J Vet Clin 29(1) : 68-77 (2012)

Journal of Veterinary Clinics
www.ksvc.or.kr

의학연구에서 표본크기 계산

박선일[1] · 오태호*

강원대학교 수의과대학 및 동물의학종합연구소, *경북대학교 수의과대학

(게재승인: 2011년 12월 13일)

Sample Size Calculation in Medical Research

Son-Il Pak[1] and Tae-Ho Oh*

College of Veterinary Medicine and Institute of Veterinary Science, Kangwon National University, Chuncheon 200-701, Korea
**College of Veterinary Medicine, Kyungpook National University, Daegu 702-701, Korea*

Abstract : Whenever planning a study design or preparing a research proposal it is highly recommended that investigators decide the optimum sample size that is required to yield an outcome of interest with a predetermined level of precision. This is because that, all else being equal, if a study with less than the optimum sample size would not detect the significance of differences in reality, and similarly, if a study with more than the optimum sample size will be costly. For these reasons, the majority of peer reviewed biomedical journals assess the adequacy of sample size requirements. The calculated sample size is used as a target number of samples to be collected to provide an estimate of the parameter with the desired and predetermined level of accuracy, and the sample size is a major determinant of the probability of detecting diseased animals from the population. There is no single method of calculating sample size for any given study design. In this context, the purpose of this article is to provide a collection of formulas and examples for some typical situations likely to be encountered in veterinary clinical practice and to highlight the importance of performing prospective sample size calculations when planning a research. Specifically, this paper is concerned with the basic principle of sample size calculation, and considerations for methodological applications were illustrated for a given data set. Also included in this paper is factors influencing sample size calculations using a statistically valid techniques. Appropriate methods to consider these factors are presented.

Key words **:** prevalence, sample size, statistical power.

서 론

실험이나 조사를 계획할 때 연구자는 연구에 필요한 대상자 수 (표본크기)를 결정하고 이들을 어떻게 선발할 것인지에 일차적으로 관심을 갖는다. 표본크기(sample size)는 약물의 효과 평가, 질병 발생률 비교, 새로운 진단검사의 정확도 평가, 감염개체 검출, 수술법의 효과 비교, 질병 청정증명(disease freedom) 등 임상의학을 비롯한 다양한 연구 분야에서 모든 연구자가 항상 직면하는 문제다(2,4,11,22). 대부분의 실험연구에서는 모집단 전체를 대상으로 하지 않고 모집단의 부분집합으로 선발된 표본을 대상으로 수행되며, 실험으로 얻은 결과가 참인지 아니면 우연(chance)에 기인한 것인지를 판단하기 위해 통계적 추론을 사용한다. 실험결과가 우연에 기인한 것이 아니라는 것을 확신하기 위해서는 무작위(randomization), 이중맹검법(double blinding) 등과 같은 실험계획의 편견(bias)을 제거하기 위한 노력과(12) 표본크기 등을 포함한 연구계획을 신중하게 작성해야 한다.

표본크기가 부족하면 아무리 엄격하게 실험을 수행한다고 하더라도 처리효과를 검출하는데 실패하거나 실험으로 얻는 결과의 정확도를 신뢰할 수 없게 된다. 반면에 적정 수준이상으로 표본크기가 크면 비용이 증가하고 임상적으로 의미가 없는 미미한 효과를 유의하게 판정하는 오류를 범할 가능성이 증가한다. 적정 수준의 표본크기는 연구결과의 유용성과 연구 목적 달성 여부에 영향을 미치고, 실험 개체수의 증가는 실험동물의 윤리적인 측면에 위배되기 때문에(11) 연구 계획을 수립하는 단계에서 신중하게 검토해야 한다. 이를테면 복통을 완화하는 두 약물의 효과를 비교하는 연구에서 실험에 사용한 개체수가 부족하면 실제로 모집단에 처리효과가 있음에도 불구하고 이러한 효과의 차이를 표본에서 검출하지 못하게 되어 실험 자체의 유용성이 상실된다. 마찬가지로 어느 우군에 감염된 개체가 없다는 결과에 근거하여 청정농장으로 인증을 받는 상황에서 표본크기가 너무 적으면 모집단에서 감염이 높은 수준으로 존재할지라도 감염개체를

[1]Corresponding author.
E-mail : paksi@kangwon.ac.kr

검출하지 못하게 되며, 반면에 표본크기가 적정수준 이상으로 매우 크면 실제로 감염된 개체가 매우 적을지라도 이를 검출하는 능력은 증가하지만 이에 따른 검진비용이 증가하여 조사의 효율성이 저하된다.

연구목적을 달성하는데 필요한 최적의 표본크기는 연구계획 단계에서 결정하는 것이 일반적이다(16). 그러나 연구에 소요되는 시간과 비용, 이용 가능한 자원의 한계, 발생률이 매우 낮은 질병인 경우 연구 대상 개체수 선발의 어려움 등으로 연구를 수행한 후 분석시점에서 연구에 사용한 표본크기에 대하여 검정력 분석(statistical power analysis)을 수행하기도 한다. 이러한 사후(post-hoc) 분석은 두 군간 실제로 차이가 있지만 연구에서 이러한 차이를 검출하지 못하였을 때 그 원인의 하나로 표본크기의 부족에 기인하였는지를 검토하는 목적으로 수행한다. 표본크기는 연구계획의 종류, 자료의 형태, 연구목적, 표본추출방법 등에 따라 다양한 방법으로 계산된다. 국제 의학 학술지 편집인 위원회(International Committee of Medical Journal Editors, ICMJE)에서는 투고 논문에 대한 심사항목에서 표본크기 산출근거를 기술하도록 권고하고 있으며(18), 이러한 항목은 임상시험이나 생명공학 분야의 논문에 대한 심사에서 필수적으로 요구하고 있다(3,7,9,13,19,24). 본 연구에서는 표본크기 계산에 필요한 통계적 원리와 수의학에서 흔히 접하는 연구 상황에 대한 예시를 통해 표본크기의 중요성을 설명한다.

결 론

연구 상황 1 : 개에서 중증의 안과질환 X에 대한 외과적으로 수술법 A와 B의 치료효과를 비교하는 연구에서 성공률이 각각 A = 45% (n_1 = 40두)와 B = 40% (n_2 = 20두)로 나타나 저자는 분석결과 두 군간 성공률에 차이가 없다(p = 0.7125)는 결론을 내렸다고 하자. 여기에서 두 수술법 간 성공률에 차이가 없다는 저자의 결론이 어느 정도 정확한지에 대해서 독자는 알 수 없다. 즉 두 집단의 표본크기 40두와 20두를 대상으로 수행한 연구결과 성공률에서 5%의 차이를 검출하기에 충분한 능력을 갖추고 있는지와 5%의 차이가 임상적으로 의미가 있는지를 묻는 것이다. 이러한 질문은 연구에 사용한 표본크기의 적절성으로 평가할 수 있다.

연구 상황 2 : 심장질환 X로 진단 받은 개의 내과적 처치방법으로 기존의 약제 A와 새로운 약제 B의 치료효과를 평가하였다. 두 약제의 성공률이 각각 A = 80% (n_1 = 50두)와 B = 70% (n_1 = 40두)이고 분석결과 성공률에 차이가 없다(p = 0.2727)는 연구논문을 투고하였다. 심장질환 A에 대하여 약물로 치료할 때 성공률은 평균 80%로 알려져 있고 성공률에서 10% 이상의 변화가 있을 때 임상적으로 가치가 있다고 가정할 때 이 논문은 출판될 수 있을까? 본 연구의 경우 성공률에서 10% 이상의 차이를 검출하는 것을 80% 확신 (검정력)하기 위해서는 유의수준 5%에서 실험군 당 219두 (총 438두)가 필요하다. 따라서 90두의 표본을 사용한 연구에서 두 군간 성공률이 동일하다는 저자의 결론에는 심각한 문제가 있음을 알 수 있다.

