주요개념:표본크기의 결정,유효크기,검정력

간호학 연구에서의 표본크기 결정 방법에 대한 고찰

이재원 · 박미라 · 이정복 · 이숙자 · 박은숙 · 박영주

이재원 · 이정복:고려대학교 통계학과 교수
박미라:을지의과대학 의예과 교수
이숙자 · 박은숙 · 박영주:고려대학교 간호대학 교수

Ⅰ. 서 론

간호학 및 의약학 분야의 임상시험 연구에 있어서 가장 중요한 사항중의 하나가 필요한 환자수(표본의 크기)를 결정하는 문제이다. 연구대상환자수는 연구의 설계단계에서 미리 고려되어야 하는 것임에도 불구하고 많은 연구에서 이러한 원칙이 무시되고 일정기간 동안에 확보할 수 있는 정도를 대상환자수로 결정해 버리는 경우가 있다. 검정력 분석(power analysis)은 필요한 표본크기를 계산하기 위해서 연구를 시작하기 전에 하는 것이 보통이지만, 연구가 끝나고 데이터를 분석한 후에 연구결과의 정당성을 보여주기 위해서 하는 경우도 많이 있다. 특히 기대한 정도의 차이가 발견되지 않은 연구에서 검정력 분석을 필수적이다. 왜냐하면 사전에 대상환자수를 고려하지 않은 임상연구 중에는 임상적으로 의미있는 차이가 있는데도 검정력이 떨어져서 이를 찾아내지 못하는 경우가 많기 때문이다. 검정력이 작은 연구의 위험성은 유익하게 쓰여질 수 있는 치료법 등이 충분히 검증되지 못하고 기각되어 두 번 다시 고려될 수 없다는데 있다.

간호학 분야의 연구자들은 최근에서야 검정력 분석을 통하여 연구대상자 수를 구하기 시작하였는데, 대부분의 경우에 연구결과를 발표할 때 검정력 분석과정을 보고하지 않고 있다. Polit & Sherman(1990)은 1989년에 Nursing Research와 Research in Nursing and Health에 발표된 62편의 논문에 나타난 표본크기를 평가하였는데, 한 그룹당 평균 83명의 환자가 할당되고 전체 논문의 2/3에서 그룹당 환자수가 100명 이하인 것을 발견하였다. 평가된 논문에서 그룹간 비교를 할 때 가장 약한 수준인 80%정도의 검정력을 확보하기 위해서는 그룹당 평균 218명이 필요한한 것을 감안할 때, 대부분의 연구에서 표본크기가 불충분하였다는 것을 감안할 때, 대부분의 연구에서 표본크기가 불충분하였다는 것을 쉽게 알 수가 있다. 62편의 논문 중에서 단지 한 편만이 검정력 분석과정을 보고하고 표본크기가 충분하다는 것을 보였다. 또한 최근의 한 연구에서는 1988년 1992년까지 5년동안 3개의 간호학저널에 발표된 논문을 검토하였는데, 단지 Nursing Research에서 8편의, Research in Nursing and Health에서 9편, Western Journal of Nursing Research에서 3편만이 검정력 분석과정을 보고하였다는 것을 보였다.

연구대상자 수를 구하는 방법은 연구시작 전에 미리 고정된 수를 할당하여 연구를 시행하는 고정설계법(fixed design)과 연구를 시행해가면서 나타나는 결과에 따라 표본의 수가 결정되는 축차설계법(sequential design)에서 각각 다르다. 연구대상자 수를 결정할 때에는 유의수준의 검정력의 크기, 연구자가 기대하는 차이의 정도, 그리고 사용되어지는 통계기법 등이 고려되어야 한다. 또한 단 한번의 분석만을 할 것인지 연구도중에 여러번의 중간분석(interim analysis)을 하게 될것인지에 따라서도 연구대

상자 수의 계산이 달라지게 된다. 중간 분석을 실시하도록 계획되어 있는 경우에는 필요한 표본의 크기가 증가하게 된다. 그러나 연구가 일찍 종료될 가능성도 있기 때문에 실제로 연구에 참가하는 환자의 수는 오히려 더 작아질 수도 있게된다. 중간분석을 수행할 경우에 어떻게 연구대상자 수를 결정하는가 하는 문제애 대해서는 Kim and DeMets(1987)나 Jennison and Turnbull(1989)에서 자세히 논의되었고, 여기에서는 고정설계에 의한 연구에서 단 한번의 최종분석이 시행되는 경우에 한하여 연구대상자 수를 구하는 몇가지 유용한 방법들을 연구형태별로 소개하기로 한다.

