MAE Vet Fak Derg, 6 (1): 28- 33, 2021
DOI: 10.24880/maeuvfd.823493

Review

The use of effect size in veterinary medicine

Pınar AMBARCIOĞLU[1]

[1]Department of Biostatistic, Faculty of Veterinary Medicine, Mustafa Kemal University, Hatay/TURKEY

Key Words:
clinical significance
effect size
hypothesis testing
power analysis
statistical significance

Anahtar Kelimeler:
etki büyüklüğü
güç analizi
hipotez testleri
istatistiksel anlamlılık
klinik anlamlılık

Received : 09.11.2020
Accepted : 25.12.2020
Published Online : 30.04.2021
Article Code : 823493

Correspondence:
P. AMBARCIOĞLU
(kpambarcioglu@mku.edu.tr)

ORCID:
P. AMBARCIOĞLU : 0000-0001-6572-4219

ABSTRACT

Effect size is a statistical index that measures the magnitude of the effect generated by the variable of interest in a study, in a sense, reflecting the practical or clinical value of the study in addition to the statistical results. In recent years, it has become preferable to report the effect size expressing practical significance in addition to the statistical significance expressed by the p-value in hypothesis tests in scientific research, and even it has been required by some scientific journals. By reporting the effect size, it is possible to use it in statistical power analysis, to compare the results of the studies, and to determine the amount of the effect in the study. In this study, by mentioning the concept of effect size, the main effect size indices used according to research types are introduced. In addition, the calculation methods of the effect size indices commonly used for continuous and categorical outcome variables were given and interpreted with scenarios from the field of veterinary medicine. In conclusion, in order to be able to interpret the results of a study in clinical or practical terms, to present the analyzed data in more detail than the p-value, and to ensure its use in power analysis, it was suggested that researchers report effect size in their studies.

Etki büyüklüğünün veteriner hekimliği alanında kullanımı

ÖZ

Etki büyüklüğü, bir çalışmada ilgilenilen değişkenin meydana getirdiği etkinin büyüklüğünü ölçen, bir anlamda çalışmanın istatistiksel sonuçlarına ek olarak, pratik veya klinik anlamdaki değerini de yansıtan istatistiksel bir indekstir. Son yıllarda bilimsel araştırmalarda hipotez testlerinde p değeri ile ifade edilen istatistiksel anlamlılığa ek olarak pratik anlamlılığı ifade eden etki büyüklüğünün de raporlanması tercih edilir hale gelmiş, hatta bazı bilimsel dergiler tarafından zorunlu kılınmıştır. Etki büyüklüğünün raporlanması ile istatistiksel güç analizinde yararlanılması, çalışmaların sonuçlarının karşılaştırılması ve çalışmada belirlenen etkinin miktarının belirlenmesi mümkün olmaktadır. Bu çalışmada etki büyüklüğü kavramından bahsedilerek, araştırma türlerine göre kullanılan başlıca etki büyüklüğü indeksleri tanıtılmıştır. Ayrıca, sürekli ve kategorik sonuç değişkenler için yaygın olarak kullanılan etki büyüklüğü indekslerinin hesaplama yöntemleri verilerek, veteriner hekimliği alanından senaryolar ile örneklendirilmiş ve yorumlanmıştır. Sonuç olarak, klinik veya pratik anlamda çalışma sonuçlarını yorumlayabilmek, incelenen veriyi p değerinden daha ayrıntılı şekilde sunabilmek, güç analizlerinde kullanımını sağlamak amacıyla, araştırmacıların çalışmalarında etki büyüklüğü raporlaması önerilmiştir.

INTRODUCTION

Statistical significance tests have a history dating back to the 1700s. It was first used by the Scottish physician John Arbuthnot in evaluating the birth rate in London according to the sex of the newborn babies, but its use was not widespread until the 1900s (1). The use of hypothesis tests has become widespread with the development of Karl Pearson's Chi-Square Goodness of Fit Test in 1900, William S. Gossett's Student t-Test in 1908, and Ronald Fisher's Analysis of Variance (ANOVA) in 1918 (2,3,4). It became popular after Fisher published his first book in 1925, Statistical Methods for Research Workers, and then The Design of Experiments in 1935 (1,5,6).