표본크기 계산에서 고려해야 할 요인

전술한 두 예에서 보듯이 표본크기와 검정력은 연구결과의 정확도에 결정적인 영향을 미치기 때문에 반드시 연구계획 단계에서 적절한 표본크기를 산정하는 것이 중요하다. 표본크기를 계산할 때에는 많은 요인에 대한 추정치를 필요로 하고 경우에 따라서는 다소 연구자의 주관적인 추정치를 필요로 한다. 중요한 것은 통계적 혹은 임상적인 타당성이 없이 임의적인 가정에 근거하여 표본크기를 결정하는 것이 아니라 이용 가능한 최대한의 정보에 근거하여 계산해야 한다. 표본크기는 실험에서 얻는 측정 자료의 정밀도(precision) 혹은 변동성(variability), 유효크기(effect size), 유의수준(significance level), 검정력(statistical power)에 영향을 받는다. 전자의 두 요인은 실험내용에 따라 다르게 설정되며 후자의 두 요인은 흔히 고정된 값을 사용한다.

(1) 측정치의 정밀도와 변동성

정밀도(precision, d)는 2회 이상 측정값이 서로 동일한 결과를 보이는 정도를 나타내는 지표다. 표본크기 계산과 관련하여 정밀도는 모집단의 참값이 위치할 것으로 기대되는 신뢰구간(confidence interval)으로 흔히 ± 5%, ± 10% 등으로 표현한다. 예를 들어 어느 연구자가 병원에 내원하는 환자의 80%가 백신접종을 완료하였음을 95% 신뢰수준에서 ± 5% 의 정밀도로 추정하는 표본크기를 사용하여 표본조사를 수행하였다면 이 조사에서 백신접종율의 신뢰구간은 75-85%로 추정된다는 결론을 얻을 수 있으며, 여기에서 보듯이 신뢰구간의 폭(width)은 정밀도의 2배 (2d)이다(20,25).

신뢰구간은 표본크기(n)의 역수와 관련이 있기 때문에 표본크기가 증가하면 모집단의 참값에 더 근사하는 (정밀도가 높은) 추정치를 얻는다. 표본으로부터 계산된 모집단 평균 추정치를 점추정치(point estimate)라고 하며 신뢰구간은 모집단의 모수가 포함될 구간을 표본으로부터 확률적으로 추정한 구간으로, 95% 신뢰수준에서 신뢰구간은 다음과 같다(1,2,10).

$$\text{신뢰구간: 점추정치} \pm 1.96 \times SE,\ [SE = SD / \sqrt{\mathrm{n}}\,]$$

모집단에서 측정하고자 하는 속성의 분포를 변동성(degree of variability) 혹은 산포성(spread)이라고 한다. 모집단에서 연구자가 관심을 두고 있는 속성이 이질성이 높을수록 주어진 정밀도에서 더 많은 표본크기를 필요로 한다. 예컨대 특정 질병으로 진단받은 120두의 개를 대상으로 blood urea nitrogen (BUN, mg/dl) 농도를 측정한 결과 평균 34, 표준편차 3.8을 얻었다면 표준오차(standard error, SE)는 $3.8/\sqrt{120} = 0.35$ 로 계산된다. 이는 모집단으로부터 동일한 크기의 표본을 반복하여 선발할 때 평균이 34이고, 표본평균의 약 95% (1.96SE)는 [33.31 – 34.69] 사이에 있을 것으로 추정할 수 있음을 의미한다. 만일 표본크기 60, 30, 10에서 전술한 추

정치를 얻었다고 가정하면 표준오차는 각각 $3.8/\sqrt{60}=0.49$, $3.8/\sqrt{30}=0.69$, $3.8/\sqrt{10}=1.20$ 이 되어 표본크기가 감소할수록 표준오차가 증가한다. 따라서 자료의 변동성이 작을수록 신뢰구간이 좁아지며 결과적으로 모집단의 특성 (평균, 비율 등)을 보다 정밀하게 추정할 수 있게 된다. 표본 변동성이 낮으면 모집단을 대표할 가능성이 높아지므로 유의한 차이를 검출하는데 요구되는 표본크기가 감소한다.

정밀도 결정시 고려사항: 유병률 추정에 필요한 표본크기를 계산할 때 정밀도를 어느 수준으로 설정해야 하는지에 대한 가이드라인은 없고 흔히 연구자가 설정한다. 일반적으로 유병률이 10-90% 범위일 경우 5%의 정밀도 (신뢰구간으로 표현하면 ±10%)를 사용하며 이 범위를 벗어난 경우 5%의 정밀도를 설정하면 문제를 초래한다. 예를 들어 유병률이 1%로 매우 드문 질병(rare disease)에 대하여 5%의 정밀도를 설정하면 추정 유병률의 95% 신뢰구간의 상한값(upper limit)이 1을 초과하거나 하한값(lower limit)이 음수를 보일 수 있다. 따라서 유병률(p)이 10% 이하일 때 정밀도(d)는 $d=0.5p$, 90% 이상인 경우 $d=0.5(1-p)$ 의 관계를 대안으로 사용할 수 있다(25). 예를 들어 $p=0.04$ 일 때 $d=0.02$, $p=0.98$일 때 $d=0.01$이 된다. 필요하다면 이 보다 더 작은 수준의 정밀도를 지정할 수 있다.

(2) 유의수준

예를 들어 두 약물의 치료효과 (완치율)를 비교하는 연구에서 연구자는 두 처리 간 완치율에 차이가 없다는 가설 (귀무가설)과 두 처리 간 완치율에 차이가 있다는 가설 (대립가설)을 고려할 수 있다. 치료율에 차이가 없다는 가설을 수용하거나 (귀무가설 수용) 치료율에 차이가 있다는 가설을 수용할 때 (귀무가설 기각, 대립가설 수용) 두 종류의 오류를 범하게 된다(1,2,10). 제1형 오류(type I error)는 참인 귀무가설을 잘못하여 기각하는 오류로 α오류라고 한다(Fig 1). 예를 들어 사료첨가제가 증체율에 미치는 효과를 평가하는 연구에서 제1형 오류는 실제로 두 집단 간 증체율에 차이가 없음에도 불구하고 차이가 있는 것으로 판정하는 경우이다. 유의수준을 5%로 설정한다는 것은 유의한 결과를 얻었을 때 이러한 결과를 관찰할 기회를 5% 이하로 설정한다는 것을 의미한다. 이는 모집단에 실제로 존재하지 않는 효과를 표본을 대상으로 하는 연구에서 잘못 검출할 확률로 5%는 수용한다는 것을 의미한다(false positive). 유의수준을 낮게 설정하는 이유는 가양성 결과를 검출할 가능성을 줄이고자 하는 연구자의 의지이며 유의수준을 낮게 조정할수록 표본크기는 증가하여 역(inverse)의 관계가 있다.

한편 $1-\alpha$ 를 신뢰수준(confidence level)이라고 한다. 모집단에서 표본을 반복하여 선발할 때 표본에서 얻은 추정치는 중심극한정리(Central Limit Theorem)에 의해 정규분포를 따르며 일부 추정치는 참값 보다 크거나 작은 값을 보일 수 있으나 표본평균은 모집단 평균에 근사한다. 정규분포에서 표본 추정치의 95%는 모집단 참값의 2SD (standard deviation)

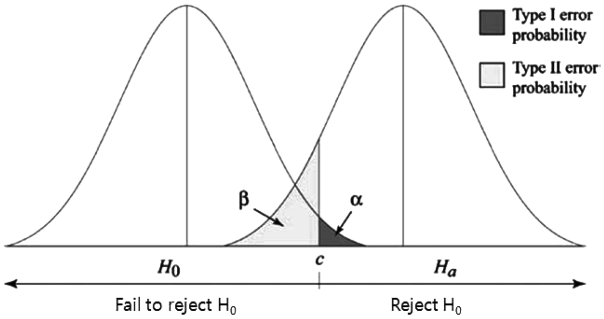

Fig 1. Type I (α) and II (β) error under the null (H_0) and alternative (H_a) hypotheses testing. c = critical value.

