소아마취 : 2003; 7: 68～70
Korean J Ped Anesth 2003

ISSN 1226-6361

과제의 표본수는 연구자 마음대로 정해도 된다?

서울의대 마취과

안 원 식

임상 의학자로서 통계를 언급할 때 이해하기 어려운 부분 중에 하나가 표본수, 검정력, 제 2종 오류 등에 관한 문제라고 생각한다. 그래서, 이 종설에서는 지금까지 애기되는 표본수와 검정력에 대한 것을 알아보고자 하였다. 그런데, 표본수에 대한 학설은 크게 2가지 종류가 있다. 하나는 전통적으로 사용되어 왔고, 지금도 널리 이용되는 방법이다. 이 종설에서는 '전통적인 방법'이라고 언급하겠다. 또 다른 방법은, 몇몇 학자들이 주장하는 방법이면서, 간혹 논문에 등장하는 방법으로 이 종설에서는 '새로운 방법'이라고 언급하겠다. 전통적인 방법은 많은 책자에서 언급하고 있고, 대한마취과학회지에도 종설이[1] 실려 있기에 간략하게 개념만 언급하겠다.

표본수를 구하는 전통적인 방법

표본수를 결정하는 데에는 여러 연구 모형에 따라 세부적인 항목은 다를 수 있지만, 크게 4가지의 변수를 알거나 정해야 한다. 네 종류의 변수는, 제 1종 오류(α), 제 2종 오류(β), 산포도(분산,σ), 차이의 정도(d)이다. 이를 개념적 수식으로 나타내면,

표본수 = '제 1종 오류'×'제 2종 오류'×'산포도'/ '차이의 정도'

상기의 식에서 이해해야 할 사항은, '제 1종 오류', '제 2종 오류', '산포도'는 표본수와 비례 관계가 있어서 이들 항목이 커지게 되면, 연구의 표본수도 크게 해야 한다는 사항이다. 반면에, '차이의 정도'는 표본수와 반비례의 관계가 있어서 검출하고자 하는 차이가 작으면 표본수를 크게 해야 한다. 상기의 4가지의 항목 중에서 제 1종 오류에 대해서는 많은 논문에서 기술하고 있으나 다른 3가지에 대해서는 기술하고 있는 논문들이 드물다. 2003년에 개정된 대한마취과학회지 투고규정에서도 이들의 기술에 대해 규정하고 있지 않다.[2] 하지만 이 항목들은 논문의 결과를 전혀 다르게 도출시킬 수 있으므로 꼭 기술하여야 한다고 생각된다. 예를 들어서 생각해 보도록 하겠다. 어떤 연구자가, desflurane 2 MAC을 투여하였을 때, 1 MAC을 투여하였을 때보다 기도저항이 증가하였는가 여부를 알아보는 연구를 기획하였다. 이 연구에서 어떤 연구자는 1 mmHg/ml/sec 이상이 증가하면 증가하였다고 판정하고, 다른 연구자는 0.5 mmHg/ml/sec가 증가하였을 때 증가하였다고 판정하기로 연구계획을 세우고 연구를 진행하였는데, 연구결과가 0.7 mmHg/ml/sec가 나왔다고 한다면, 같은 연구결과를 제시하면서도 정반대의 연구결론을 제시할 것이다. 이러한 형태의 연구 결과가 많은 의학적 논란(controversy)의 원인 중에 하나가 되지 않을까 생각된다. 이러한 논란을 줄이기 위해 몇몇 생물통계학자들은, 두 군에서의 효과 차이를 공통의 표준편차로 나눈 값(effect size)를 기술해 주자는 의견이 있다.[3,4]

통계학 이해를 위한 논리학

표본수를 구하는 새로운 방법을 이해하기 위해서는 기본적인 논리학에 대한 이해를 하는 것이 필요하다. 우리가 중, 고교 수학 시간에 배운 논리학을 다시 생각해보자. 우리가 어떤 연구에서 통계 처리를 하면서 내리는 결론은 대개

'…의 연구결과, P값이 얼마이므로 가정(Ho)이 맞다(또는 틀리다). 그러므로, …라고 생각한다.'

로 이루어진다. 그런데, 이를 약간 자세히 살펴보면 심각한 오류가 있음을 알 수 있다. 우리가 원하는 것은,

'주어진 자료에 의하면, 가정(Ho)이 참일 확률'

이다. 그런데, 보통의 연구에서 구하는 것은,

'가정(Ho)이 참이면, 이러한 자료가 나올 확률'

이다. 이 두 가지는 명제는 전혀 다른 것이다. 약간 다른 관점에서 좀 더 자세히 살펴보면,

'만약 귀무가설이 맞다면, 이러한 자료(결과)가 일어나지 않는다. 그러나, 이러한 결과가 일어났다. 그러므로 귀무가설이 틀리다.'