Although the decision-making process based on observation data in scientific research is actually a complex process, the fact that the deterministic algorithm of hypothesis testing approach has reduced this process to a dichotomous form of accepting and rejecting the hypothesis. For this reason, the hypothesis testing approach has spread rapidly as an easy-to-use process for researchers (7). However, over time, as researchers perceived this process only as a tool used to reach statistical significance, it has become difficult to reach scientific generalizations, and consequently, it received serious criticism (8,9,10). Criticisms of the hypothesis testing approach argue that the effect size should be used to express practical or clinical significance, in addition to the statistical significance expressed by hypothesis testing. Indeed, organizations such as the American Educational Research Association (AERA) and the American

Bu makaleye atıfta bulunmak için/ To cite this article:
Ambarcıoğlu P. The use of effect size in veterinary medicine. MAE Vet Fak Derg. 2021; 6 (1) :28-33.

Psychological Association (APA) have made it mandatory to report effect sizes and confidence intervals in their publishing guidance (11). In summary, hypothesis tests provide information only about the probability of confirming the null hypothesis of observed data, as mentioned above. This information, used as the p-value, forces the researcher to make a dichotomous decision in the form of rejecting or not being able to reject the null hypothesis (12). In order to go beyond this process, additional values such as statistical power, effect size, and confidence interval should be evaluated (13).

1. What is effect size?

The effect size is the statistical value showing the deviation level between the results obtained from the sample and the expectations defined in the null hypothesis (14). Effect size is also defined as a statistical index that measures the magnitude of the effect created by the variable of interest in a study and, in a sense, reflects the practical or clinical value of the study in addition to the statistical results.

Including the effect size while reporting the research results generally serves three main purposes;

- The first of these is the use of the effect size in the statistical power analysis to calculate the sample size at the beginning of the study. Effect size is an important part of statistical power analysis. Although not applied consciously, the effect size contributes to a good experiment design (12). In other words, during the power analysis, the required sample size is chosen on purpose, taking into account the importance of the effect between the phenomena of interest, the sensitivity of the tools used to detect this effect, and the research design (13).

- The second purpose of using effect size is to allow comparison between studies answering the same hypothesis. These studies may have been done using different test statistics, different sample sizes, and designs. Therefore, effect size, which is a standardized index that eliminates different features between studies, is needed in order to compare study results (13).

- Reporting the effect size also makes it possible to interpret the magnitude of effect determined in the studies. In addition to making comparative interpretations of different studies, it is also possible to classify a single effect size as small, medium, large effect size as determined by Cohen. Cohen states that the cut-off values he gives for the interpretation of the effect level will be useful in new areas where there are not many studies. That is, when an effect is observed in a study, it is functional if there are no studies that can be compared to understand its magnitude (14). The classification of some effect size indices of most common used statistical tests are given in Table 1 (14, 15)

2. The calculation and interpretation of effect size

Just as there are different hypothesis tests used for different research designs in inferential statistics, effect size calculations also vary according to the structure of the variables. It is possible to evaluate the frequently used effect size indices under two main titles: those used for continuous outcomes and dichotomous outcomes.

Table 1. Effect size classifications for common used test

		Classification		
	Test	**Small**	**Medium**	**Large**
Cohen's d	t-Test	0.20	0.50	0.80
Cohen's f	Variance analysis	0.10	0.25	0.40
f^2	Regression analysis	0.02	0.15	0.35
Odds ratio (OR)	Contingency tables (2x2)	1.5	2	3
Risk ratio (RR)	Contingency tables (2x2)	2	3	4
W (Φ)	Contingency tables	0.10	0.30	0.50
Cohen's h	Contingency tables	0.20	0.50	0.80
r	Correlation	±0.20	±0.50	±0.80
r^2	Variance analysis/ Regression analysis	0,04	0,25	0,64

2.1. Effect size indices for continuous outcomes

In the research design where the means of the two independent groups are compared, the effect size can be calculated by the mean difference or standardized mean difference.