Ⅱ . 가설검정의 개념

연구대상자 수를 추정하는데 있어서는 가설검정(hypothesis testing)이나 유의수준(significance level), 검정력(power)등의 기본개념에 대한 이해가 필수적이므로, 먼저 이들에 대한 설명을 간략하게 한 후에 대상환자 수의 추정에 대한 구체적인 방법들을 설명하기로 한다.

실험군과 대조군의 두 그룹을 비교하는 문제를 생각해 보자. 만약 반응변수가 질병의 재발여부 등과 같이 이상형(discrete)인 경우라면 연구자가 관심을 갖는 것은 흔히 대조군에서의 반응률(ρ_1)과 실험군에서의 반응률(ρ_2)에 차이가 있는지에 관한 문제일 것이다 먼저 두 반응률에 차이가 없다는 가설을

$$H_0 : \rho_1 - \rho_2 = 0$$

으로 표현 할 수 있다. 이러한 가설은 연구자의 기대가 무위로 돌아가는 가설이며, 이를 귀무가설(null hypothesis)이라고 한다. 연구자의 목적은 귀무가설이 기각되는 지의 여부를 결정하는데 있으며, 귀무가설은 반증이 되기 전까지는 사실인 것으로 가정한다. 반응률에 대한 하나의 추정치만이 구해지므로 실제로는 두 집단의 반응률의 차이가 없는데도 불구하고 관측된 값에서 반응률의 차이가 크게 나타날 수 있다. 이 값이 우연히도 매우 크게 나타났다면 연구자는 사실과 다르게 귀무가설을 기각하게 되는 잘못을 범하게 된다. 이러한 종류의 오류를 제 1종 오류(Type Ⅰ error)라고 한다. 이러한 오루가 일어날 확률의 최대 허용치를 유의수준(significance level)이라고 하며 통상 α로 표기한다. 귀무가설이 옳을 때 반응률의 차이가 관측된 차이보다 크거나 같을 확률을 유의확률 또는 ρ-값이라고 한다. 따라서, ρ-값이 미리 정해진 유의수준 α조다 작거나 같을 때 귀무가설을 기각하게 된다. α의 선택은 이론적으로는 어떤 값을 택해도 무관하나 많은 경우 0.01이나 0.05를 사용한다. 예컨데 α를 0.01로 선택하고 ρ-값이 이보다 작게 계산되었을 때 “유의수준 1%에서 귀무가설은 기각된다”고 표현한다.

이에 대해 두 집단의 반응률에 차이가 있다는 가설은

$$H_1 = \rho_1 - \rho_2 = \delta\ (\neq 0)$$

으로 표현된다. 이와같이 연구자가 증명하고 싶은 주장을 대립가설(alternative hypothesis) 또는 연구가설에도 불구하고 관측된 차이가 우연히 매우 작은 값이 될 수도 있다. 이때 연구자는 이에 근거하여 귀무가설을 기각하지 않는 오류를 범하게 되는데, 이러한 오류를 제2종 오류(Type Ⅱ error)라고 하며 이러한 오류가 일어날 확률을 β로 표기한다. β의 값은 두 그룹간 반응률의 차이(δ), 그리고 표본의 크기(여기서는 대상환자수)와 α에 의존한다. 귀무가설이 사실이아닐 때 귀무가설을 기각하게 될 확률은 1-β가 되는데 이값을 검정력(power)이라고 한다. 검정력은 다양한 δ에 대해서 옳고 표본의 크기가 주어졌을 때 1-β도 역시 이들의 함수가 된다. 표본의 크기가 주어졌을 때 1-β와 δ의 그래프를 검정력 곡선(power corve)이라고 한다. 간호학 및 의약학 연구의 설계시 검정력을 0.80에서 0.95 사이로 정하는 것이 보통인데, 이것은 실제로 δ라는 차이가 존재하는 할 때에 반응률간에 통계적으로 유의한 차이를 찾아낼 확률을 80%에서 95%사이로 정하겠다는 의미이다.