이내에 위치한다. 즉 95% 신뢰수준이라 함은 100개의 표본 중 95개는 지정한 정밀도에서 모집단의 참값을 포함한다는 것을 의미한다(1). 표본조사 결과는 모집단의 참값을 포함하지 않을 수도 있기 때문에 신뢰수준을 흔히 위험수준(risk level)이라고도 하며 만일 이러한 위험수준을 줄이고자 한다면 신뢰수준을 99%로 높이고, 반대로 신뢰수준을 90%로 낮추면 위험수준은 증가한다.

제2형 오류(type II error)는 거짓인 귀무가설을 기각하지 못하는 오류로 β 오류라고 한다. 이를테면 두 군간 실제로 차이가 있지만 이를 기각하지 못하고 차이가 있다고 잘못 판정하거나, 두 집단 간 증체율의 차이가 실제로 존재함에도 불구하고 이러한 유의한 차이를 검출하는데 실패하는 경우다. α와 β 오류는 반대방향으로 작용하는 특성이 있어 α 오류를 5%에서 10%로 증가시키면 β 오류는 감소한다(Fig 1). 검정력 계산(power calculation)이란 두 종류의 오류를 회피하는데 필요한 표본크기 (엄밀히 말하면 지정한 유의수준과 표본크기에서 제2형 오류의 가능성을 회피하면서 두 군간 차이를 검출할 확률)를 계산하는 것이라 할 수 있다(17). 유의수준과 마찬가지로 β 오류와 표본크기는 역(inverse)의 관계가 있다.

(3) 유효크기

연구자가 검출하기를 원하는 혹은 검출 가능한 효과의 크기(magnitude of effect)를 유효크기(effect size)라 한다(14). 예를 들어 혈압을 조절하는 약물의 효과를 평가하는 연구에서 유효크기를 10 mmHg의 감소로 설정한다는 것은 이 정도의 변화량은 임상적으로 의미가 있다고 판단하는 연구자의 의지이다. 다른 예로 특정 종양 환자의 1년 생존율이 50%로 알려져 있다고 할 때 새로 개발된 약제의 1년 생존율이 70%로 높아진다면 20%의 변화량 (유효크기)은 임상적으로 매우 의미가 있는 것으로 이 값을 유효크기로 지정할 수 있다. 유효크기는 상황에 따라 다를 수 있다. 예를 들어 어느 감염증에 대한 두 종류 항생제를 이용한 치료효과에 있어 70%와 80%의 치료율을 비교하는 연구에서 이러한 10%의 차이는 임상적으로 큰 의미를 갖지 못할 수 있지만, 반면에 치명적인 질병인 경우에는 1%나 2%의 근소한 차이도 중요한 의미가 있을 수도 있다.

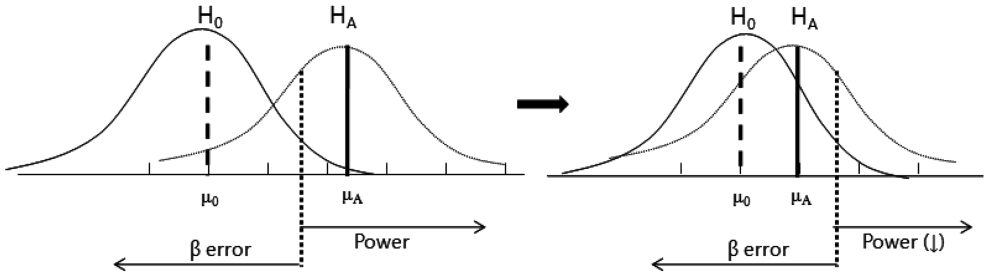

Fig 2. Power change when mean (μ_A) of the alternative hypothesis closes to the mean (μ_0) of null hypothesis.

유효크기는 기존의 연구결과, 예비실험, 임상적인 중요성 등을 고려하여 설정한다. 어느 정도의 차이가 검출할만한 의미가 있는 것인지에 대한 기준은 없지만 첫째, 환자의 입장에서 이 정도 유효크기 (예: 혈압감소)를 얻기 위하여 치료비용을 지불할 필요가 있는가 (비용 대비 치료효과) 둘째, 임상가의 입장에서 의미를 부여하기 어려운 유효크기의 범위는 어느 정도인가 (임상적 중요성) 셋째, 어느 정도의 유효크기를 기대하고 있는가 (연구자의 기대) 넷째, 15%의 혈압감소가 환자에게 중요한 의미를 갖는가 (환자의 치료 및 관리) 등과 같은 질문을 통하여 중요한 단서를 얻을 수 있다. 유효크기가 클수록 이러한 차이를 검출하기가 용이하므로 β 오류가 감소하여 검정력이 증가한다(Fig 2). 대립가설의 평균이 귀무가설의 평균과 근사한 경우 즉 유효크기가 작으면 β 오류가 증가하므로 검정력은 유의수준에 근사해질 정도로 감소한다. 이는 귀무가설과 거의 동일한 대립가설을 수용할 확률은 귀무가설을 잘못 기각할 확률 즉 유의수준과 같아지기 때문이다.

요약하면 연구자가 검출하기를 원하는 차이 즉 유효크기가 작을수록 이러한 차이를 검출하는 것이 쉽지 않고 매우 정밀한 추정치를 필요로 하기 때문에 표본크기는 증가한다. 동일한 표본크기에서 유효크기가 증가할수록 통계적 유의한 결과를 얻을 기회가 높아지므로 두 군간 유효크기가 임상적으로 중요하지 않은 상황에서 단순히 통계적으로 유의한 결과를 얻기 위하여 표본크기를 증가시키는 것은 가치가 없다. 따라서 연구를 계획할 때 최소 유효크기에 대하여 문헌고찰 등을 통하여 타당성을 제시해야 한다.

(4) 검정력

흔히 검정력을 실험에 서용한 개체 수인 표본크기와 동일한 의미로 사용하고 있다. 이를테면 "본 연구에서 잠재적인 혼란변수(confounder)를 보정하는데 검정력이 충분하지 못하였다"고 기술하는 것은 "가능한 다양한 효과를 설명하는데 개체수가 충분하지 못하였다"로 해석된다. 다른 예로 "본 연구는 유의수준 5%에서 상대위험도(relative risk) 1.1을 검출하기 위하여 80%의 검정력을 사용하였다"는 "효과를 검출하는데 충분한 개체수를 대상으로 하였다" 의미로 이해할 수 있다. 이러한 해석이 완전히 잘못된 것은 아니지만 검정력에 대한 개념을 충분히 이해하고 있어야 표본크기를 계산할 때 활용할 수 있다.

전술하였듯이 β 오류는 귀무가설이 거짓일 때 이를 기각하지 못하는 확률이므로 거짓 귀무가설을 올바르게 기각할 확률은 $1-\beta$ 가 되며 이는 두 실험군 간 실제로 차이가 있을 때 이러한 차이를 올바르게 검출할 수 있는 확률을 의미하고 이를 검정력이라고 한다. 다시말해 검정력은 표본검사에 근거한 가설검정을 통하여 모집단에 존재하는 진정한 효과를 검출하는 능력 (진양성, true positive)으로 틀린 귀무가설을 올바르게 기각할 확률이다. 가설검정에서 귀무가설은 "차이가 없다"고 설정하기 때문에 귀무가설을 기각한다는 것은 차이가 있다는 것을 의미한다. 유의수준(α)은 귀무가설을 잘못 수용할 확률 즉 효과가 있는 것으로 잘못 수용할 확률 (가양성, false positive)로 모집단에서 효과가 없는 상황에 대해서만 관심을 갖는다는 점이다. 따라서 가장 이상적인 상황은 충분히 높은 검정력과 충분히 낮은 유의수준이 보장되는 표본크기를 사용하여 실험하면 가음성(false negative)과 가양성 결론에 이를 가능성을 최소화하게 되어 두 군간 차이가 없다는 결론을 내릴 때 연구자가 실제로 차이가 없다는 것을 높은 신뢰도로 확신할 수 있는 것이다. 예를 들어 두 치료법의 효과를 비교하는 연구에서 검정력이 90%라는 것은 실제로 두 치료법 간의 효과에 차이가 있을 때 연구에서 이러한 치료효과를 검출할 가능성이 적어도 90%는 유지된다는 것을 의미한다. 즉 연구자가 주장하고자 하는 대립가설 즉 치료효과에 차이가 있다는 주장이 참일 때 통계적 검정에서도 이를 뒷받침해주는 능력의 정도가 된다. Fig 3은 두 군간 비율의 차이에 대한 가상의 분포로 유의수준 5%에서 두 군간 비율

