Hanyang Medical Reviews

Hanyang Med Rev 2015;35:40-43
http://dx.doi.org/10.7599/hmr.2015.35.1.40
pISSN 1738-429X eISSN 2234-4446

효과크기의 이해

남상건
분당서울대학교병원 마취통증의학과

Understanding Effect Sizes

Francis Sahngun Nahm

Department of Anesthesiology and Pain Medicine, Seoul National University Bundang Hospital, Seongnam, Korea

In most medical research the P value is commonly used to describe test results. Because the power of statistical test is influenced by sample size, the null hypothesis can be rejected ($P < 0.05$) in most cases if the sample size is tremendously big even if the real difference (or relationship) is extremly small. To overcome the weakness of using the P value, effect size can be used in the statistical analysis. Effect size can be defined as the "degree to which the phenomenon (difference or relationship) is present in the population". The effect size is used in sample size calculation, data interpretation and conducting meta-analysis. This manuscript describes limitations in using the P value and further introduces the concept of effect size.

Key Words: Data Interpretation, Statistical; Meta-Analysis; Research Design

Correspondence to: Francis Sahngun Nahm
우463-707, 경기도 성남시 분당구 구미로 173번길 82, 분당서울대학교병원 마취통증의학과
Department of Anesthesiology and Pain Medicine, Seoul National University Bundang Hospital, 82 Gumi-ro 173 beon-gil, Bundang-gu, Seongnam 463-707, Korea
Tel: +82-31-787-7499
Fax: +82-31-787-4063
E-mail: hiitsme@snubh.org

Received 30 November 2014
Revised 14 December 2014
Accepted 31 December 2014

서 론

대부분의 의학 논문에서 연구 결과를 보고할 때 주로 이용되는 방법은 P value를 표시하는 방법이다. 의학 연구 논문은 먼저 귀무가설을 정하고, 유의 수준을 미리 설정한 후 통계적 유의성을 검정하는 방식으로 연구 결과를 기술하게 된다. 그러나 최근에는 이러한 통계 분석 과정의 한계점을 인식하고 effect size (효과크기)를 통해 실제 존재하는 유의성을 살펴야 한다는 주장도 있다[1-3]. 이번 지면을 통해 효과크기의 의미와 이용에 대해 소개하려고 한다.

효과크기사용의 필요성

의학 연구에서는 주로 P값을 이용하여 가설을 검정하게 된다. 통상적으로 $P < 0.05$인 경우 통계적으로 유의 하다고 간주되나, 이 방법은 표본 수에 많은 영향을 받는다는 단점이 있다. 예를 들어 적절한 표본 수 산출 과정을 통해 연구에 필요한 표본 수를 계산한다 하였더라도, 표본 수 계산 과정이 많은 가정을 바탕으로 이루어지기 때문에 실제 연구 과정에서 기존 가정과 다른 상황이 발생하게 되는 경우 $P > 0.05$인 결과가 나오게 되어 연구자들을 난감하게 하는 경우가 있다. 이와는 반대로 표본 수가 수천만 개인 상황을 가정하면, 비교하려는 집단 사이의 실제 차이 혹은 연관성이 매우 적은 상황에서도 overpower의 문제가 발생하여 $P < 0.05$인 현상이 발생하게 된다. 이는 P값이 표본 수에 많은 영향을 받기 때문이다. 따라서 기존의 P값을 사용하는 방법에 대한 대안으로 효과크기를 사용하는 방법을 고려할 수 있다. 효과크기란 비교하려는 집단 사이에 "얼마나" 차이(혹은 연관성)가 있는지를 나타내 주는 지표를 말한다.

P값은 "관찰된 현상이 기존에 알려진 확률 분포와 비교해 볼 때 어느 정도로 희귀하게 발생하는가"에 대한 확률적 의미를 가지고 있지만, 효과크기는 실제 관찰된 데이터에서 비교하려는 집단 사이의 차이(혹은 연관성)을 직접적으로 기술해 준다는 장점이 있다. 예를 들어, 고혈압 약제 A와 B 의 효과를 비교하는 연구를 진행함에 있어 어떤 연구자는 10 mmHg의 기준으로 유의하다고 판정하고, 다른 연구자는 5 mmHg의 기준으로 유의하다고 판정한다고 할 때, 실제 연구에서 두 약제의 혈압 강하 효과가 8 mmHg 차이가 관찰되었다면, 같은 연구 결과에 대해 두 연구자는 정반대의 연구 결론을 내리게 된다. 이러한 논란을 줄이기 위해 비교하려는 집단 사이의 실제 차이를 기술하는 것이 필요하다. 이렇듯 추론 통계에서 주로 P값이 사용된다면, 기술 통계에서는 효과크기가 사용되는 것이다.