유의수준 α는 작아야 하고 (예컨대 0.05 또는 0.01) 검정력 1-β는 커야 하는데(예컨데 0.80또는 0.90), 변화될 수 있는 값은 δ와 표본의 크기이다. 임상시험을 설계하는데 있어서 연구자들은 특정수준의 차이 δ 또는 그 이상의 차이를 찾기를 원하는데, δ의 선정에 있어서 고려해야 할 요소중의 하는

임상적으로 중요하다고 생각되는 최소의 차이이다. 간호학 분야의 연구자들은 이와 비슷한 개념으로 표준화된 효과의 크기로 정의되는 유효크기(effect size)를 사용한다. 두 평균을 비교하는 실험인 경우 표준화된 평균 차이로 유효크기를 정의하며, 원하는 가설검정의 형태에 알맞게 유효크기를 정의할 수도 있다(cf. Cohen, 1977). 유효크기에 관해서는 다음 절에서 자세히 논의하도록 하겠다. 이제 α, $1-\beta$, δ가 주어졌을 때 표본의 크기를 계산하는 문제에 대하여 생각해 보자, 먼저 확률화에 의해 각 그룹에 동일한 수(N)의 표본을 할당한다고 가정하자. 두 그룹에서 반응변수의 변동(variation)이 근사적으로 같을 때에는 같은 수를 할당하는 것이 가장 큰 검정력을 가진다. 또한 동일한 수를 할당하는 문제가 훨씬 간단하기 때문에 보다 선호되고 있다. 그러나 경우에 따라서는 윤리적 또는 경제거 이유등으로 서로 다른 환자수를 할당하는 것이 바람직 할 때도 있다.

표본의 크기(연구대상자 수)를 계산하기 전에, 연구자는 실험군에 대해 한쪽 방향의 차이에만 관심이 있는지, 아니면 실험군에서 반응률이 증가 또는 감소되었는가 하는 문제와 같이 양쪽 방향으로 모두 관심이 있는지를 먼저 결정해야 한다. 전자의 경우를 단측검정(one-sides test), 후자의 경우를 양측검정(two-sided test)이라고 한다. 이러한 결정을 해야하는 이유는 유의수준 α가 이의 경향을 받게 되기 때문이다. 일반적으로 한쪽 방향으로의 차이점만을 고려할 만한 충분한 이유가 없을 때에는 양측검정을 사용하게 된다. 연구자는 새로운 실험이 유익할 수도, 유해할 수도 있다는 사실을 항상 명심해야 한다. 그러나 만약 단측검정이 정당화 될 수있다면 동일한 수준의 α, $1-\beta$에 대하여 요구되는 표본의 크기가 양측검정에 비해 작아지게 된다.

위에서 언급했듯이 총 표본의 크기(총 대상환자의 수)는 유의수준(α)과 검정력($1-\beta$), 그리고 찾아내야 할 반응률의 차이(δ)의 함수가 된다. α, $1-\beta$ 또는 δ를 바꾸면 총 표본의 크기가 바뀌게 된다. δ가 작아지면 그 차이를 찾아내기 위한 표본의 크기는 커지게 된다. 만약 계산된 대상환자 수가 현실적으로 구할 수 있는 것보다 많다면 연구설계에서 α, $1-\beta$ 또는 δ를 수정하지 않으면 연구를 수행할 수 없게 된다. 유의수준은 통상 0.05나 0.01로 고정되기 마련이므로 연구자는 δ를 보다 큰 값으로 재조정하거나 마련이므로 연구자는 δ를 보다 큰 값으로 재조정하거나 아니면 δ를 유지하고 연구의 검정력이 떨어지는 것을 감수해야 한다.

지금까지 비교해야 할 그룹이 실험군과 대조군의 두 그룹으로 나뉘어지는 경우에 대하여 설명하였으나, 이러한 방법들을 비교해야 할 그룹이 둘 이상인 경우로의 확장이 가능하다. 또한 자료의 분석이 임상실험의 최종단계에서 단 한번 수행되는 것을 전제로 하여 설명하였는데, 만약 반응변수의 자료가 연구기간 동안 주기적으로 분석된다면 우연히 유의한 차이를 발견하게 될 기회가 그만큼 많아진다고 보아야 한다. 따라서 제 1종 오류의 확률이 증가하는데 대한 보완이 필요하게 되므로 이러한 경우에는 유의수준을 α를 재조정해야 한다.