2.1.1. Mean difference

Let us assume that one compares the monthly live weight gains of Angus and Simental cattle in a breeding farm. The mean and standard deviation values of the live weight gain of two breeds are given in Table 2. It is seen that the difference between the means of the two groups, ie the effect size, is d = 9.03-7.46 = 1.57. However, it is difficult to comment on the difference between groups based on the pure mean difference. Because this difference is also related to the variation in the dependent variable. If the dependent variable is distributed with a wide variation, the difference of 1.57 units represents a very small effect, while the dependent variable is distributed in a narrow range may infer that the difference of 1.57 units is a significant effect.

Table 2. The measurements of live weight gain of Angus ve Simental cattle

	N	Mean	Standard Deviation	Variance
Angus	150	9.03	1.96	3.84
Simental	150	7.46	2.13	4.53

2.1.2. Standardized mean difference

If there is a predetermined standard of measurement for the variable of interest, it may be possible to comment on the effect of the difference between the two groups. However, as it is seen in the above example and most studies, generally there is no standard scale for the variable of interest. Therefore, in order to comment on the amount of difference between means, it is necessary to evaluate the means together with the variations of the distributions (16). Accordingly, the effect size of two independent group designs is calculated as in Equation 1.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{S_{pooled}} \quad (1)$$

$$S_{pooled} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \quad (2)$$

In the formula for the effect size expressed as "Cohen's d" $\bar{x}_i$ refers to the mean, S_i^2 refers to the variance and n_i refers to the sample size of each group.

In the example in Table 2, Angus and Simental beef cattle in a farm were intended to be compared in terms of monthly live weight gains.The variance (V) and standard error of d are calculated as in Equations 3 and 4, respectively (17).

$$V_d = \frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)} \quad (3)$$

$$SE_d = \sqrt{V_d} \quad (4)$$

Herefrom;

$$S_{pooled} = \sqrt{\frac{(150-1)3{,}84 + (150-1)4{,}53}{150+150-2}} = 2{,}04$$

$$d = \frac{9{,}03-7{,}46}{2{,}04} = 0{,}77$$

When the above formulas are examined, it does not seem difficult to estimate d, if the population parameters are known or it is possible to obtain the data of interest. However, as is frequently encountered today, there may be a new variable that has not been subjected to any experiment, and has no available data. Under these conditions, it is not possible to obtain the required mean difference and standard deviation information for calculating the effect size. For similar cases, Cohen developed the categories of "small", "medium" and "large" effect size and enabled an approximate interpretation (14). For example, a new nutrition program is aimed to compare, which is thought to affect the milk yield of Holstein breed cows, with the standard nutrition program in terms of milk yield. The effect size was around d = 0.2-0.3 in expectation of small and the effect size could be around d = 0.8-1.00 in expectation of large. When the effect of the nutrition program on milk yield is expected to be moderate, the effect size may be about d = 0.5. If this interpretation is to be generalized, it can be expressed as (14);

d ≅0.20 small effect size

d≅0.50 medium effect size

d≅0.80 large effect size

Hedges suggested a degree of freedom correction as in Equation 5 because Cohen's d overestimates the effect size when the sample size is small (18).

$$J(df) = 1 - \frac{3}{4df-1} \quad (5)$$

$$g = j(df)d \quad (6)$$

$$V_g = [J(df)]^2 V_d \quad (7)$$

$$SE_g = \sqrt{V_g} \quad (8)$$

2.2. Effect size indices for dichotomous outcomes

If both dependent and independent variables are dichotomized, the most frequently used effect sizes are; risk difference, risk ratio, or odds ratio.

