BORNEO EPIDEMIOLOGY JOURNAL | BEJ

REVIEW ARTICLE Open Access

Use of Effect Size Measures along with *p*-Value in Scientific Publications

Sandheep Sugathan*, Lilli Jacob

Abstract

Background: To describe various measures for estimation of effect size, how it can be calculated and the scenarios in which each measures of effect size can be applied.

Methods: The researchers can display the effect size measures in research articles which evaluate the difference between the means of continuous variables in different groups or the difference in proportions of outcomes in different groups of individuals. When p-value alone is displayed in a research article, without mentioning the effect size, reader may not get the correct pictures regarding the effect or role of independent variable on the outcome variable.

Results: Effect size is a statistical concept that measures the actual difference between the groups or the strength of the relationship between two variables on a numeric scale.

Conclusion: Effect size measures in scientific publications can communicate the actual difference between groups or the estimate of association between the variables, not just if the association or difference is statistically significant. The researchers can make their findings more interpretable, by displaying a suitable measure of effect size. Effect size measure can help the researchers to do meta-analysis by combining the data from multiple research articles.

Keywords: Effect size, *p*-Value, Research Publication, statistical concept.

*Correspondence Email: drsandheep@gmail.com

[1]*Community Medicine, Quest International University, Perak, Malaysia*

Received: 30/11/2021

Accepted: 21/12/2021

Introduction

The researchers can display the effect size measures in research articles which evaluate the difference between the means of continuous variables in different groups or the difference in proportions of outcomes in different groups of individuals. Effect size is a statistical concept that measures the actual difference between the groups or the strength of the relationship between two variables on a numeric scale.

> The effect size measures represent the actual magnitude of the relationship between the independent variable and dependent variables (Sullivan, G.M. *et al.*, 2012).
>
> "You should describe the results in terms of measures of magnitude –not just, does a treatment affect people, but how much does it affect them" - Gene V. Glass (Kline, R.B., 2004).
>
> "The primary product of a research inquiry is one or more measures of effect size, not P values." - Jacob Cohen (Cohen, J., 1990).

When p-value alone is displayed in a research article, without mentioning the effect size, reader may not get the correct pictures regarding the effect or role of independent variable on the outcome variable. In many of the research publications, effect size measures are not represented as the p-value is presented. Purpose of this review article is to describe various measures for estimation of effect size, how it can be calculated and the scenarios in which each measures of effect size can be applied.

Methods

Absolute measures of effect size

The effect size can be presented as the raw difference between group means or absolute effect size, as well as the standardized measures of effect, which are calculated to transform the absolute effect size to an easily understood scale. Absolute effect size measure is useful when the variables under study have intrinsic meaning (for example, if we are estimating the number of hours of sleep or the difference between mean hours of sleep among individuals from different groups).

Standardized measures of effect size

Standardized measures of effect size are useful when the measurements have no intrinsic meaning, such as the raw scores on a Likert scale; or when the studies used different scales so no direct comparison is possible; or when effect size is examined in the context of variability in the population under study.

The effect size measures differ based on the type of comparison.

For example,

- When two means are compared, the effect size displays the actual difference in means divided by the standard deviation. That is known as standardized mean difference.
- When two means are compared, the effect size displays the actual difference in means along with 95% confidence interval of the difference in means.

- When two proportions are compared, the effect size displays the actual difference in proportions along with the 95% confidence interval of the difference in proportions.
- When studying the linear relationship between two continuous variables, the effect size is represented by the correlation coefficient (r).
- Comparing the odds of exposure or outcome in different groups – using odds ratio and 95% confidence interval of odds ratio.
- Incidence of disease among exposed group and incidence of disease in the non-exposed group can be displayed.
- Relative risk or risk difference can be calculated by dividing the incidence of disease among exposed group by the incidence of disease in the non-exposed group.

In statistics analysis, the effect size is usually measured in three ways: standardized mean difference, odds ratio or correlation coefficient (Complete Dissertation by Statistics Solution).