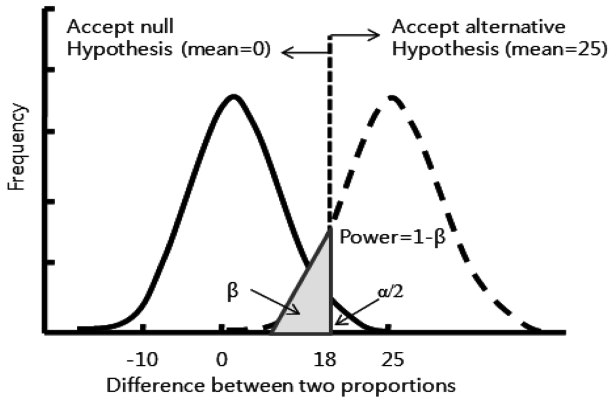

Fig 3. Relationship between significance level, β error and statistical power.

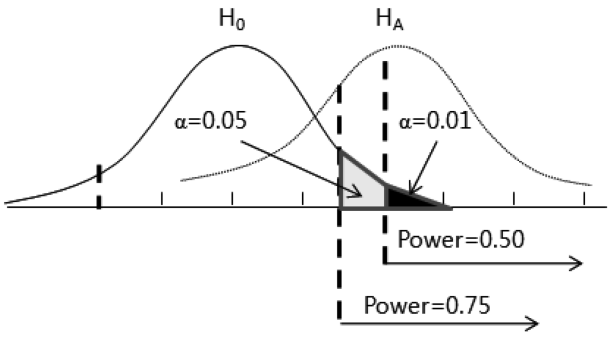

Fig 4. Relationship between significance level and power at equal sample size.

의 차이에 대한 상한 임계값이 18%인 상황을 예시한 것이다. 이 경우 귀무가설은 두 군간 차이가 없다 (즉 차이 = 0)는 것이고 대립가설은 두 군간 차이가 25%라는 것이다. 두 비율의 차이 25%를 검출하는 검정력은 표본 추정치가 적어도 상한 임계값 이상에 위치할 확률을 추정하는 것과 동일하므로 예컨대 대립가설하에서 추정치의 10%가 임계값 이하라면 즉 β = 10%라고 하면 검정력은 90%가 된다.

검정력 선택: 검정력을 최대화하는 일반적인 방법은 α를 0.1, 0.05, 0.01 등으로 고정시킨 상태에서 β를 최소화시킴으로써 검정력을 최대화하는 방법을 사용한다. α오류로 흔히 0.05를 사용하지만 연구 목적이나 상황에 따라 다양한 값을 지정할 수 있다. 예를 들어 연구자가 가음성(false negative) 결과를 회피하는데 관심을 더 두는 경우에는 0.05 대신에 0.1 (유의수준 증가, 완화된 기준)로 설정하고, 반대로 가양성(false positive) 결과를 회피하는데 관심을 더 두는 경우에는 0.05 대신에 0.01 (유의수준 감소, 보수적인 기준)로 설정한다. 유의성을 판단할 때 보수적인 기준을 사용할수록 우연에 의한 결과를 사실인 것으로 잘못 해석할 위험을 줄일 수는 있지만 매우 작은 유의확률을 얻기 위해서는 표본크기가 매우 커야 한다. 요약하면 동일한 표본크기에서 유의수준을 증가시키면 검정력은 증가하지만 반대로 가양성 결과를 얻을 확률이 증가할 수 있음을 주의해야 한다(Fig 4).

표본크기와의 관계: 표본크기는 검정력과 직접적인 관계가 있으며 표본크기가 클수록 검정력은 증가한다. 반면에 표본크기가 작으면 두 군간 실제로 차이가 있음에도 이러한 차이를 발견할 기회는 감소하고 그 결과 유의확률은 커지고 신뢰구간이 넓어져 두 군간 차이가 없다는 잘못된 판단을 범하게 된다. 표본크기가 충분하지 못해 유의하지 못한 결과를 얻는다면 비용과 자원의 낭비이기 때문에 실험계획 단계에서 검정력을 달성하는 표본크기를 계산하여 실험을 진행하는 것이 바람직하며 일반적으로 의학연구에서는 80%의 검정력을 유지할 것을 권고하고 있다(21). 이는 모집단에 실제로 존재하는 효과를 표본을 대상으로 하는 연구에서 검출하지 못할 확률로 20%를 수용한다는 것을 의미한다 (가음성). 효과를

Table 1. Values of Z_α and Z_β for sample size calculation

α error	Z_α		β error	$1-\beta$	Z_β (1-tailed)
	1-tailed	2-tailed			
0.01	2.326	2.576	0.01	0.99	2.326
0.05	1.645	1.960	0.05	0.95	1.645
0.10	1.282	1.645	0.10	0.90	1.282
0.20	0.842	1.282	0.20	0.80	0.840

검출하지 못할 확률을 줄이기 위하여 검정력을 90%로 증가시키면 표본크기는 증가한다. 예컨대 혈압을 조절하는 두 종류 약물의 효과를 비교하는 연구에서 가설검정 결과 유의하지 않다는 결론은 얻은 경우 이러한 결과가 반드시 두 군간 차이가 존재하지 않는다는 것을 의미하는 것이 아니고 차이를 검출할만한 검정력이 부족하여 유의하지 않은 결과가 나타났을 가능성을 고려할 필요가 있다. 이는 "증거의 없음이 반드시 없음의 증거가 되는 것은 아니다"는 것을 의미하며 (absence of evidence is not evidence of absence) 두 군간 차이가 없다는 결과에 대해서는 검정력이 적정 수준에 미치지 못하여 진정한 차이를 올바르게 검출하지 못하였을 가능성을 고려하는 것이다(28). 1994-2004년 기간 동안 소동물 마취분야에서 처리효과가 없다고 결론을 내린 46편의 논문 중 22편을 분석한 결과 20%의 처리효과 (결과변수의 증가 혹은 감소)를 검출하기에 충분한 검정력을 유지하는 논문은 18%에 불과한 것으로 조사되었다(16). 이러한 결과는 연구에서 유의한 차이가 없다는 결론을 내리기 이전에 연구방법이나 자료의 변동성 등을 포함하여 결과에 영향을 미칠 수 있는 다양한 요인을 검토할 필요가 있음을 시사한다. 검정력이 매우 높은 연구는 차이를 검출할 기회는 높이지만 매우 큰 표본크기를 필요로 하기 때문에 일반적으로 유의수준은 5%, 검정력은 최소 80%를 사용한다.

표본크기 계산에서 흔히 사용하는 유의수준과 검정력에 대한 표준정규분포의 Z_α와 Z_β를 요약하면 Table 1과 같다. 여기에서 Z_α는 귀무가설하에서 표준정규분포 곡선에서 유의수준 α (예, 0.1, 0.05, 0.01 등)에 해당하는 Z값이고, Z_β는 대립가설하에서 표준정규분포 곡선에서 β오류에 해당하는 Z값이다($\beta < 0.5$일 경우 Z_β는 음수).

검정력 곡선: 검정력은 유효크기, 유의수준 및 표본크기에 좌우되므로 검정력은 이들의 함수로 표현할 수 있다. 즉 주어진 통계량에서 다양한 대립가설의 평균 μ_1에 대한 검정력을 표본크기 (혹은 유효크기)의 관계로 나타낸 그래프를 검정력 곡선(power curve)이라 한다. 예를 들어 심장질환으로 진단받은 개의 경우 혈청 CK (creatine kinase, U/L) 농도가 낮다고 알려져 있는데 이를 확인하기 위하여 100두의 심장병 개를 대상으로 CK 농도를 측정하여 평균(μ_A) 115와 표준편차 25를 얻었다고 하자. 일반모집단에서 평균 농도(μ_0)는 120이고 표준편차가 25라고 할 때 검정력은 약 63.9%가 되며 이 자료에 대한 검정력 곡선은 Fig 5와 같다.