효과크기란 무엇인가?

효과크기는 'standardized measure of the size of the mean difference or the relationship among the study groups'이라고 정의할 수 있다[4]. 즉, 비교하려는 집단들 사이의 차이 혹은 관계를 나타내는 '표준화 된 지표'를 의미한다. 효과크기가 0이라는 것은 비교 집단들 사이의 차이(혹은 연관성)가 없다는 것을 의미하여 귀무가설 하에서 "효과크기 = 0"이 된다. 평균치 비교의 경우 비교하려는 집단 사이에 평균 차이가 클수록 효과크기는 커지게 된다. 효과크기의 쉬운 예를 들면, 귀무가설 하에서 남녀의 성별의 비율이 50:50이라고 할 때, 어떤 집단에서 남녀의 비율이 53:47이라고 하면 효과크기는 3%가 된다. 또한 전 인구의 평균 IQ가 100이라고 할 때, 어떤 집단에서의 평균 IQ가 105라고 하면, 효과크기는 5 IQ unit이 된다. 이렇게 효과크기란 비교하려는 집단 사이의 실제 차이를 나타내는데 개별 측정 단위에 따라 해석이 어려워질 수 있다. 따라서 개별 측정 단위에 영향을 받지 않도록 효과크기를 그 산포 정도(표준편차 등)로 나누어 단위에 상관없이 사용할 수 있도록 한 index가 사용되며 이를 효과크기 인덱스(effect size index) 혹은 표준화 된 효과크기(standardized effect size)라고 한다. 그러나 실제로는 효과크기와 효과크기 인덱스 및 표준화 된 효과크기라는 용어는 혼용되며 일반적으로는 효과크기라는 용어가 더욱 널리 사용되고 있다. 이 논문에서도 효과크기라는 용어를 사용하기로 한다. 효과크기의 예는 다음과 같다(Table 1).

1. d (t distribution): 비교하려는 집단의 평균의 차이를 표준화한 효과크기(standardized mean difference)로 두 집단의 평균값의 차이를 통합표준편차(pooled standard deviation)로 나누어 얻게 된다. Cohen's d는 두 집단 평균의 차이를 표준화하기 위해 두 집단의 표준 편차를 고려하는 반면에, Glass' Δ는 두 집단의 차이를 통제 집단의 표준 편차로 나눈다는 점이 약간 다르다.
2. ρ (Product moment): 변수 사이의 연관성을 나타내는 효과크기로 상관계수를 기본으로 한다. 특히 피어슨의 상관계수 r은 자주 사용되는 지표이다.
3. h (Z distribution): Z 분포를 따르는 두 집단 사이의 비율의 차이를 나타내는 효과크기이다.
4. ω (χ^2 distribution): 카이제곱 검정 등 χ^2분포에서 사용되는 효과크기이다.
5. f (F distribution): 분산 분석 등 f 분포에서 사용되는 효과크기이다.
6. f^2 (multiple regression): 다중회귀 등에서 사용되는 효과크기이다.
7. 기타: 승산비(odds ratio), 상대위험도(relative risk) 등도 효과크기의 일종이다.

위에서 소개된 효과크기들은 다양한 계산 과정을 통해 서로 변환이 가능하기 때문에 메타 분석 등에서는 여러 가지 방법으로 측정된 효과크기를 종합하여 하나의 수치로 나타낼 수 있다.

