be emhasized, as it is the added value to any research representing a step toward the scientific trueness. For this reason, researchers should be encouraged to show the ES in their work, particularly reporting it any time the P-value is mentioned. It should be also advisable to provide CI along with ES, but we are aware that in many situations it could be rather discouraging as there is still no accessible means for its computation as it is with ES. In this regard, calculators might be of great help, although the researchers should always bear in mind formulae to recall what each ES is suited for and what information it actually provides.

In the introduction of this paper, we were wondering whether negative findings were actually decreasing in scientific research, or rather we were observing a kind of yet unexplained bias. Of course, the dictating paradigm of P-value is leading to forgetting what is scientific evidence and what is the meaning in its statistical assessment. Nonetheless, through the ES we could start teaching ourselves of weighting findings against both chance and magnitude, and that would be a huge help in our appreciation of any scientific achievement. By the way, we might also realize that the bias probably lays in the way we conceive negative and positive things, the reason why we tend to mean the scientific research as nothing but a "positive" endeavour regardless to the size of what it comes across.

Potential conflict of interest

None declared.

References

1. *Fanelli D. Negative results are disappearing from most disciplines and countries. Scientometrics 2011;90:891-904. http://dx.doi.org/10.1007/s11192-011-0494-7.*
2. *Lehmann EL, editor. Fisher, Neyman, and the creation of classical statistics. New York, NY: Springer, 2011.*
3. *Lin M, Lucas HC, Shmueli G. Too big to fail: large samples and the p-value problem. Inform Syst Res 2013;24:906-17. http://dx.doi.org/10.1287/isre.2013.0480.*
4. *Cohen J, editor. Statistical power analysis for the behavioral sciences. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 1988.*
5. *Cohen J. A power primer. Psychological bulletin 1992;112:155-9. http://dx.doi.org/10.1037/0033-2909.112.1.155.*
6. *Armitage P, Berry G, Matthews JNS, eds. Statistical methods in medical research. 4th ed. Osney Mead, Oxford: Blackwell Publishing, 2007.*
7. *Lieber RL. Statistical significance and statistical power in hypothesis testing. J Orthop Res 1990;8:304-9. http://dx.doi.org/10.1002/jor.1100080221.*
8. *Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. J Educ Stat 1981;6:106-28. http://dx.doi.org/10.2307/1164588.*
9. *Zakzanis KK. Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. Arch Clin Neuropsychol 2001;16:653-67. http://dx.doi.org/10.1093/arclin/16.7.653.*
10. *Steiger JH, Fouladi RT. Noncentrality interval estimation and the evaluation of statistical models. In: Harlow LL, Mulaik SA, Steiger JH, eds. What if there were no significance tests? Mahwah, NJ: Lawrence Erlbaum Associates, 1997. p. 221-258.*
11. *Steiger JH. Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. Psychological methods 2004;9:164-82. http://dx.doi.org/10.1037/1082-989X.9.2.164.*
12. *Thalheimer W, Cook S. How to calculate effect sizes from published research articles: A simplified methodology. Available at: http://www.bwgriffin.com/gsu/courses/edur9131/content/Effect_Sizes_pdf5.pdf. Accessed February 1st 2016.*
13. *Dunst CJ, Hamby DW, Trivette CM. Guidelines for calculating effect sizes for practice-based research syntheses. Centerscope 2004;2:1-10.*
14. *Tomczak A, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends Sport Sci 2014;1:19-25.*

15. Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. Crit Care 2003;7:451-9. http://dx.doi.org/10.1186/cc2401.
16. Olejnik S, Algina J. Measures of effect size for comparative studies: Applications, interpretations, and limitations. Contemp Educ Psychol 2000;25:241-86. http://dx.doi.org/10.1006/ceps.2000.1040.
17. Ferguson CJ. An effect size primer: a guide for clinicians and researchers. Prof Psychol Res Pract 2009;40:532-8. http://dx.doi.org/10.1037/a0015808.
18. Olejnik S, Algina J. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. Psychological methods 2003;8:434-47. http://dx.doi.org/10.1037/1082-989X.8.4.434.
19. Pierce CA, Blochk RA, Aguinis H. Cautionary note on reporting eta-squared values from multi factor anova designs. Educ Psychol Meas 2004;64:916-24. http://dx.doi.org/10.1177/0013164404264848.
20. Levine TR, Hullett CR. Eta squared, partial eta squared, and misreporting of effect size in communication research. Hum Commun Res 2002;28:612-25. http://dx.doi.org/10.1111/j.1468-2958.2002.tb00828.x.
21. Keppel G, Wickens TD, eds. Design and analysis: A Researcher's Handbook. 4th ed. Englewood Cliffs, NJ: Prentice Hall, 2004.
22. McHugh ML. The odds ratio: calculation, usage, and interpretation. Biochem Med (Zagreb) 2009;19:120-6. http://dx.doi.org/10.11613/BM.2009.011.
23. Kleinbaum DG, Klein M, eds. Logistic regression: a self-learning text. 2nd ed. New York, NY: Springer-Verlag, 2002.
24. Simundic AM. Confidence Interval. Biochem Med (Zagreb) 2008;18:154-61. http://dx.doi.org/10.11613/BM.2008.015
25. Fritz CO, Morris PE, Richler JJ. Effect size estimates: Current use, calculations, and interpretation. J Exp Psychol Gen 2012;141:2-18. http://dx.doi.org/10.1037/a0024338.
26. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc 2007;82:591-605. http://dx.doi.org/10.1111/j.1469-185X.2007.00027.x.
27. O'Keefe DJ. Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. Commun Methods Meas 2007;1:291-9. http://dx.doi.org/10.1080/19312450701641375.
28. Levine M, Ensom MH. Post hoc power analysis: an idea whose time has passed? Pharmacotherapy 2001;21:405-9. http://dx.doi.org/10.1592/phco.21.5.405.34503.
29. Coe R. It's the effect size, stupid: what effect size is and why it is important. Available at: http://www.cem.org/attachments/ebe/ESguide.pdf. Accessed February 1st 2016.
30. McHugh ML. Power Analysis in Research. Biochem Med (Zagreb) 2008;18:263-74. http://dx.doi.org/10.11613/BM.2008.024.
31. Ioannidis JP, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. Am J Epidemiol 2006;164:609-14. http://dx.doi.org/10.1093/aje/kwj259.
32. McCrum-Gardner E. Sample size and power calculations made simple. Int J Ther Rehabil 2009;17:10-4. http://dx.doi.org/10.12968/ijtr.2010.17.1.45988.
33. Statsoft STATISTICA Help. Available at: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Power/Indices/PowerAnalysis_HIndex. Accessed February 1st 2016.
34. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39:175-91. http://dx.doi.org/10.3758/BF03193146.
35. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009;41:1149-60. http://dx.doi.org/10.3758/BRM.41.4.1149.
36. Lenhard W, Lenhard A. Calculation of effect size 2015. Available at: http://www.psychometrica.de/effect_size.html. Accessed February 1st 2016.
37. Wilson DB. Practical meta-analysis effect size calculator. Available at: http://www.campbellcollaboration.org/resources/effect_size_input.php. Accessed February 1st 2016.
38. Lyons LC, Morris WA. The Meta Analysis Calculator 2016. Available at: http://www.lyonsmorris.com/ma1/. Accessed February 1st 2016.
39. Harmon M. Effect size for single-factor ANOVA 2014. Available at: http://blog.excelmasterseries.com/2014/05/effect-size-for-single-factor-anova.html. Accessed February 1st 2016.