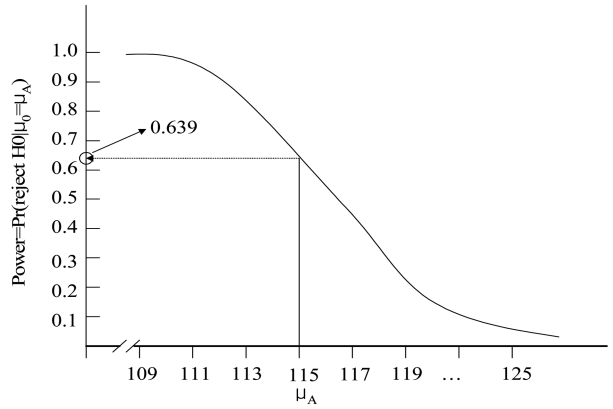

Fig 5. Power curve for creatine kinase (CK, U/L) for dogs with heart disease 'x'. μ_A represents mean CK level under the alternative hypothesis.

(5) 기대유병률

유병률 (모집단에서 감염된 개체의 비율) 추정에 필요한 표본크기를 계산할 때 기대유병률(expected prevalence, design prevalence)을 가정해야 하는데 이는 표본조사에서 검출할 수 있는 가장 낮은 수준의 유병률을 의미하며 만일 질병이 지정된 수준 이하로 존재한다면 조사에서 이를 검출할 수 없다. 기대유병률은 조사대상 질병의 역학적 특성에 대한 문헌고찰에 근거하여 연구자가 지정한다. 이를테면 조류인플루엔자와 같이 전염성이 매우 높고 전파속도가 매우 빠른 질병에 대하여 감염된 개체를 검출하는 경우 모집단에 감염이 존재한다면 감염된 개체의 비율이 높을 것이므로 기대유병률을 높게 설정하고, 요네병, 결핵 등과 같이 전염성이 낮거나 유전 질환과 같이 드문 질병이라면 기대유병률을 낮게 설정한다. 유병률에 대한 정보를 모른다면 최악의 시나리오 (모집단의 이질성)를 가정하여 계산한다. 이항분포인 비율의 분산은 $var = p(1-p)$이므로 $p = 0.5$일 때 최대값을 보이므로 표본크기는 최대가 된다. 예를 들어 유병률에 대한 정보로 30-40%를 얻었다면 40%, 70-80%를 얻었다면 70%를 적용하여 표본크기를 최대화시키는 것이 좋다. Fig 6에서 보듯이 동일한 유병률에서 정밀도를 높이면 표본크기는 증가하고, 신뢰수준을 높이면 표본크기는 증가한다. 예를 들어 50%의 기대 유병률을 95% 신뢰수준에서 ∓10%의 정밀도로 추정한다면 표본크기는 약 96두로 계산되지만 ∓5의 정밀도로 추정할 경우 표본크기는 약 384두로 증가한다.

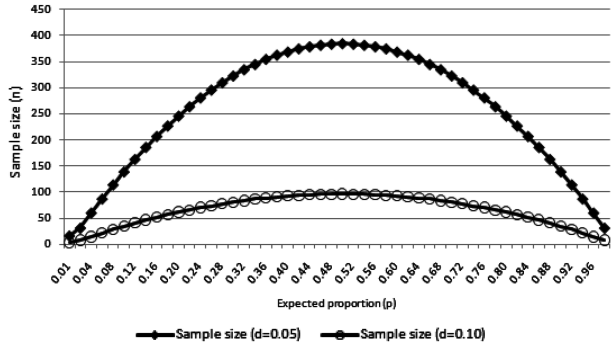

Fig 6. Relationship between sample size (n) and prevalence (p), assuming a confidence level of 95%, precision (d) 5%, and simple random sampling.

(6) Design effect

일반적으로 표본크기를 계산할 때 모집단의 구성원 즉 표본추출 단위(sampling unit)가 상호 독립적이라는 가정을 전제로 한다. 예를 들어 집락추출(cluster sampling)에서는 결과(비율)가 집락의 수준과 연관되어 있어 독립성 가정을 위반하고 이 경우 분산이 과소평가되기 때문에 분산을 팽창시키기 위하여 집락 내 상관계수(intraclass correlation coefficient, ρ)를 이용하여 분산팽창보정계수(design effect, variance inflation correction factor)로 표본크기를 보정해주어야 한다(23,27). 즉 design effect (DE)는 단순무작위추출을 사용할 때 기대되는 분산에 대한 집락추출 분산의 비(ratio)로 집락크기를 m이라 하면 $DE = 1 + \rho(m-1)$ 로 계산된다. 여기에서 ρ는 집락 간과 비교할 때 집락 내에서 표본추출 단위의 동질성(homogeneity)을 나타내는 지표로 총 변동 중 집락 간 변동으로 설명되는 비율이다(23). ρ는 문헌고찰이나 예비실험 등을 통하여 추정하며 이 값이 클수록 의존성이 매우 높아 결과적으로 DE가 증가한다. 집락추출법에 필요한 표본크기는 단순무작위추출에서 독립성을 가정하여 계산된 표본크기에 DE를 곱하여 계산한다. 예를 들어 DE = 2일 경우 단순무작위추출에서 계산된 표본크기에 2배를 적용해야 동일한 정밀도를 달성할 수 있다는 것을 의미한다.

(7) 단측검정과 양측검정

연구자는 실험군에서의 치료율이 대조군에서의 치료율과 동일한지에 관심을 갖기도 하지만 증가 혹은 감소여부를 동시에 관심을 가질 수도 있다. 전자를 단측검정(one-sided test, one-tailed test), 후자를 양측검정(two-sided test, two-tailed test)이라고 한다. 검정의 형태를 결정해야 하는 이유는 연구의 목적과 표본크기 계산에서 유의수준과 직접적인 관련이 있기 때문이다. 예를 들어 개에서 췌장염과 체중에 대한 연구에서 연구의 목적이 정상 개체와 비교할 때 췌장염 환자가 과체중(obesity)과 관련이 있는지를 평가하는 것이라면 단측검정에 해당하고, 연구의 목적이 췌장염과 체중 (과체중과 저체중) 간의 연관성이라면 양측검정에 해당한다. 양측검정에 비하여 단측검정에서 표본크기는 증가하며, 표본크기를 고정시킬 때 양측검정에 비하여 단측검정에서 검정력이 증가하는데 그 이유는 유의수준이 증가하면 β는 감소하므로 결과적으로 검정력 $1-\beta$는 증가하기 때문이다.

(8) 기타 요인

특히 전향적 연구(prospective study)에서는 연구기간 동안 연구대상 개체가 탈락하는 경우가 많은데 특히 생존분석 자료에서 흔히 볼 수 있다. 생존분석에서 중도탈락 관찰치(censoring)는 연구대상자를 추적-관찰함에 있어 계획된 연구

Table 2. Factors that affect sample size calculation

Factor	Magnitude	Impact on detecting effect	Required sample size
Variability	Small	Easy to detect	Small
	Large	Not easy to detect	Large
α error	Small	Not easy to detect	Large
	Large	Easy to detect	Small
Effect size	Small	Not easy to detect	Large
	Large	Easy to detect	Small
Power ($1-\beta$)	Low	Not easy to detect	Small
	High	Easy to detect	Large
Precision	High		Large, compared to low precision
1-tailed test			Large, compared to 2-tailed test

기간 중에 생존여부를 완전하게 관찰하지 못하는 것이다. 연구계획을 수립하는 과정에서 표본크기를 완벽하게 결정하였다고 하더라도 연구 대상자가 많이 탈락한다면 연구결과의 신뢰도를 훼손한다. 따라서 표본크기 결정단계에서 연구 대상자가 처리(intervention)나 결과와 무관하게 폐사 등의 이유로 탈락할 것으로 예상되는 경우 이를 보정한 표본크기를 계산해야 한다. 이를테면 탈락율이 10%로 예상되는 경우 계산된 표본크기를 0.9로 나누어 최종 표본크기를 계산한다.

