Open lecture on statistics

ISSN 2234-7658 (print) / ISSN 2234-7666 (online)
http://dx.doi.org/10.5395/rde.2015.40.4.328

RDE Restorative Dentistry & Endodontics

Statistical notes for clinical researchers: effect size

Hae-Young Kim*

Department of Health Policy and Management, College of Health Science, and Department of Public Health Sciences, Graduate School, Korea University, Seoul, Korea

In most clinical studies, *p* value is the final result of data analysis. A small *p* value is interpreted as a significant difference between the experimental group and the control group. However, reporting *p* value is not enough to know the actual difference. Problem of *p* value is that it depends on the sample size, *n*. Even a trivial meaningless difference can result in an extremely small *p* value when sample size is large. To make up this weak point, we need to report the 'effect size' as well as the *p* value. Effect size is a simple way to show the actual difference, which is independent of the sample size.

1. Reporting *p* value is not enough

In statistical testing we set a null hypothesis first and calculate the test statistic such as *t* values under an assumption of the null hypothesis. Finally, a *p* value is obtained which represents the probability of observing the current data due to chance when the null hypothesis is true. In most scientific articles, we usually make conclusion based on *p* values compared to the alpha error level chosen, e.g., 0.05. A smaller *p* value than alpha level is interpreted as a statistical significance. However, there are serious problems in relying on the *p* value only.

First, depending on the sample size, a wide range of *p* values can be obtained with the same size of difference, which can lead to contradictory results: either statistically significant or insignificant conclusions. Examples 1 and 2 in Table 1 have the same trivial difference of 3 between before and after treatments, assuming a clinically meaningful difference as 10. Two results were contradictory: statistically significant ($p = 0.001$, Example 2) and insignificant ($p = 0.382$, Example 1) depending on whether the sample size is large ($n = 10{,}000$) or small ($n = 100$). Moreover, as appeared in Example 2, it is a serious problem that clinically meaningless condition is concluded as statistically significant. The treatment in example 2 is clinically insignificant but statistically significant! What would you reasonably conclude on this case? This is a problem caused by using inappropriately large sample sizes.

Second, the information provided by the size of *p* value is confusing, because it is confounded by the sample size. We may expect that a small *p* value can tell us some information on how much difference exists between the observed data and the assumption of null hypothesis. However, the same size of *p* values can be obtained from quite different situations. Example 2 with a trivial effect and larger sample size and Example 3 with a substantial effect and smaller sample size both show the same *p* value 0.001 in Table 1. The result shows that *p* values are confounded with the sample size.

Two problems above can be overcome by controlling the sample size. To avoid this discordant situation, sample size determination procedure must be performed in the design stage in an experimental study. We generally need to calculate appropriate sample size in consideration with difference, SD, alpha error and power in the study design stage. The conclusion of significance testing is reliable only when an appropriate sample size was applied in a study. When we analyze a survey data with a

***Correspondence to**
Hae-Young Kim, DDS, PhD.
Associate Professor, Department of Health Policy and Management, College of Health Science, and Department of Public Health Sciences, Graduate School, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, Korea 02841
TEL, +82-2-3290-5667; FAX, +82-2-940-2879; E-mail, kimhaey@korea.ac.kr

large sample size, we need to consider the effect of large sample size in the interpretation of the test results.

Also the weakness of *p* value can be compensated by considering the effect size coincidently. As shown in Table 1, effect sizes exactly reflect the magnitude of actual effect, as displayed by 0.03 for a trivial difference and 0.3 for a substantial one.

Table 1. Examples of results of significant testing using *p* value and comparative effect size

Example	Before	After	SD*	Diff.	*n*	*t* value	*p* value	Effect size	Characteristics
1	145	142	100	3	100	$=\frac{0.3}{3/\sqrt{100}}$	0.382	$=\frac{0.03}{3/100}$	Trivial effect & insignificant
2	145	142	100	3	10,000	$=\frac{3}{3/\sqrt{10,000}}$	0.001	$=\frac{0.03}{3/100}$	Trivial effect & significant
3	145	115	100	30	100	$=\frac{3}{3/\sqrt{100}}$	0.001	$=\frac{0.3}{30/100}$	Substantial effect & significant

*SD, standard deviation.

2. What is effect size?

'Effect size' is simply a way of quantifying difference between compared groups, in other words, the actual effect.[1] While a *p* value has an important meaning in statistical inference, an effect size is expressing a descriptive importance. In Table 1, the effect sizes were expressed as the difference between two group means divided by the standard deviation of the group. When we compared Example 2 and Example 3, their effect sizes are a quite different as 0.03 and 0.3, while their *p* values are the same. Let's suppose clinicians generally think a change of at least '10' is clinically meaningful while a change of 3 after treatment is negligible. Therefore, they would not apply the treatment for the small change 3, even though the statistical significance test concluded the treatment is effective based on highly significant *p* value. Contrarily, they would apply the treatment in Example 3 because they can expect a substantial change of '30', and the statistical test concluded its significance. The results show that effect size exactly reflects the actual difference or effect. Therefore, reporting both the *p* value and the effect size is necessary in order to consider both statistical significance and actual clinical significance.

3. Types of effect size

Generally, there are two types of common effect size indices: standardized difference between groups and measures of association between groups. Table 2 shows the types of effect size indices and general standards of small, medium, and large effect for each type of effect size.