상기의 말은 지극히 정당한 논리이고, 아리스토텔레스가 말한 modus tollens의 전형적인 예이다. 그런데, 이를 약간 변형하여,

'만약 귀무가설이 맞다면, 이러한 자료(결과)가 일어나지 않을 가능성이 많다. 그러나, 이러한 결과가 일어났다. 그러므로 귀무가설이 틀릴 가능성이 많다.'

는 논리를 사용한다면 언뜻 보기에 맞는 것처럼 보이지만 매우 심각한 오류가 존재한다. 즉, 논리에 확률이 추가되면서 이를 인식하지 못하면 오류가 발생한다. 구체적인 예를 들어 생각해보자.

'만약, 어떤 남자가 이철수라면, 그 남자는 국회의원이 아니다.
저기 오는 남자는 국회의원이다. 그러므로, 그 사람은 이철수가 아니다.'

상기 예는 modus tollens의 예이다. 그런데,

'만약, 어떤 남자가 한국인이라면, 이 남자는 국회의원이 아니다(가정이 틀렸음).
저기 오는 남자는 국회의원이다. 그러므로, 그 사람은 한국인이 아니다.'

라는 말을 하였다면, 가정이 틀렸기에 결과도 틀린 논리이다. 그러면, 이것에다가 확률을 부여해 보자.

'만약, 어떤 남자가 한국인이라면, 이 남자는 국회의원이 아닐 가능성이 높다.
저기 오는 사람은 국회의원이다. 그러므로, 그 사람은 한국인이 아닐 가능성이 높다.'

명제에 논리를 부여하여, 가정이 있을 법하게 되었다. 그러나 결론은 전혀 엉뚱한 것이 나왔다. 이러한 명제에 해당사항을 바꾸면 앞에서 기술한, 기존의 전통적인 방법에서 사용하는 다음과 같은 명제인 것이다.

'만약 귀무가설이 맞다면, 이러한 자료(결과)가 일어나지 않을 가능성이 많다. 그러나, 이러한 결과가 일어났다. 그러므로 귀무가설이 틀릴 가능성이 많다.'

다시 처음의 명제에 대해 살펴 보도록 하자.

'주어진 자료에 의하면, 가정(H_o)이 참일 확률'---- (1)
'가정(H_o)이 참이면, 이러한 자료가 나올 확률'---- (2)

(1)식과 (2)식을 기호로 나타내면,

$P\ (H_o|D)$ ---------- (1)
$P\ (D|H_o)$ ---------- (2)

우리가 원하는 것은 $P\ (H_o|D)$인데, 보통 연구에서 주어지는 것은, $P\ (D|H_o)$이다. 이 두 가지가 서로 다르다는 것은 Bayes's theorem을 언급하지 않더라도 직관적으로도 이해할 수 있을 것이다. 상기 논리에 대한 것은 참고 문헌을 숙독해 보면 보다 이해가 쉬울 것이다.[5-17] 또한, 이러한 논리에 대한 반박하는 논문도 있으니 참고하기 바란다.[18]

P값과 귀무가설

어떤 연구의 치료 효과(treatment effect)보다 계산되어서 나온 P값에 보다 큰 비중을 두고 있는 경우가 종종 있다. 하지만, P값은 차이의 정도를 기술하거나 임상적 의미의 중요성을 설명해 주지 못한다. P값은 연구 모델에서 자료의 수에 영향을 받게 된다. 매우 작은 차이를 보이는 모델이라도 자료의 양을 많이 취하게 되면 매우 유의한 P값을 얻게 되고, 비록 임상적

으로 매우 큰 차이가 존재하더라도 자료의 수가 적으면 P값이 유의하지 않게 나올 수 있다. P값은 확률에 대한 수학적 기술에 불과하다. 그러므로 치료 효과나 차이에 보다 많은 의미를 두는 것이 필요하다.[19]

상기 기술에 대한 이해를 돕기 위해 예를 하나 들어보자. isoflurane 흡입농도(예, 0 MAC에서 3 MAC까지 0.5 MAC 간격)와 cerebral blood flow (CBF)와의 선형관계의 정도를 분석하는 상관분석을 보자. 우리가 여기서 알아보고자 하는 것은 기울기가 0인가 여부이다. 귀무가설은,

'서로의 상관관계가 없어 기울기가 0이다.'

이고, 이것이 참이면 주어진 자료가 일어날 확률(주어진 자료에 의하면, 가정(H_o)이 참일 확률이 아님)을 구하게 된다. 이 때, 상기 문단에서 언급한대로, 자료의 수를 많이 뽑으면, P값이 적어지게 된다. 또한, 기울기가 0이 아닌 정도(다른 임상 연구에서는, 치료 효과나 두 군의 차이 등)에 대해서는 아무런 언급이 없이 단지 0이 아님 여부만을 구하게 된다. 그러므로, 제목에 있는 것과 같은 극단적인 표현을 사용하는 연구자도 있다. 그래서 새로운 방법이 등장하게 되었다.