이상의 내용을 요약하면 표본크기는 변동성이 증가할수록, 유의수준이 감소할수록, 유효크기가 작을수록, 검정력이 증가할수록 증가한다(Table 2). 유의수준을 증가시키면 결과적으로 검정력이 증가하지만 반대로 가양성 결과를 얻을 확률이 증가한다. 가장 이상적인 경우는 표본크기가 무조건 커야 좋은 것이 아니고 두 군간 차이가 있을 때 이러한 차이를 올바르게 검출할 수 있는 수준의 표본크기를 계산하는 것이다. 표본크기를 계산하기 위해서는 전술한 항목에 대한 다양한 정보를 필요로 하는데 이들 항목에 대하여 신뢰할만한 자료를 확보하는 것이 쉽지 않지만 관련 정보가 없다면 예비실험을 통하여 자료를 확보하는 방안을 검토할 필요가 있다.

표본크기 계산 사례

(1) 단일 집단 비율

이항분포에 대한 정규분포 근사성과 오차한계 (e = 신뢰계수 × 표준오차)를 이용하여 계산공식을 유도할 수 있다. 오차한계는 추정치 (비율)의 분포가 정규분포를 따른다면 표준정규분포에서 추정치의 68%는 참값에서 $p(1-p)/n$ 이내의 차이를 보인다. 이를 비율의 표준오차(standard error, SE)라고 하며 표본 추정치의 95%는 참값에서 1.96SE 이내의 차이를 보인다(8,26). 이항분포에서 비율을 p ($p = x_i/n$, $p = 1-q$)라 할 때 표준오차 [$SE(p)$]와 오차한계는 다음과 같다.

$$e = \text{신뢰계수} \times \text{표준오차} = z_{1-\alpha/2} \times SE(p) = z_{1-\alpha/2}\sqrt{p(1-p)/n}$$

이 식에서 n에 대하여 정리하면 다음의 공식이 유도된다. 이 공식은 단순무작위추출에서 진단검사의 특성이 완벽하다고 가정한 것이다.

$$n = \frac{z_{1-\alpha/2}^2 p(1-p)}{e^2}$$

이 공식은 이항분포에 대한 정규분포 근사성을 활용한 것으로 비율 p가 $0.2 < p < 0.8$ (혹은 $0.1 < p < 0.9$)의 범위를 보일 때 유효하게 사용할 수 있으며 p가 0이나 1에 접근하는 경우 정확(exact) 이항분포를 이용하여 계산한다(5,15,20). 또한 모집단크기(N)에 비하여 표본크기(n)가 5% 이상이라면 유한모집단보정계수(finite population correction)를 적용하며 다음의 공식을 사용한다.

$$n = \frac{N z_{1-\alpha/2}^2 p(1-p)}{e^2(N-1) + z_{1-\alpha/2}^2 p(1-p)}$$

한편 검정력을 고려할 때 양측검정에서 가설상의 수치 (즉 두 실험군 중 어느 한 군의 비율을 알고 있을 때)와 비교하는데 필요한 표본크기를 계산해야 하는 경우가 있다. 이를테면 특정 약제 A의 치료효과에 대한 대규모 연구에서 20%의 성공률이 보고되었다고 하자. 동일한 조건에서 새로 개발한 약제의 치료 성공률을 평가한다고 할 때 약제 A의 성공률 20%를 알고 있기 때문에 신약 효과를 평가하는 연구에 약제 A를 포함시킬 필요가 없다. 이러한 연구 형태를 historical control study라고 하며 실험군이 1개인 상황이다. 따라서 단일 비율에 대한 이항검정을 위한 표본크기를 계산하게 되며 다음의 공식을 사용한다(21).

$$n = \frac{p_0 q_0 \left[z_{1-\alpha/2} + z_{1-\beta}\sqrt{\frac{p_1 q_1}{p_0 q_0}} \right]^2}{(p_1 - p_0)^2} = \frac{\left[z_{1-\alpha/2}\sqrt{p_0 q_0} + z_{1-\beta}\sqrt{p_1 q_1} \right]^2}{(p_1 - p_0)^2}$$

감염검출: 무한모집단의 크기를 N, 모집단에서 감염된 개체 수를 d, 표본크기를 n이라할 때 유병률이 p인 모집단에서 표본을 추출한다고 하자. 크기가 1인 표본을 선발하여 검사결과가 이분형 (양성 혹은 음성)인 이러한 과정을 무수히 반복할 때 검사 양성 개체수가 x두일 확률은 이항분포를 따른다(8).

$$P(T+=x)=\binom{n}{x}p^x(1-p)^{n-x}$$

만일 전 두수 검사음성이라면 $x=0$이 되므로 $P(T+=0)=(1-p)^n$로 정리된다. 따라서 n두의 표본에서 적어도 1두가 감염되어 있을 신뢰수준(C)은 $C=1-(1-p)^n$이 되며, 이 식을 n에 대하여 정리하면 다음과 같다.

$$n=\frac{\log(1-C)}{\log(1-p)}$$

이 공식에서 분모는 검사결과 음성인 개체 수고, 분자의 $1-C$는 유의수준(α)으로 제1형 오류다. 만일 검사의 민감도(Se)가 완벽하지 않다면 $p=p\times Se$로 대치한다.

예제 1: 어느 지역에서 개의 심장사상충 항원양성률이 20%로 알려져 있다고 할 때 표본추정치가 모집단 유병률 참값의 4% 이내로 추정하기 위해서는 몇 두를 검사해야 하는가? 95% = 2SE의 관계에서 2SE는 4%를 이내가 되어야 하므로 1SE는 2%에 해당한다. 따라서 400두를 검사하면 된다.

$$n=\frac{0.2\times(1-0.2)}{0.02^2}=400$$

예제 2: 특정 품종의 개에서 호발하는 종양 'X'의 발생률은 2%로 알려져 있다. 유전적 소인이 있을 때 종양 발생의 위험이 5%로 증가하는 것을 확인하기 위해서는 (즉 3%의 차이 검출) 5% 유의수준에서 90%의 검정력을 달성하기 위해서는 몇 두를 조사해야하는가? 일반 모집단에 비하여 유전적 소인이 있는 집단에서 암 발생률이 2.5배 높다고 할 때 95% 신뢰수준과 양측검정에서 341두를 조사하면 유의한 차이를 검출할 90%의 기회를 갖는다.

$\alpha=0.05,\ 1-\beta=0.9;\ p_0=0.02;\ p_1=0.05$이므로

$$n=\frac{(0.02)(0.98)\left[1.96+1.28\sqrt{\frac{0.05(0.95)}{0.02(0.98)}}\right]^2}{(0.02-0.05)^2}\approx 341$$

예제 3: 어느 양계장에서 2%의 닭이 추백리에 감염되어 있다고 할 때 감염된 1수를 검출하는 것을 95% 신뢰하기 위해서는 148수를 검사하면 된다.

$$n=\frac{\log(1-0.95)}{\log(1-0.2)}\approx 148$$

(2) 단일 집단 평균

단일 평균에 대한 표본크기라 함은 모집단의 평균이 표본평균과 차이가 있는지를 평가하는 것으로 표본크기를 계산하기 위해서는 첫째, 귀무가설(μ_0)과 관련된 유의수준(α), 둘째, 대립가설(μ_1)과 관련된 검정력($1-\beta$), 셋째, 임상적으로 중요한 것으로 판단하기 위한 평균치 차이($d=\mu_1-\mu_0$), 넷째, 모집단 표준편차(σ) 등 네 가지 모수에 대한 정보를 필요로 한다. 양측검정에서 단일 평균 추정을 위한 표본크기 계산공식은 다음과 같다.

$$n=\frac{(z_{1-\beta}+z_{1-\alpha/2})^2\sigma^2}{(\mu_0-\mu_1)^2}$$

이 공식에서 보듯이 표본크기는 귀무가설의 평균과 대립가설의 평균 간 절대차이가 작을수록, 분산이 증가할수록, 유의수준이 작을수록(z_α 증가), 검정력이 클수록 (즉 β오류 감소 혹은 z_β 증가) 증가한다는 것을 알 수 있다.