1. Between groups
1) Cohen's d or Glass's Δ: Defined by difference between two group means divided by standard deviation for continuous outcomes. Cohen's d is calculated by dividing pooled standard deviation under assumption of the equal variances while Glass's Δ is obtained by dividing the standard deviation of control group.
2) Odds ratio: Defined by ratio of odds of two compared groups for binary outcomes.
3) Relative ratio: Defined by ratio of proportions of two compared groups for binary outcomes.

2. Measures of association
1) Pearson's *r* correlation: Effect size representing association of two variables.
2) Pearson *r* correlation coefficient: The amount of variation explained.

Table 2. Common effect size indices[2]

Index		Description	Standard	Comment
Between groups	Cohen's d or Glass's Δ	d or Δ = ($Mean_1$ - $Mean_2$) / SD* d: use pooled SD Δ: use SD of control group	Small 0.2 Medium 0.5 Large 0.8 Very large 1.3	For continuous outcomes
	Odds ratio (OR)	OR = $odds_1$ / $odds_2$	Small 1.5 Medium 2 Large 3	Degree of association between binary outcomes
	Relative risk or risk ratio (RR)	RR = p_1 / p_2	Small 2 Medium 3 Large 4	For binary outcomes, ratio of two proportions
Measures of association	Pearson's *r* correlation	Range -1 to 1	Small ± 0.2 Medium ± 0.3 Large ± 0.5	Measures the degree of linear relationship
	Pearson *r* correlation coefficient	Range 0 to 1	Small 0.04 Medium 0.09 Large 0.25	Proportion of variance explained

*SD, standard deviation.

4. Interpretation of effect size

Then, how would we interpret the degree of effect size? An effect size is exactly equivalent to a Z score of a standard normal distribution. Assume that all data are normally distributed. If Cohen's d is calculated to be zero, it means that there is no mean difference between two comparative groups and the position of the mean of experimental group is exactly the same with the mean of control group. Therefore, 50% of observations in control group locate below the mean of experimental group (Table 3). The relative 'small' effect size '0.2' means the mean of experimental group is located at 0.2 standard deviation above the mean of control group. The Z score of 0.2 is at 58th percentile which have 58% of observations below in control group (Figure 1). Similarly, the Cohen's d values 0.5 and 0.8 locate at 69th and 79th percentile of the distribution of the control group, respectively.

Table 3. Interpretation of Cohen's d which represents a standardized difference [($Mean_1$ - $Mean_2$) / SD*][1,3]

Relative size	Effect size	% of control group below the mean of experimental group
	0.0	50%
Small	0.2	58%
Medium	0.5	69%
Large	0.8	79%
	1.4	92%

*SD, standard deviation.

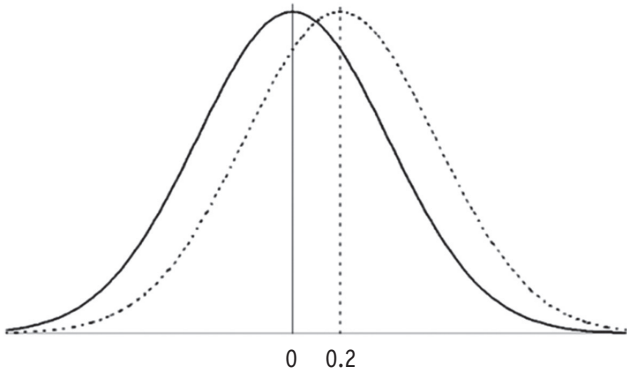

Figure 1. Distribution of control group (solid line) and experimental group (dotted line), and position of Cohen's d = 0.2.[1]

http://dx.doi.org/10.5395/rde.2015.40.4.328

5. Conversion of effect sizes to Pearson *r* correlation coefficient

Pearson *r* correlation coefficient is an effect size which is widely understood and frequently used. Converting various statistic values including *t* or *F* into Pearson *r* correlation coefficient may be advantageous because it facilitates interpretation. Also Cohen's d can be converted into *r*. Table 4 provides the conversion formula and a brief explanation.

Table 4. Conversion from various statistics to Perason *r* correlation coefficient association measures[3]

Statistic	Conversion formula	Comment
$\chi^2 df = 1$	$r = \sqrt{\frac{\chi^2 df = 1}{N}}$	A single degree of freedom chi-square value divided by the number of cases
t	$r = \sqrt{\frac{t^2}{t^2 + df}}$	From *t* value to *r* correlation coefficient
F	$r = \sqrt{\frac{F(df = 1,\_)}{F(df = 1,\_) + df(error)}}$	From *F* value with single freedom numerator to *r*
Cohen's d	$r = \sqrt{\frac{d^2}{d^2 + 4}}$	From Cohen's d to *r*

6. Summary

Though *p* values give information on statistical significance, they are confounded with the sample size. Effect size can make up the weak point, by providing information on the actual effect which is independent of the sample size. Therefore, reporting the effect size as well as the *p* value is recommended.

References

1. Coe R: It's the effect size, stupid: what effect size is and why it is important. Paper presented at the 2002 Annual Conference of British Education Research Association, University of Exeter, Exeter, Devon, England, September 12-14, 2002. Available from: http://www.leeds.ac.uk/educol/documents/00002182.htm (updated 2015 Sep 6).
2. Sullivan GM, Feinn R. Using effect size – or why *p* value is not enough. *J Grad Med Educ* 2012;4:279-282.
3. Becker LA: Effect size (ES). Available from: http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Lehre/StatIIKrim/EffectSizeBecker.pdf (updated 2015 Sep 6).