이상과 같이 이산형 반응변수에 대해 두 그룹의 비율을 비교하는 문제를 예로 들어 대상환자 수를 계산하는데 필요한 여러 개념들을 설명하였다. 반응변수가 혈압등과 같이 연속형(continuous case)일때에는 대조군에서의 모평균(μ_1)과 실험군에서의 모평균(μ_2)을 비교하는 문제등을 생각할 수 있다. 이밖에도 여러 다양한 상황들이 있을 수 있다. 여러상황하에서 연구대상환자수를 결정하는 문제를 구체적으로 다루어보기 전에 유효크기에 대해서 논의하기로 하자.

Ⅲ. 유효크기

간호학 분야의 연구자들은 특정수준의 차이인 δ와 비슷한 개념으로 표준화된 효과의 크기로 정의되는 유효크기인 경우 표준화된 평균차이로 유효크기를 정의할 수 있다. Cohen(1977)은 자주 사용되는 다섯 종류의 검정에 대해서 <표1>과 같이 유효크기를 계산하는 공식을 제안하였고, 또한 작고 큰 정도를 나름대로 정의하였다.

앞 절에서 언급하였듯이 표본의 크기를 결정할 때에는 유의수준, 검정력의 크기, 연구자가 기대하는 차이의 정도 또는 유효크기, 단측 또는 양측 검정여부, 연구의 형태, 그리고 사용되어지는 통계기법 등이 고려되어야 한다. 통상적으로 간호학 및 의약학 연구에서는 5%의 유의수준과 80%의 검정력을 사용하는데, 연구를 시작하기 전에 위의 요소중에서 가장 결정하기가 어려운 것이 바로 유효크기이다.

Polit와 Sherman(1990)은 유효크기를 결정하기 위한 다음과 같은 네가지 방법을 설명하였다. 첫째, 과거의 비슷한 연구로부터 유효크기를 추정하는 것이다. 관계되는 연구가 많이 있을 때는 메타분석(meta analysis)을 사용해서 추정하는 것이 가장 바람직하다고 하겠다. 둘째, 관계되는 연구가 거의 없다면 작은 규모의 시험연구(pilot study)를 실시하여 추정하는 것이 좋다. 셋째, 시험연구를 실시할 형편이 안되면 '가상표(dummy table)'를 작성하여 임상적으로나 이론적으로 가치가 있기에 충분히 크다고 간주할 수 있는 가장 작은 유효크기를 계산한다(cf. yaran야, 1991). 예를 들어, 어떤 금연요법이 적어도 10%의 금연효과를 가져올 때 효과적이고 가치있는 것이라고 간주할 수 있다면, 유효크기는 이 값에 근거해서 추정할 수 있다. 넷째, 방법은 앞의 세가지 방법을 적용할 수 없을때 마지막으로 사용하는 방법인데, 과거의 연구 경험으로 단지 유효크기가 작은지(small), 적당한지(medium), 큰지(large)만을 결정하여 그에 따라 정해진 추정치를 할당하는 것이다(cf. Cohen, 1988). 이 방법을 사용하는데 있어서 강조할 점은 새로운 연구분야에서는 보통 유효크기가 작다는 것이다.

유효크기가 작을수록 많은 표본이 필요하다는 것은 당연하다. Pilot와 Sherman(1990)은 그들이 조사한 연구로부터 계산된 유효크기들의 52.7%가 작으면 유효크기가 작은 연구의 평균 검정력은 부족한 환자수로 인해서 30%도 안된다는 것을 알아냈다. 또한 그들은 유효크기가 적당한 연구도 검정력의 평균이 70%밖에 되지 않아서 최저수준인 80%에도 못 미치고 있으며, 심지어 유효크기가 큰 연구중의 11%의 연구의 검정력이 80%가 되지 않았고, 전체 연구중의 단지 15%만이 충분한 검정력을 가진다는 것도 발표하였다. 그들이 권위있는 간호학저널을 조사했다는 것을 감안해 볼 때 이는 가히 충격적인 결과이며, 앞서 언급한대로 저널에 발표되지 않은 연구까지 고려한다면 대부분의 간호학 연구에서 표본크기가 부족한 실정이라고 결론지을수 있다.