예제 4: 산란계 농장에서 계란의 두께가 평균 0.32 mm(표준편차: 0.08 mm)로 얇아 파손율 증가로 손실이 증가하고 있다. 계란의 두께는 사료섭취와 밀접한 관련이 있다는 보고에 따라 경영주는 특별히 제조한 사료를 30일간 제공한 후 계란을 회수하여 두께를 측정하고자 한다. 계란의 두께가 0.36 mm를 넘으면 매우 의미가 있는 결과로 간주한다고 할 때 유의수준이 5%에서 평균 차이 0.4 mm를 검출할 확률이 적어도 90%를 유지하기 위해서는 몇 개의 계란을 검사해야 하는가? 유의수준 5%에서 양측검정의 임계값은 $z_{1-\alpha/2}=1.96$이고 검정력 90%일 때 임계값은 $z_{1-\beta}=1.282$ [즉 $P(Z<1.282)=0.9$]이므로 약 35개의 계란을 측정하면 된다.

$$n=\frac{(1.645+1.282)^2\cdot 0.08^2}{(0.36-0.32)^2}\approx 35$$

(3) 두 집단 비율

두 집단 비율 (p_1, p_2)의 차이에 대한 표본크기를 계산할 때 두 군의 할당비(allocation ratio)를 k라 하면 각 집단에서 필요한 표본크기(n)는 다음의 두 공식으로 계산된다(6).

$$n=\frac{\left[z_{1-\alpha/2}\sqrt{\left(1+\frac{1}{k}\right)\times\bar{p}\bar{q}}+z_{1-\beta}\sqrt{p_1q_1+\frac{p_2q_2}{k}}\right]^2}{(p_1-p_2)^2}$$

또는

$$n=\frac{(z_{1-\alpha/2}+z_{1-\beta})^2[p_1(1-p_1)+p_2(1-p_2)]}{(p_1-p_2)^2}$$

$$\left[q_1=1-p_1,\ q_2=1-p_2,\ \bar{p}=\frac{p_1+kp_2}{1+k},\ \bar{q}=1-\bar{p}\right]$$

두 집단의 표본크기가 동일한 경우($k=1$) 위의 공식은 간단히 다음과 같이 정리된다.

$$n=\frac{\left[z_{1-\alpha/2}\sqrt{2\times\bar{p}\bar{q}}+z_{1-\beta}\sqrt{p_1q_1+p_2q_2}\right]^2}{(p_1-p_2)^2}$$

예제 5: 개의 질병 'x'에 대한 새로운 치료제의 효과를 폐사율로 측정하여 비교하는 연구에서 기존 약제는 40%, 신약은 31%의 폐사율을 보인다고 할 때 (즉 9%의 폐사율 차이를 유의하다고 판단할 때) 유의수준 5%에서 검정력 90%를 달성하기 위해서는 각 군 당 590두 (총 1,180두)가 필요하다.

$p_1 = 0.4$, $q_1 = 0.6$, $p_2 = 0.31$, $q_2 = 0.69$, $\bar{p} = (0.4 + 0.31)2 = 0.35$, $\bar{q} = 1 - 0.35 = 0.65$이므로

$$n = \frac{[1.96\sqrt{2 \times 0.35 \times 0.65} + 1.282\sqrt{0.4 \times 0.6 + 0.31 \times 0.69}]^2}{(0.4 - 0.31)^2} \approx 590$$

만일 두 군의 표본크기가 다를 경우 이를테면 $k = 2$ 로 가정하면 제 1군은 446두, 제 2군은 $n_2 = kn_1$의 관계에 의해 $n_2 = 2 \times 446 = 892$ 두가 된다.

$$n = \frac{\left[1.96\sqrt{0.35 \times 0.65 \times \left(1 + \frac{1}{2}\right)} + 1.282\sqrt{0.4 \times 0.6 + \frac{0.31 \times 0.69}{2}}\right]^2}{(0.4 - 0.31)^2} \approx 446$$

(4) 두 집단 평균

정규분포를 따르는 두 집단 평균의 차이(d)를 비교하는 연구에서 표본크기(n_1, n_2)는 두 집단의 표본크기가 동일할 때와 동일하지 않을 때로 구분할 수 있다. 두 표본 집단이 각각 정규분포를 따르는 모집단에서 표본추출한다고 가정하자.

표본크기가 동일할 때: 만일 두 집단의 분산이 서로 다르면 각각의 분산을 적용해야 하므로 각 집단에서 필요한 표본크기는 다음과 같이 계산된다.

$$n = \left[\frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \cdot (\sigma_1^2 + \sigma_2^2)}{d^2}\right]$$

이 공식에서 표준편차에 대한 정보를 모르는 경우 문헌고찰이나 예비실험을 통하여 추정해야 한다. 표준편차에 대하여 보수적인 추정치를 적용하면 검정력이 저하되기 때문에 연구 계획 단계에서 가능한 최상의 값을 추정해야 한다.

예제 6: 개에 특정 약제를 투여할 때 혈청 화학검사 항목 'a'가 변하는지를 조사할 계획이다. 약제 투여군과 대조군 (비투여군)에 대하여 검사항목 'a'의 평균을 측정한 결과 두 집단의 평균과 표준편차가 각각 $\bar{x}_1 = 132.86$; $s_1 = 15.34$; $\bar{x}_2 = 127.44$; $s_2 = 18.23$이라고 할 때 유의수준 5%, 양측검정에서 80%의 검정력을 유지하기 위해서는 각 군 당 152두 (총 304두)가 필요하다.

$$n = \left[\frac{(1.96 + 0.84)^2 \cdot (15.34^2 + 18.23^2)}{(132.86 - 127.44)^2}\right] \approx 152$$

예제 7: 어느 연구자는 질병 'x'에 대한 새로운 약제를 투여할 때 혈압이 증가하는 부작용이 있는지에 관심을 두고 있다. 기존약제를 투여하면 평균혈압은 81 mmHg (표준편차 18 mmHg)를 보인다. 신약을 투여할 때 14 mmHg 이상의 차이가 있을 경우 임상적으로 의미가 있는 변화로 간주한다면 유의수준이 5%에서 이러한 차이를 검출하는 검정력이 80%가 되기 위해서는 몇 두가 필요한가? 실험군 당 표본크기가 동일할 때 평균혈압의 공통분산은 324 mmHg, 유의수준이 5%에서 양측검정의 임계값은 $z_{1-\alpha/2} = 1.96$, 검정력 80%의 임계값은 $z_{1-\beta} = 0.84$ 이므로 각 군 당 26두 (n_1, 총 표본크기 52두)가 필요하고, 동일한 가정에서 유의수준을 1%로 가정하면 $z_{1-\alpha/2} = 2.58$ 이므로 각 군 당 39두 (n_2, 총 표본크기 78두)가 필요하다. 한편 두 집단의 표준편차를 각각 15와 12 mmHg라고 하면 각 군당 15두 (n_3, 총 표본크기 30두)가 필요하다.

$$n_1 = \frac{(1.96 + 0.84)^2 \times 2 \times 18^2}{14^2} = 26$$

$$n_2 = \frac{(2.58 + 0.84)^2 \times 2 \times 18^2}{14^2} = 39$$

$$n_3 = \frac{(1.96 + 0.84)^2 \cdot (15^2 + 12^2)}{14^2} = 14.8$$

n_1 (공통분산)과 비교할 때 n_3에서 표본크기가 30두로 감소한 것은 각 군의 표준편차가 15와 12 mmHg로 비교적 변동이 적고 두 군의 차이인 14 mmHg에 비하여 큰 차이가 없으므로 표본크기가 감소한 것이며 실험군의 혈압에 대한 표준편차가 클수록 표본크기는 증가한다.