Table 1. Various types of effect sizes [5]

Test	ES Index	Effect size		
		Small	Medium	Large
Comparison of two independent means (m_A vs m_B)	$d = \frac{m_A - m_B}{\sigma}$	0.20	0.50	0.80
Significance of product-moment r (correlation coefficient)	ρ	0.10	0.30	0.50
Differences between correlation coefficient (Z_A vs Z_B)	$q = Z_A - Z_B$	0.10	0.30	0.50
Differenes between proportions (P_A vs P_B)	$h = \phi_A - \phi_B$	0.20	0.50	0.80
Chi-square for goodness of fit and contingency	$\omega = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	0.10	0.30	0.50
One-way ANOVA	$f = \frac{\sigma_m}{\sigma}$	0.10	0.25	0.40
Multiple regression and correlation	$f^2 = \frac{R^2}{1 - R^2}$	0.02	0.15	0.35

효과크기가 갖는 의미

평균 비교에서 효과크기가 0이라는 의미는 비교하려는 집단 사이의 평균이 동일하다는 뜻이고, 효과크기의 값이 양수를 갖게 되면 비교 집단이 대조 집단에 비해 평균치가 크다는 의미이며, 음수의 값을 갖게 되면 비교 집단의 평균이 대조 집단에 비해 작다는 것을 의미한다. 그러나 연구 결과로 제시되는 효과크기는 일반인들이 이해하기 어렵다는 단점이 있다. 이러한 단점을 극복하기 위해서 Cohen은 그의 저서에서 효과크기를 임의로 small, medium, large로 구분하여 직관적으로 이해하기 쉽도록 소개하였다[5]. 그러나 여전히 Cohen이 제시한 값들은 상대적인 값이고 모집단이나 변수의 특징에 따라서 달라질 수 있으므로, 참고만 할 뿐 절대적으로 받아들여지는 데에는 한계가 있다. 예를 들어 두 집단의 평균 비교에 있어 effect size = 0.5 (medium)라면 어느 정도의 차이가 있는지에 대해서 짐작하기가 어렵다. 그런데, 효과크기의 원래 의미가 비교하려는 집단들 사이의 차이를 나타내는 '표준화 된 지표'라는 점을 생각하면 좀 더 이해가 쉽다. 즉, 효과크기의 값의 크기는 정규 분포의 Z값과 동일한 의미를 가지므로, 만일 effect size값이 1.96이라면 실험군의 평균은 대조군의 분포에서 상위 5%에 해당함을 의미한다 (Fig. 1).

효과크기 사용의 장점

첫째, 효과크기는 연구 결과의 해석을 이분법적인 방법이 아닌 연속 선상에서 할 수 있게 해준다. 기존에 사용되던 P값은 비교하려는 집단 사이에 미리 정해진 기준(일반적으로 유의 수준 alpha)에 따라 "유의한 차이가 있다/없다"로 해석할 수 밖에 없었으나, 효과크기를 이용하면 실제로 "얼마만큼의 차이(혹은 연관성)가 있는가"를 구체적인 수치로 보여줄 수 있다. 둘째, 효과크기는 P값과는 달리 표본 수에 의한 영향을 받지 않는다. P값은 표본 수가 적은 경우에는 검정력이 떨어져 유의하지 않은 결과가 얻어질 수 있고, 이와는 반대로 표본수가 매우 큰 경우에는 큰 의미가 없는 결과도 "통계적으로 유의"하게 얻어지는 단점이 있다. 셋째, 효과크기는 다양한 형태의 결과들을 비교 가능한 공통의 단위로 변화 시켜줌으로써 각기 다른 통계적 방법에 의해 시행된 연구 결과들을 수량적으로 통합하는 메타 분석을 시행할 때 하나의 공통 척도로 이용될 수 있다.

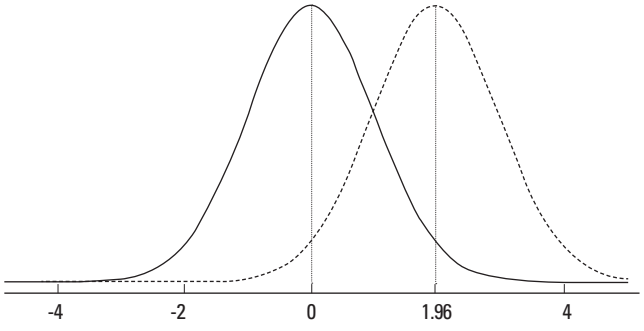

Fig. 1. Example of effect size = 1.96. The mean of the treatment group is in the range of upper 5% of the control group.