다음절부터는 연구형태별로 표본의 크기를 결정하는 문제를 구체적으로 다루어보도록 하겠다.

<표1> 다섯가지 검정에서의 유효크기 계산공식과 유효크기 정도에 대한 정의

검 정	유효크기	유효크기 값 소	중	대
두 집단 평균에 대한 t-검정	$d = \frac{\lvert X_1 - X_2 \rvert}{\sigma}$	.20	.50	.80
k개의 독립평균에 대한 F-검정	$f = \frac{\sigma m}{\sigma}$	.10	.25	.40
ρ ($\rho \neq 0$)	ρ	.10	.30	.50
두 비율치 차에 대한 감정	$\phi i = 2arscine\sqrt{Pi}$ $h = \lvert \phi_1 - \phi_2 \rvert$	.20	.50	.80
F($R^2 \neq 0$)(다중회귀)	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

Ⅳ. 연구형태별 표본크기의 결정

4.1 단일집단에서의 평균

만약 어떤 확률 변수 X의 평균이 μ_1인지에 대해서 검정하기 위해서 귀무가설로 평균이 $\mu_0(\neq \mu_1)$라고 하고, 대립가설로 평균이 μ_1이라고 하였다. 이를

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu = \mu_1 (\neq \mu_0)$$

으로 표현할 때, 이에 대한 검정통계량

$$Z = \frac{x - \mu}{\sigma / \sqrt{n}}$$

은 근사적으로 표준정규 분포를 따르게 된다. 제 1종 오류 (또는 유의수준)를 α로 하고, 제 2종 오류를 β라고 할 때, 위의 통계량에 근거하여 이들 오류에 대한 표준정규분포의 임계치 $Z_{\frac{\alpha}{2}}$는 유의 수준 α에 해당하는 표준정규분포 양측 임계체이고, Z_β는 제 2종 오류에 해당하는 표준정규분포 단측 상한의 임계치가 된다. $\sigma = \mu_1 - \mu_0$ 를 평균의 차이라고 할 때, 위의 두 임계치로부터 표본크기 n을 정리하면

$$n = \left[\frac{(Z_{\frac{\alpha}{2}} + Z_\beta)\sigma}{\delta} \right]^2$$

와 같이 공식을 유도할 수 있다.

유의수준과 제2종 오류가 주어졌을 때, 위의 공식은 δ와 σ에 따라 변하게 된다. 만약 평균의 차이 δ가 크면 표본크기는 감소할 것이고, 작다면 표본크기는 작이질 것이고, 크다면 표본크기도 역시 커질 것이다.

만약 위에서 제시했던 양측 검정의 문제와 다르게 단측검정일 경우 $Z_{\frac{\alpha}{2}}$는 Z_α로 대치하여 계산하면 된다. 즉 α=0.05에서 양측검정시에는 $Z_{\frac{\alpha}{2}}$는 1.96일 것이고, 단측검정시에는 Z_α가 1.645가 된다. 통계학 교과서에 나와있는 정규분포표로부터 원하는 Z_α와 Z_β의 값들을 찾아낼 수 있다. <표3>은 자주 쓰이는 값들을 정리한 것이다.

<표> Z_α값

유의수준(α)	단측검정(Z_α)	양측검정($Z_{\alpha/2}$)
0.10	1.282	1.645
0.05	1.645	1.960
0.025	1.960	2.240
0.01	3.236	2.576

한 예를 들어보자. 일반 환자에 대한 평균 최대흡입압력(maximal inspiratory mouth pressure)은 110cm H_2O이라 할 때, 한 연구자는 척주후만증(kyphoscoliotic)환자의 경우도 일반 환자와 차이가 있는지 알아보기 위해 평균의 차이를 10으로 하고 이에 필요한 표본크기를 계산하였다. 유의수준을 0.05, 검정력을 90%로 하면 $Z_{\frac{\alpha}{2}}$는 1.96이고 Z_β는 1.282이고, 선행연구에 의해 표준편차가 20인 것을 알았다. 이때 표본의 크기는

$$n = \left[\frac{(1.96 + 1.282)20}{10} \right]^2 = 42.0297$$

로 약 43명의 척추후만증 환자의 최대흡입압력을 측정할 필요가 있다.