2.2.1. Risk difference

The effect size which expresses the difference between the two proportions P_1 and P_2 is shown as

$$j = P_1 - P_2 \quad (9)$$

But for example, let be $P_1 = 0.65$ and $P_2 = 0.45$, the effect size is calculated as $j = 0.20$; and let be $P_1 = 0.25$ and $P_2 = 0.05$, the effect size is still calculated as $j = 0.20$. This situation shows that the index j is insufficient to scale equal units. Therefore Cohen developed the index h in Equation 11, which he obtained with a non-parametric transformation on P values (14).

$$\Phi = 2 \arcsin \sqrt{P} \quad (10)$$

$$h = \Phi_1 - \Phi_2 \quad (11)$$

A generalization can be made about the interpretation of the index h as follows (14):

h≅0.20 small effect size

h≅0.50 medium effect size

h≅0.80 large effect size

2.2.2. Risk ratio

Risk ratio or relative risk (RR) is another effect size index frequently used in cross-sectional or prospective studies (17). It expresses the ratio of the probability of observing the event of interest in two independent samples.

$$RR = {}^{P_1}/_{P_2} \quad (12)$$

$$SE_{\ln(RR)} = \left(\frac{1-P_1}{n_1 P_1} + \frac{1-P_2}{n_2 P_2}\right)^{1/2} \quad (13)$$

As an illustrative example, let the data of a research design investigating the efficacy of the drug A developed for the treatment of Feline Infectious Peritonitis (FIP) disease seen in cats, compared to placebo, given in Table 3.

According to Table 3, the risk of disease occurrence in cats treated with placebo is calculated as $P_1 = 40 / 50 = 0.80$, while the risk of disease occurrence in cats treated with drug A is calculated as $P_2 = 5 / 50 = 0.10$. Accordingly, the risk ratio is found as RR= $P_1 / P_2 = 0.80 / 0.10 = 8$. This result is interpreted as the risk of disease in cats treated with placebo is 8 times higher than in cats treated with drug A. The point to be considered in relative risk is that one of the ratios of interest should belong to the unpreferable situation and the other to the preferred situation (17).

Table 3. The distribution of FIP positive and FIP negative cases in Drug A and Placebo

	Y		
X	**FIP positive**	**FIP Negative**	**Total**
Placebo	40	10	50
Drug A	5	45	50
Total	45	55	100

2.2.3. Odds ratio

Odds is defined as the ratio of the probability of occurrence of an event to the probability of non-occurrence. And the odds ratio (OR) is defined as the ratio of the odds of two groups (eg, treatment and placebo groups) whose effects were examined (19). While the risk ratio is an effect size measure used in cross-sectional and prospective studies, the odds ratio can also be used in retrospective research design (17). The observed positive and negative values of the X and Y variables are given in Table 4, and the calculation of the odds ratio according to these values in Equation 14 and the calculation of its standard error (SE) in Equation 15 are shown.

Table 4. Odds Ratio

	Y		
X	**Positive**	**Negative**	**Total**
Positive	n_{11}	n_{12}	$n_{1.}$
Negative	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} \quad (14)$$

$$SE_{OR} = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2} \quad (15)$$

Odds ratio based on the example in Table 4 is calculated as;

$$OR = 40x45/5x10 = 36$$

According to this result, it is interpreted that the likelihood of being positive for FIP disease in cats treated with placebo is 36 times more than cats treated with drug A. In other words, cats treated with drug A had 36 times more likelihood of recovery than cats treated with placebo.

CONCLUSION

In this study, we evaluated not only the effect size of calculations that can be used in different research designs but also evaluated how these calculated indexes should be interpreted. Furthermore, Cohen's effect size classification as "small", "medium" and "large" is also mentioned. Some researchers attribute Cohen's popularity about effect size to this classification system that he brought to the interpretation of effect size (10, 20). However, Vacha-Haase and Thompson argued that the use of this classification system is unreasonable, as it resembles the rigidity in the $p < 0.05$ system used in the hypothesis testing approach (10). Therefore, it was stated that a specific evaluation should be made for each study without sticking to a classification in the interpretation of the effect size. For example, in a study investigating the effect of smoking on lifetime, even if the effect size is found to be low, this is considered a valuable result. Because first of all, the outcome we are interested in is, the lifetime, clinically very important and it would also be seen that it is approximately similar to the effect size found in previous studies conducted on the same subject. Accordingly, while interpreting the effect size, interpretation should be made by considering both the characteristics of the outcome evaluated in the study and the effect sizes found in previous studies on the same subject.