표본수를 구하는 새로운 방법

표본수를 구하는 새로운 방법의 핵심은 '치료효과'의 선정에 있다. 전통적인 방법에서는 치료효과라는 개념이 없이 단지 치료효과의 있고 없음, 또는, 두 군간의 차이가 있고 없음으로 귀무가설과 대립 가설을 새우고 있다. 이러한 가설은, **P값과 귀무가설**에서 살펴본 바와 같이 표본수를 늘리면 모두 차이가 나는 연구 결과를 가지고 오며, 논리적으로도 오류가 있다. 그래서 치료 효과라는 개념을 두어, 예를 들어, 두 군간의 차이가 '매우 적은 차이(무시할만한 차이), 적은 차이, 중등도 차이, 많은 차이'로 구분하여 주어진 자료에 의해 각 가정이 나올 확률(귀무가설이 생길 확률이 아님)을 구하자는 것이다. 고혈압 치료제를 예로 들어 보자. 어떤 신약이 개발되었는데, 기존의 약보다 얼마나 효과가 좋은지를 알고 싶다고 치자. 전통적인 방법에서는 '기존의 약보다 효과가 좋다(또는 동일하다)'라는 결론이 나왔지만, 새로운 방법에서는, '기존의 약보다 효과가 무시할만한 차이(또는 적은 차이, 중등도 차이, 많은 차이)가 있다.'로 결론이 난다. 이 때 차이의 정도 (매우 적은 차이[무시할만한 차이], 적은 차이, 중등도 차이, 많은 차이)가 '치료 효과(effect size)'라는 것으로 각 연구 방법별로 설정되어 있고, 대개는 '효과 차이를 공통의 표준편차로 나눈 값'이 된다.

그러면, 이 새로운 방법으로는 표본수를 어떻게 정하게 될까? 새로운 방법에서도 전통적인 방법에서 사용하는 네가지 항목, 제 1종 오류(α), 제 2종 오류(β), 산포도(분산,σ), 차이의 정도(d)가 모두 필요한데, 이들에 대한 기술이 치료효과라는 것으로 변형되어, 산포도와 차이의 정도가 한 항목으로 합쳐져 있다. 새로운 방법도 상기 항목을 선정하면 간단한 표에 의해서 구해볼 수 있다.[3,4]

글을 맺으며

아직 전통적인 방법과 새로운 방법 간의 통계학적인 논쟁이 종료되지 않은 새로운 분야에 대한 설명을 해 보았다. 이 종설을 많은 의학자들이 완전히 이해할진 못할 것을 생각되지만, 기존의 전통적인 통계학적 방법에서 상당한 문제를 포함하고 있고, 이것이 의학의 많은 논란거리의 한가지 요인이 될 수도 있다는 인식을 하는 것만으로도 이 종설은 큰 성과를 거두었다고 생각된다.

참 고 문 헌

1. 김 호: 적절한 연구대상수의 산출. 대한마취과학회지 2002; 42: 1-10.
2. 대한마취과학회: 투고규정. 대한마취과학회지 2003; 45: 1권 부록.
3. Cohen J: Statistical power analysis for the behavioral sciences. 2nd ed. New Jersey, Lawrence Erlbaum 1988.
4. Murphy KR, Myors B: Statistical power analysis. New Jersey, Lawrence Erlbaum 1998.
5. Pollard P, Richardson J: On the probability of making type I errors. Psychological Bulletin 1987; 102: 159-63.
6. Cohen J: The earth is round (p＜.05). American Psychologist 1994; 49: 997-1003.
7. Valen L: Null hypotheses and prediction. Nature 1985;

314: 230.
8. Frick RW: Accepting the null hypothesis. Memory & Cognition 1995; 23: 132-8.
9. Serlin RC, Lapsley DK: Rationality in psychological research. The good-enough principle. American Psychologist 1985; 40: 73-83.
10. Chow SL: Significance test or effect size? Psychological Bulletin 1988; 103: 105-10.
11. Schmidt FL: What do data really mean? American Psychologist 1992; 47: 1173-81.
12. Murphy KR: If the null hypothesis is impossible, why test it? American Psychologist 1990; 45: 403-4.
13. Rouanet H: Bayesian methods for assessing importance of effects. Psychological Bulletin 1996; 119: 149-58.
14. Cohen J: Things I have learned (so far). American Psychologist 1990; 45: 1304-12.
15. Rozeboom WW: The fallacy of the null-hypothesis significance test. Psychological Bulletin 1960; 57: 416-28.
16. O'Quigley J, Baudoin CE: Null hypotheses and the misuse of statistics. Nature 1985; 316: 582.
17. Schervish MJ: P-Values: What they are and what they are not. American Statistician 1996; 50: 203-6.
18. Hagen RL: In praise of the null hypothesis statistical test. American Psychologist 1997; 52: 15-24.
19. Myles PS, Gin T: Statistical Methods for anaesthesia and intensive care. Woburn, Reed educational and professional publishing 2000, pp 131.