Advantages of Effect Size Measures

Effect size can tell you:

- How large the difference is between groups.
- The absolute effect (the difference between the average outcomes of two groups).
- The standardized effect size for an outcome.

An example of absolute effect size could be: patients taking a drug for depression might see a mean improvement on a depression test (like Beck Depression Inventory) by 10 points. Standardized effect sizes are estimated in a way that some scores are standardized using z-scores; it makes the result more interpretable and comparable to those research articles using another measure for data collection.

How to Present Effect Size in Publications

According to the 7th edition of Publication manual of American Psychological Association, it is highly recommended to include measures of effect size along with a confidence interval for each effect size in the Results section for the readers to appreciate the magnitude or importance of a study's findings. Confidence interval for each effect size is used to display the precision of the effect size measures (Publication Manual of the American Psychological Association).

Effect sizes may be expressed in the original units.

- Actual mean number of questions answered correctly along with 95% confidence intervals of mean.
- Difference in mean values between groups along with 95% confidence intervals of difference in mean.
- Actual increase in the value of dependent variable for each unit increase in the value of independent variable – whenever presenting a logistic regression result.

It is valuable to also report an effect size in some standardized or units-free or scale-free unit (e.g., Standardized mean difference, Cohen's d, Hedge's g, Glass's Δ, Cohen's f, Somer's d, f^2, R^2, Phi and Cramer's V).

As commonly used effect size measures with the scenarios, it continues to go through the commonly used effect size measures.

a. Standardized Means Difference:

Standardized mean difference is used as a measure of effect size, when a research study is based on the population mean and standard deviation.

The effect size of the difference in the subgroups in the population can be known by dividing the mean differences in the sub-populations or subgroups by their standard deviation in the whole population.

$$\theta = \frac{\mu_1 - \mu_2}{\sigma}$$

b. Cohen's d:

Cohen's d is known as the difference of two population means and it is divided by the standard deviation from the data. It can be applied for independent samples t test, one sample t-test and paired t-test (Effect Size in Statistics).

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

Numerator is the difference between means of the 2 groups. 's' is the pooled standard deviation which can be calculated using the following formula

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

n_1 = number of samples in group 1 and n_2 = number of samples in group 2.

c. Pearson Correlation Coefficient:

Pearson correlation coefficient is a measure on effect size which shows the strength of linear relationship between two continuous variables and the direction of linear relationship.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Pearson correlation coefficient can also calculated using the following formula

$$r = \frac{\text{Covariance between x and y}}{(\text{Standard deviation of x}) * (\text{standard deviation of y})}$$

d. Point-Biserial Correlation Coefficient:

It is used to measure the strength and direction of the association that exists between one continuous variable and one dichotomous or binary variable. It can be used for independent samples t-test since the independent variable is dichotomous.

e. Hedge's g Method:

This method is the modified method of Cohen's d method

$$\text{Hedges' } g = \frac{M_1 - M_2}{SD}$$

Numerator: difference between means of 2 groups. SD: pooled & weighted standard deviation

Pooled and weighted standard deviation for 2 groups can be calculated using the following formula (Effect Size in Statistics).

$$SD^*_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

Pooled and weighted standard deviation if there are 3 or more groups (consider that k groups are present)

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2 + \ldots + (n_k - 1)s_k{}^2}{n_1 + n_2 + \ldots + n_k - k}}$$

If sample sizes are equal in multiple groups, pooled and weighted standard deviation can be calculated as

$$s_{\text{pooled}} = \sqrt{\frac{s_1{}^2 + s_2{}^2 + \ldots + s_k{}^2}{k}}$$

Interpretation of Cohen's d and Hedge's g

A 'Hedge's g' of 1 indicates the two groups differ by 1 standard deviation, a g of 2 indicates they differ by 2 standard deviations, and so on. Standard deviations are equivalent to z-scores (1 standard deviation = 1 z-score). Hedges' g and Cohen's d are similar. Both have an upwards bias in results of up to about 4%. When sample sizes are below 20, when Hedges' g is better measure of effect size compared to Cohen's d. So Hedges' g is therefore sometimes called the corrected effect size. For very small sample sizes (<20) choose Hedges' g over Cohen's d and for sample sizes >20, the results for both statistics are roughly equivalent (Stephanie, G., 2017).