The recomendations using effect size in addition to p-value aim to overcome the deficiencies of p-value. The most important limitation of the p-value is that it is affected by the sample size. Even though the effect size is zero or very small, p-value would indicate a statistically significant difference, if the sample size is adequately big. Statistically significance depends upon both effect size and sample size, while effect size is independent of sample size. The other limitation of p-value is that it is provided information only about the existance of the effect, not its effect. Thus, reporting only the p-value is not sufficient to fully understand the results (15).

Finally, it should be noted that, even though Cohen's small-medium-large effect size classification seems like it prevents to avoid the inflexibility of the p-value, it can be used as a rough guide in the absence of any preliminary information during the design phase of the research. In addition, researchers should prefer to report effect size to give information about the amount of the effect revealed in the intervention.

DECLARATIONS

Ethics Approval

Not applicable.

Conflict of Interest

The authors declare that they have no competing interests.

Author Contribution

Idea, concept and design: P Ambarcıoğlu

Data collection and analysis: P Ambarcıoğlu

Drafting of the manuscript: P Ambarcıoğlu

Critical review: P Ambarcıoğlu

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Thompson B. "Statistical", "Practical", and "Clinical":How many kinds of significance do counselors need to consider?. Journal of Counseling&Development, 2002; 80: 64-71.

2. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 1900; 50: 157-175.

3. Student. The probable error of a mean. Biometrika, 1908; 6: 1–25.

4. Fisher RA, The correlation between relatives on the supposition of mendelian inheritance. Transactions of the Royal Society of Edinburg, 1918, 53: 399–433.

5. Fisher, R.A. (1925). Statistical methods for research workers. (11th ed. rev.). Edinburgh: Oliver & Boyd.

6. Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver & Boyd.

7. Yıldırım HH, Yıldırım S (2011). On hypotesis testing, confidence interval, effects size and noncentral probability distributions. Elementary Education Online, 2011; 10(3): 1112-1123.

8. Cohen J. The earth is round ($p<0.05$). American Psychologist, 1994; 49: 997-1003.

9. Rosnow RL, Rosenthal R. Statistical procedures and the justidication of knowledge in psychological science. American Psychologist, 1989; 44: 1276-1284.

10. Vacha-Haase T, Thompson B. How to estimate and interpret various effect sizes. Journal of Counselig Psychology, 2004; 51(4): 473-481.

11. American Psychological Association (APA) Guide, Sixth Edition, 2010.

12. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Review, 2007; 82: 591-605.

13. Işık İ. Yokluk hipotezi anlamlılık testi ve etki büyüklüğü tartışmalarının psikoloji araştırmalarına yansımaları. Eleştirel Psikoloji Bülteni, 2014; 5: 55-80.

14. Cohen J (1988). Statistical power analysis for the behavioral sciences. (Second edit.). New Jersey: Lawrence Erlbaum Associates

15.Sullivan GM, Feinn R. Using effect size-or why the p value is not enough? Journal of Graduate Medical Education, 2012; 4 (3): 279-282.

16. Coe R. (2002). It's the effect size, stupid: What effect

size is and why it is important. Annual Conference of the British Educational Research Association, 12-14 September 2002, England.

17. Cooper H, Hedges LV, Valentine JC (1994). The handbook of research synthesis and meta-analysis. New York, Richard Sage Foundation. 221-253

18. Hedges LV. Distribution Theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 1981; 6(2): 107-128.

19. Süt N, Şenocak M. Relatif risk ölçütünün odds oranı, atfedilen risk ve tedaviye gerekli sayı ölçütleriyle karşılaştırılması. Trakya Üniversitesi Tıp Fakültesi Dergisi, 2007; 24(3): 213-221.

20. Kirk RE (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurements, 1996; 56: 746-759.