Effect size

Shengping Yang PhD, Gilbert Berdine MD

The concept of effect size is widely used in biomedical research, for example, sample size/power calculation in randomized trials, estimation statistics in data reporting and presentation, and an overall effect size estimation in meta-analysis. Would you provide a brief introduction on effect size?

In statistics, effect size is defined as a number that measures the strength of a relationship between two variables at the population level, or an estimation of such a quantity using samples. Larger effects can be detected with smaller sample sizes. Smaller effects require larger sample sizes to detect at any degree of confidence or significance.

1. Different types of effect sizes

Depending on the nature of a study, the effect size can be measured in the following ways:

1.1 Effect sizes based on differences between means

One of the widely used effect size is the standardized mean difference between two populations. In practice, because population level information is often not known, this quantity can be estimated from samples collected from the populations. Considering data on a continuous variable for two independent groups, and that the difference between the two groups is of interest, then several different effect sizes could be measured.

1.1.1 Cohen's *d*

Cohen's *d* is the difference between two means, divided by the standard deviation of either group when the variances of the two groups are homogeneous, although in practice, the pooled standard deviation is commonly used. Specifically, Cohen's *d* is,

$$\boldsymbol{d} = \frac{\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2}{\boldsymbol{s}},$$

where $\bar{x}_1$ and $\bar{x}_2$ are the two sample means, and $\boldsymbol{s} = \sqrt{\frac{(\boldsymbol{n_1}-1)\boldsymbol{s_1^2}+(\boldsymbol{n_2}-1)\boldsymbol{s_2^2}}{\boldsymbol{n_1}+\boldsymbol{n_2}}}$, where n_1 and n_2 are the sample sizes, and s_1 and s_2 are the sample standard deviations of the two groups, respectively. Cohen's *d* indicates the magnitude of the difference between two means in units of standard deviation. For example, a Cohen's *d* of 1 indicates that the means of two groups are 1 standard deviation apart. A positive value of Cohen's *d* indicates that the treatment group has a greater value of whatever was measured than the control group. Negative values indicate that the treatment group has a lower mean than the control group.

1.1.2 Hedges's *g*

Hedges' *g* is defined as

$$\boldsymbol{g} = \frac{\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2}{\boldsymbol{s}^*},$$

where $\boldsymbol{s}^* = \sqrt{\frac{(\boldsymbol{n_1}-1)\boldsymbol{s_1^2}+(\boldsymbol{n_2}-1)\boldsymbol{s_2^2}}{\boldsymbol{n_1}+\boldsymbol{n_2}-2}}$. The main difference between Cohen's *d* and Hedges's *g* is that the former uses pooled standard deviations while the latter uses pooled weighted standard deviations. Note that, both Cohen's *d* and Hedges's *g* are positively biased on estimating population effect size because the standard deviation estimated from a sample tends to be smaller than the true standard deviation of the population, although such biases are negligible for moderate or large sample sizes. To correct for this bias, Hedges's *g** is proposed,

$$\boldsymbol{g}^* = \boldsymbol{J}(\boldsymbol{n_1} + \boldsymbol{n_2} - 2)\boldsymbol{g},$$

Corresponding author: **Shengping Yang**
Contact Information: Shengping.Yang@pbrc.edu
DOI: 10.12746/swrccc.v9i40.901

where $J(a) = \frac{\Gamma(a/2)}{\sqrt{a/2}\Gamma((a-1)/2)}$, and Γ is the gramma function.

1.1.3 Glass's Δ

Glass's Δ is defined as

$$\Delta = \frac{\overline{x}_1 - \overline{x}_2}{s_2},$$

where s_2 is the standard deviation of the second group, which is often the control group.

There are also other methods to calculate effect size based on differences between means, such as root-mean-square standardized effect and Mahalanobis distance.

1.2 Effect sizes based on "variance explained"

In situations in which data are approximately normally distributed and are expected to be evaluated using an ordinary linear regression model, or an ANOVA, the followings are the commonly used effect size measurements.

1.2.1 Coefficient of determination R^2

The coefficient of determination R^2 is the square of the Pearson correlation between two continuous variables and measures the proportion of the variance for one variable that is shared with another. Note that, R^2 can be calculated for both simple and multiple-variable linear regression models, and a high R^2, i.e., a large effect size, often indicates good model fit. However, there are exceptions - sometimes a high R^2 does not means a good fit, and a non-random model residual pattern could tell a different story.

1.2.2 Cohen's f^2

Cohen's f^2 is defined as

$$f^2 = \frac{R^2}{1 - R^2}.$$

This is an effect size that can be used in both ANOVA and ordinary linear regression.

1.3 Effect sizes for associations among categorical variables

There are many types of categorical variables, we focus on situations that the outcome variable is binary, and the exposure variables are categorical.

1.3.1 Odds ratio

The odds ratio is a measure of association between an exposure and a binary outcome, and represents the odds that an outcome will occur given a particular exposure category, compared to that in another exposure category. Note that odds ratio can also be calculated for a continuous exposure variable, however, we will not discuss that in this article.