f. Glass's Δ or Glass's Delta (Complete Dissertation by Statistics Solution)

Glass's Delta is a measure for effect size, when standard deviations are significantly different between groups. Glass's delta uses only the control group's standard deviation (SDC).

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

Numerator is the difference between means of the 2 groups; S_2 is the standard deviation in the control group.

g. Cohen’s f statistic

Cohen’s f statistic is used as a measure of effect size in ANOVA and ANCOVA. The statistic gives an estimate of the proportion of variance explained by the categorical variable. Cohen’s f is a ratio between two interim sums of squares: the treatment sum-of-squares and error-sum-of-squares (The proportion of variance explained by the analysis model relative to the proportion of variance not explained by the analysis model). Cohen suggested the following interpretation for f when used in ANOVA / ANCOVA: 0.10 = Small effect size, 0.25 = Medium effect size, 0.40 = Large effect size, When f = 0, that’s an indication that the population means are all equal. As the means get further and further apart, f will grow indefinitely larger.

h. Somers’ Delta (Somers’ D)

Somers’ Delta (Somers’ D) is a measure of agreement between pairs of ordinal variables. Ordinal variables are ordered, like highest to lowest or smallest to greatest (the Likert scale is one of the more popular ordinal scales.)
A measure of agreement tells you something about how two pairs of variables are connected. This connectivity is defined by concordance and discordance. Concordant pairs “match” while discordant pairs don’t “match”.

Effect size measures which can be used for (simple and multiple) linear regression are f^2, R^2 and adjusted R^2 (Stephanie, G., 2018)

> R^2 can be used as the measure of effect size (for entire model) – is the coefficient of determination. It is the proportion of the variation in the dependent variable that is predictable from the variation in the independent variable in simple linear regression.
> R^2 is the proportion of the variation in the dependent variable that is predictable from the variation in the combination of independent variables added to the model in multiple linear regression.
>
> f^2 can be used as the measure of effect size for entire model and for individual predictor variable.

f^2 can be calculated using the following formula

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

$\mathbf{R^2_{inc}}$ is the increase in R^2 value while adding the independent variable to the model as compared to the model which is not having the particular independent variable added. When only one predictor variable is added in the linear regression model, R^2_{inc} will be $R^2 – 0 = R^2$

Effect size measures for Chi square test

There are three ways to measure effect size: **Phi (φ), Cramer’s V (V), and odds ratio (OR).**

a. Phi (φ)

Phi is calculated as $\varphi = \sqrt{(X^2 / n)}$.

X^2 is the Chi square value and n is the sample size.

It's appropriate to calculate φ only when you're working with a 2 x 2 contingency table (i.e. a table with exactly two rows and two columns).

A value of **φ = 0.1** is considered to be a small effect, 0.3 a medium effect, and 0.5 a large effect.

b. Cramer's V (V)

Cramer's V is calculated as $V = \sqrt{(X^2 / (n*df))}$

where X^2 is the Chi-Square test statistic, n = total number of observations and
df = degree of freedom for Chi square test = (no. of rows - 1) * (no. of columns - 1)

It's appropriate to calculate V when you're working with any table larger than a 2 x 2 contingency table.