효과크기의 이용

효과크기는 비교하려는 집단 사이의 차이/혹은 연관성을 나타내 주기 때문에 표본 수 계산과 메타 분석에서 유용하게 이용할 수 있다.

1. 표본 수 계산: 표본 수 계산 과정에서 필요한 것은 단측/양측 검정 여부, alpha error, beta error, 비교하려는 집단 사이의 예상 평균 차이, 예상 표준 편차이다. 이 중 단측/양측 검정, alpha error와 beta error는 모두 주어진 값이므로 실제 필요한 것은 예상 평균 차이와 예상 표준 편차뿐이다. 예상 평균 차이를 예상 표준 편차로 나누어준 값이 바로 효과크기이기 때문에, 결국에는 표본 수 계산 과정은 효과크기를 추정하고 계산하는 과정이라 할 수 있다. 표본 수 계산을 위해 효과크기를 추정하는 방법으로는 1) 기존의 비슷한 연구 결과로부터 추정하거나 2) 비슷한 연구가 없는 경우에는 pilot study를 실시하여 추정하는 방법, 3) 이 모든 방법이 어려운 경우 단지 효과크기가 클 것인지(large)/중간 정도일지(medium)/작을지(small)만을 예상하여 이에 할당된 효과크기를 적용하는 방법이 있다. 이 방법은 Cohen에 의하여 제시된 방법으로[5] 현재까지 이용되고 있다.
2. 메타 분석에서의 이용: 효과크기는 비교하려는 집단 사이의 차이를 수치화하여 나타내 주는 지표이므로 여러 연구에서 발표된 결과를 수치화한 후 종합하면 메타 분석에서 이용될 수 있다. 개별 연구 결과의 기술 방법이 모두 다른 경우에도, 집단의 평균/표준편차, 검정통계량, 상관계수, 셀 빈도, 승산비값 등을 이용하면 효과크기의 계산이 가능하며, 이를 종합하여 메타 분석의 결과를 도출할 수 있다. 예를 들어, 두 집단의 평균과 표준 편차가 제시된 경우에는 Table 1에서와 같이 효과크기를 계산할 수 있으며, 아래와 같은 다양한 경우에도 효과크기를 쉽게 계산할 수 있다.

t값과 표본 수를 알고 있는 경우의 효과크기 $= t\sqrt{\left(\frac{1}{n_1}\right)+\left(\frac{1}{n_2}\right)}$

2 × 2 분할표에서 셀 빈도가 (a, b, c, d)인 경우의 효과크기

$$= \frac{(ad-bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

결 론

의학 연구 결과를 보고하는 데 널리 사용되고 있는 P값은 통계적 유의성에 대한 이분법적인 판단만을 제공하며, 표본 수에 영향을 받는다는 한계를 가진다. 이에 비해 효과크기는 실제 차이를 파악하여 수치화 할 수 있다는 장점을 지닌다. 더욱이 여러 개별 연구 결과를 효과크기라는 표준화된 수치로 나타내면 이를 종합하여 메타분석에 이용할 수 있다는 장점도 있다. 그러나 현재와 같이 $P<0.05$를 금과옥조로 여기는 분위기에서는 $P>0.05$인 경우에도 효과 크기≠0임에도 불구하고 $P>0.05$라는 이유로 연구 결과가 사장되는 일이 많은 것이 사실이다. 이러한 학계의 관행에 효과크기의 사용을 과감히 주장하는 것도 필요하다고 생각된다. 아직까지는 임상 의학 분야에서는 효과크기의 사용이 드물지만 심리학, 행동과학 분야에서는 많은 논문들이 효과크기를 사용하여 결과를 기술하는 경우가 많다. 임상 의학 분야에서도 연구자들이 효과크기의 장점을 인식하여 의학 연구 분야에서 널리 이용되기를 기대한다.

REFERENCES

1. Wilkinson L. Statistical methods in psychology journals: guidelines and explanations. Am Psychol 1999;54:594-604.
2. Olejnik S, Algina J. Measures of effect size for comparative studies: applications, interpretations, and limitations. Contemp Educ Psychol 2000;25:241-86.
3. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and an alternative. J wildl Manage 2000;64:912-23.
4. McGraw KO, Wong SP. A common language effect size statistic. Psychol Bull 1992;111:361-5.
5. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New York: Lawrence Erlbaum Associates; 1988.