1.3.2 Cohen's *h*

Cohen's *h* is defined as

$$h = 2\left(arcsin\sqrt{p_1} - arcsin\sqrt{p_2}\right),$$

where p_1 and p_2 are the proportions of the two samples being compared, and *arcsin* is the arcsine transformation.

1.4 Effect size for time-to-event data

In situations where the outcome of interest is not only whether or not an event has occurred, but also when the event occurred, the commonly used effect size measurement is hazard ratio.

A hazard ratio is the ratio of the hazard rates corresponding to the conditions described by two levels/categories of an exposure variable, when the exposure variable is categorical.

Note that, while many times effect sizes and treatment effects are used interchangeably, for example, the difference between groups is based on a deliberate intervention; other times, it is more appropriate to use effect size, for example, the difference between males and females.

2. Effect size and sample size/power calculation

Statistical power is defined as the probability of finding significant results from a statistical test when an effect actually exists. The factors affecting a power calculation often include sample size, type I error rate (α) and effect size. For example, with a continuous outcome approximately following a normal distribution, the sample size (per group) required for comparing two equal-sized independent groups using a 2-sided two-sample *t* test is,

$$n = 2\left(\frac{z_{1-\alpha/2} + z_{power}}{Effect\ size}\right)^2,$$

where *z* the quantile of a standard normal distribution.

In general, with a pre-specified type I error rate having a small sample size might fail to detect an important effect, while have a large sample size might lead to detect a statistically significant yet clinically insignificant effect. To meaningfully calculate the sample size, a minimal clinically important effect size needs to be specified. This is the smallest difference in outcome between the study groups that is of clinical interest to investigate.

3. Estimation statistics in data reporting and presentation

Traditionally, results from a statistical significance test are often used in data reporting and presentation. However, more and more investigators are suggesting that effect size should be included in data reporting and presentation as well.

The primary goal of statistical testing is to obtain a p value, which is the probability of observing results as least as extreme as the observed, when assuming the null hypothesis is true. If the p value of a statistical test is less than a specified value, usually set at 0.05, then we declare statistical significance. On the other hand, effect size evaluates the magnitude of the difference found independent of statistical significance. Therefore, both the significant test and the effect size are complementary and both essential for understanding the differences in a comparison. In fact, effect size has the advantage of quantifying the difference over the use of tests of statistical significance, which could confound with sample size. Many journals have recommendations for manuscript submissions that "avoid relying solely on statistical hypothesis testing, such as p value, which fail to convey important information about effect size." However, specifying only the effect size such as an odds ratio can be just as misleading as specifying only the p value. For rare events, a large odds ratio can still be a very small absolute benefit. Sometimes, a third measure, such as the number to treat to achieve a single benefit, must be added to clarify the situation.

4. Effect size in meta-analysis

A meta-analysis is a statistical analysis that combines results from multiple studies addressing the same question. One important consideration in performing a meta-analysis is to choose an appropriate effect size to be evaluated. For example, the effect size from different studies should be comparable to one another, and should have good technical properties, so that its sampling distribution is known and confidence intervals can be computed.

5. Effect size interpretation

It has been suggested by Cohen that $d = 0.2$ can be considered a small effect size, 0.5 a medium effect size, and 0.8 a large effect size. This means that, from a statistical consideration, if the difference between the two groups is less than 0.2 standard deviations, then the difference is considered to be small. On the other hand, the interpretation of effect size might be different from a clinical perspective. Depending on the nature of the outcome measured, a 0.2 standard deviation difference might be considered as clinically important for certain outcomes, and a 0.3 standard deviation difference might be considered trivial for other outcomes. There are also differences between clinical and pre-clinical studies. For example, a small but meaningful difference might be important in a clinical study, however, a difference of the same magnitude might not be of interest in a pre-clinical study due to the homogeneity nature of pre-clinical studies.

In summary, effect size is an important quantity in evaluating the strength of the relationship between two variables. Depending on the nature of the outcome variables, there are different types of effect sizes. Choosing the appropriate effect size is not only important in a power/sample size calculation, but also in a meta-analysis. The interpretation of an effect size should take into account both statistical and clinical considerations.

Keywords: effect size, Cohen's d, statistical significance, clinical significance

Article citation: Yang S, Berdine G. Effect size. The Southwest Respiratory and Critical Care Chronicles 2021;9(40):65–68
From: Department of Biostatistics (SY), Pennington Biomedical Research Center Baton Rouge, LA; Department of Internal Medicine (GB), Texas Tech University Health Sciences Center, Lubbock, Texas
Submitted: 7/6/2021
Accepted: 7/9/2021
Conflicts of interest: none

References

1. Borenstein M. Common mistakes in meta-analysis and how to avoid them. Biostat Inc. 2019.
2. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Hillsdale, N.J. Lawrence Erlbaum, 1988.
3. Hedges LV. Distribution theory for Glass' estimator of effect size and related estimators. J Educational Statistics 1981;6(2): 107–128.
4. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Orlando: Academic Press. 1985.
5. Kelley K, Preacher KJ. On effect size. Psychological Methods 2012;17(2):137–152.
6. Turner DP, Houle TT. The importance of statistical power calculations. Headache 2018;58(8):1187–1191.