Degrees of freedom	Small	Medium	Large
1	0.10	0.30	0.50
2	0.07	0.21	0.35
3	0.06	0.17	0.29
4	0.05	0.15	0.25
5	0.04	0.13	0.22

c. Odds Ratio (OR) and 95% confidence interval of odds ratio

Odds Ratio can be calculated for a 2 X 2 contingency table as follows

Exposure categories	Success (frequency)	Failures (frequency)
Treatment group (exposed)	A	B
Control group (not exposed)	C	D

Odds of success among exposed group (treatment group) = A / B

Odds of success among not exposed group (control group) = C / D

Odds ratio of success = Odds of success among treatment group / Odds of success among control group

= (A / B) / (C / D) = AD / BC

If odds ratio of success is higher than 1 and the 95% confidence interval of Odds ratio is not including 1, the treatment is having a higher odd of success.

If odds ratio of success is lower than 1 and the 95% confidence interval of Odds ratio is not including 1, the treatment is having a lower odd of success.

Odds Ratio can be calculated for a case control study design as follows.

Exposure categories	Diseased individuals (cases)	Not diseased individuals (controls)
Exposed to risk factor	A	B
Not exposed to risk factor	C	D

Odds of exposure to risk factor among cases = A / C

Odds of exposure to risk factor among controls = B / D

Odds ratio = Odds among treatment group / Odds among control group
= (A / C) / (B / D) = AD / BC

d. relative risk or risk difference

Relative risk can be calculated by dividing the incidence of disease among exposed group by the incidence of disease in the non-exposed group

Relative risk can be applied for a 2 X 2 contingency table

Exposure categories	Success (frequency)	Failures (frequency)
Treatment group (exposed)	A	B
Control group (not exposed)	C	D

Incidence of success among the exposed group (treatment group) = A / (A+B)

Incidence of success among the not exposed group (control group) = C / (C+D)

Relative risk or Risk ratio of success among exposed group as compared to the not exposed group = Incidence of success among the exposed group / Incidence of success among the not exposed group
= [A / (A+B)] / [C / (C+D)]

If relative risk of success among exposed group is higher than 1 and the 95% confidence interval of relative risk is not including 1, the treatment or exposure is having a higher risk of success.

If relative risk of success among exposed group is lower than 1 and the 95% confidence interval of relative risk is not including 1, the treatment is having a lower risk of success.

Conclusion

Effect size measures in scientific publications can communicate the actual difference between groups or the estimate of association between the variables, not just if the association or difference is statistically significant. The researchers can make their findings more interpretable, by displaying a suitable measure of effect size. Effect size measure can help the researchers to do meta-analysis by combining the data from multiple research articles.

References

Cohen, J. (1990). Things I have Learned (So far). *Am. Psychol.*, 45(12), 1304-1312. https://psycnet.apa.org/doi/10.1037/0003-066X.45.12.1304; Devroye, L. (1986). Non-

Uniform Random Variate Generation. Springer, New York, NY. https://doi.org/10.1007/978-1-4613-8643-8

Complete Dissertation by Statistics Solution. (2021, August 2), Effect size. Statistics Solutions (Retrieved November 29, 2021). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/effect-size/.

Effect Size in Statistics - Ultimate Guide. (2021). SPSS Tutorials. https://www.spss-tutorials.com/effect-size/

Kline, R. B. (2004). Beyond Significance Testing: Reforming Data Analysis Methods in Behavioural Research Washington, DC: *Am. Psychol. Assoc.*, 2004, 336, (ISBN 1-59147-118-4). https://www.apa.org/pubs/books/4316031

Publication Manual of the American Psychological Association (6th ed.). (2010). Washington, DC. *Am. Psychol. Assoc.* https://apastyle.apa.org/products/publication-manual-7th-edition

Stephanie, G. (2018). "Hedges' G: Definition, Formula." https://www.statisticshowto.com/hedges-g/

Stephanie, G., (2017). Pooled Standard Deviation, In: Elementary Statistics for the Rest of Us! https://www.statisticshowto.com/pooled-standard-deviation/

Sullivan, G.M., Feinn, R. (2012). Using Effect Size-or why the P Value is Not Enough. *J. Grad. Med. Educ.*, *4*(3), 279-282. https://doi.org/10.4300/jgme-d-12-00156.1