표본크기가 다를 때: 두 군의 표본크기 다를 때(unequal sample size) 표본크기를 각각 n_1과 n_2, 총 표본크기를 n, 두 군의 할당비 (k)를 $k = n_1/n_2$로 정의하면 $n_2 = kn_1$ 일 때 n_1군의 표본크기는 다음과 같이 계산된다.

$$n_1 = \left[\frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \cdot (\sigma_1^2 + \sigma_2^2/k)}{(\mu_1 - \mu_2)^2}\right]$$

여기에서 두 군의 분산이 동일하다면 전체적으로 표본의 분산이 감소하므로 표본크기는 감소하며 이 경우 아래의 공식을 사용한다. 총 표본크기가 고정되어 있을 때 두 군간 표본크기를 동일하게 할당하면 ($k = 1$, $n_1 = n_2$) 검정력이 가장 높다.

$$n = \frac{(k+1)^2 \cdot (z_{1-\alpha/2} + z_{1-\beta})^2 \cdot \sigma^2}{k \cdot (\mu_1 - \mu_2)^2}$$

예제 8: 전술한 예제 6에서 두 실험 군의 표본크기가 다를 경우 ($k = 2$) 유의수준 5%, 양측검정에서 80%의 검정력을 유지하기 위해서는 약제 투여군 108두 (n_1), 대조군 216두가 필요하다.

$$n_1 = \left[\frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \cdot (\sigma_1^2 + \sigma_2^2/k)}{(\mu_1 - \mu_2)^2}\right]$$

$$= \left[\frac{(1.96 + 0.84)^2 \cdot (15.34^2 + 18.23^2/2)}{(132.86 - 127.44)^2}\right] \approx 108$$

본 연구에서는 표본크기 계산과 관련한 통계적 이론을 임상에서 흔히 접하는 비율과 평균 비교를 예시하여 설명하였다. 이러한 원리는 유병률 추정, 질병 검출, 비발생 증명 등 다양한 수의학 연구 분야에 필요한 표본크기를 계산하는데 직접 활용할 수 있다. 표본크기는 연구계획을 포함하여 상황에 따라 매우 다양한 방법으로 계산된다. 표본크기를 계산할 때 흔히 개체 간 독립성(independence)을 가정하는데 이 용

어는 연구에서 관심을 두고 있는 처리(treatment)에 대하여 연구 대상자들이 반응하는 결과가 전혀 관련이 없다는 것을 의미한다. 다음의 두가지 연구계획을 살펴보자.

연구 상황 1: 실험적으로 개의 대퇴골머리 골괴사(femoral head osteonecrosis)를 유발한 후 새로운 외과적 수술법의 치료효과를 평가할 예정이다. 이 방법은 전문적인 기술을 요구하기 때문에 술자의 경험에 따라 치료효과에 차이가 있을 것으로 판단되어 2곳의 동물병원을 대상으로 각각 15마리의 개를 할당하고자 한다. 이 질병에 대한 기존의 외과적 치료효과는 대략 68%로 보고되어 있으며 연구자는 새로운 치료법에 의한 성공률이 15% 이상 증가할 때 임상적으로 가치가 있다고 판단할 계획이다. 이러한 성공률의 변화가 실제로 존재할 때 이를 검출하는 것을 90% 확신하기 위해서는 몇 두의 실험견이 필요한가?

연구 상황 2: 약제 A의 창상치유 효과에 대한 연구를 계획하고 있다. 개의 피부에 실험적으로 피부 결손창을 유발한 후 특정 약제를 투여하고 2주 후 창상부위의 면적이 60%에서 90% 이상 변화할 때 임상적으로 의미가 있는 효과로 판단할 예정이다. 이 실험을 4곳의 동물병원을 대상으로 각각 20마리를 할당하여 평가한다고 할 때 연구목적을 달성하기 위해서는 실험견 몇 두가 필요한가?

상기의 두 사례는 전형적인 집락 무작위배정 임상시험(cluster randomized trial, CRT)에 해당하며, CRT 연구에서는 단순무작위추출에 비하여 표본크기에 큰 차이를 보인다. 다음 호에서는 nomogram을 이용하여 표본크기 계산을 간편하게 수행하는 방법과 CRT 연구에 필요한 표본크기를 계산하는 과정을 예시하여 설명한다.

감사의 글

본 연구는 2010년도 농림수산검역검사본부 연구사업 (과제번호: Z-FS10-2011-11-01)과 강원대학교 동물의학종합연구소의 지원에 의해 이루어졌으며 이에 감사드립니다.

참 고 문 헌

1. 박선일, 이영원. 자료분석의 기초. 한국임상수의학회지 2009; 26: 189-199.
2. Altman DG. How large a sample? In: Gore SM, Altman DG (eds). Statistics in practice. Londin, UK: British Medical Association 1982: 6-8.
3. Anonymous. Statistical guidelines for authors. J Med Screen 2008; 15: 51.
4. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? JAMA 1990; 263: 275-278.
5. Berry G, Armitage P. Mid-P confidence intervals: a brief review. Statistician 1995; 44: 417-423.
6. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ 1995; 311: 1145-1148.
7. Christley RM. Statistical significance, power and sample size - what does it all mean?. J Small Anim Pract 2008; 49: 263.
8. Cochran WG. Sampling techniques. 3rd ed. Wiley & Sons, New York, 1977: 1-428.
9. Cummings P. Reporting statistical information in medical journal articles. Arch Pediatr Adolesc Med 2003; 157: 321-324.
10. Daly LE. Confidence intervals and sample sizes: don't throw out all your old sample size tables. BMJ 1991; 302: 333-336.
11. Dell RB, Holleran S, Ramakrishnan R. Sample Size Determination. ILAR J 2002; 43: 207-213.
12. Diwan VK, Eriksson B, Sterky G, Tomson G. Randomization by group in studying the effect of drug information in primary care. Int J Epidemiol 1992; 21: 124-130.
13. Evans RB, O'Connor A. Statistics and evidence-based veterinary medicine: answers to 21 common statistical questions that arise from reading scientific manuscripts. Vet Clin North Am Small Anim Pract 2007; 37: 477-486.
14. Florey C. Sample size for beginners. BMJ 1993; 306: 1181-1184.
15. Fosgate GT. Modified exact sample size for a binomial proportion with special emphasis on diagnostic test parameter estimation. Stat Med 2005; 24: 2857-2866.
16. Hofmeister EH, King J, Read MR, Budsberg SC. Sample size and statistical power in the small-animal analgesia literature. J Small Anim Pract 2007; 48: 76-79.
17. Jones SR, Carley S. An introduction to power and sample size estimation. Emerg Med J 2003; 20: 453-458.
18. ICMJE (International Committee of Medical Journal Editors). Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. 2008.
19. Kojima T, Barron JP. Changes in the ethos of medical publications as reflected in progressive alterations in the uniform requirements for manuscripts submitted to biomedical journals (1979-2008). Chest 2010; 137: 1479-1482.
20. Lwanga SK, Lemeshow S. Sample size determination in health studies - A practical manual. 1st ed. Geneva: World Health Organization; 1991: 1-80.
21. Machin D, Campbell MJ, Tan SB, Tan SH. Sample size tables for clinical studies. 3rd ed, BMJ Books, 2008: 1-264.
22. Martin SW, Shoukri M, Thorburn MA. Evaluating the health status of herds based on tests applied to individuals. Prev Vet Med 1992; 14: 33-43.
23. McDermott JJ, Schukken YH, Shoukri MM. Study design and analytic methods for data collected from clusters of animals. Prev Vet Med 1994; 18: 175-191.
24. Moher D, Schultz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Lancet 2001; 357: 1191-1194.
25. Naing L, Winn T, Rusli BN. Practical Issues in calculating the sample size for prevalence studies. Arch Orofac Sci 2006; 1: 9-14.
26. Snedecor GW, Cochran WG. Statistical methods. 8th ed. Ames: Iowa State Press, 1989: 1-503.
27. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. Stat Med 2002; 21: 3757-3774.
28. Whitley E, Ball J. Statistics review 4: Sample size calculations. Crit Care 2002; 6: 335-341.