<표3> Z_β값

검정력(1-β)	단측검정(Z_β)
0.50	0.00
0.60	0.25

0.70	0.53
0.80	0.84
0.85	1.03
0.90	1.282
0.95	1.646
0.975	1.960
0.99	2.326

4.2. 단일집단에서의 비율

단일표본에서 비율에 대한 표본크기의 결정문제는 앞의 평균에 대한 문제와 유사하다. 어떤 반응변수 X가 n회 시행시 성공률이 p_1인지에 대해 검정하기 위해 가설을

$$H_0 : p = p_0 \text{ vs } H_1 : p = p_1 \ (\neq p_0)$$

와 같이 세웠다. 이에 대한 검정통계량

$$Z = \frac{x - np}{\sqrt{np(1-p)}}$$

은 n이 크다면 근사적으로 표준정규분포를 따르게 된다. 제 1종 오류(또는 유의수준)를 α로 하고, 제 2종 오류를 β라고 할 때, 위의 통계량에 근거하여 이들 오류에 대한 표준정규분포의 임계치 $Z_{\frac{\alpha}{2}}$는 유의수준 α에 해당하는 표준정규분포의 임계치이고, Z_β 제 2종 오류에 해당하는 표준정규분포의 단측 상한의 임계치가 된다. 대립가설과 귀무가설의 비율의 차를 $\delta = \rho_1 - \rho_0$라고 할 때, 위의 두 임계값을 통해서 표본크기 n으로 정리하면,

$$n = \left[\frac{(Z_{\frac{\alpha}{2}}\sqrt{p_0(1-p_0)} + Z_\beta\sqrt{p_1(1-p_1)}}{\delta} \right]^2$$

와같은 공식을 유도할 수 있다. 만약 위에서 제시했던 양측검정의 문제와 다르게 단측검정일 경우 $Z_{\frac{\alpha}{2}}$는 Z_α로 대치하여 계산하면 될 것이다.

한 예를 들어보자. 호흡기질 환자에게 의사 및 간호사의 지속적인 금연 권고가 호흡기질환 환자의 금연에 도움을 주는지에 대해 알아보기 위해 연구계획을 세웠다. 사전 연구에 의해 호흡기질환 환자의 금연 실패율은 40%라는 것을 알았을 때, 금연 권로를 받은 환자의 금연 실패율이 25%인지에 대한 검정을 하기 위해 표본크기를 계산하였다. 지속적인 금연 권고를 받은 환자가 금연 실패율이 더 높아지는 경우는 없다고 가정하고 단축 검정을 실시하였다. 유의수준을 0.05로 검정력을 80%로 하면 $Z_{\frac{\alpha}{2}}$는 1.645이고 Z_β는 0.84이고

$$n = \left[\frac{1.645\sqrt{0.40 \times 0,60} + 0.84\sqrt{0.25 \times 0.75}}{0.15} \right]^2 = 60.7997$$

와 같은 표본크기를 얻을 수 있다. 따라서 약 61명의 호흡기질환 환자를 대상으로 할 것이다.

4.3. 두집단에서의 평균비교

병원에 입원한 기간이나 콜레스테롤 수준, 혈압, 폐활량 등 연속형 반응 변수를 관찰하는 실험에서 필요한 표본크기를 계산하는 방법에 대해 알아보자. 반응변수 y가 평균 y가 평균이 μ이고, 분산이 σ^2인 정규분포를 따른다고 가정하고, 그룹1과 그룹2에 각각 N명과 rN명이 확률화를 통해 할당된다고 하자. 두 그룹의 평균간에 차이가 없다는 것을 귀무가설로 하고, 대립가설을 두 그룹의 평균간에 차이가 있다는 것으로 설정하면, 이를

$$H_0: \delta = \mu_1 - \mu_2 = 0 \text{ vs } H_1: \delta = \mu_1 - \mu_2 \neq 0$$

과 같이 표현할 수 있다. 다음과 같은 검정통계량

$$Z = \frac{y1 - y2}{\sigma\sqrt{(1/N)+(1/rN)}}$$

은 분산 σ^2이 알려져 있을 때 근사적으로 표준정규분포를 따른다. 여기서, y_1와 y_2는 각각 그룹 1과 그룹 2에서 관찰된 자료들의 표본 평균이다. σ^2이 알려져 있지 않으면 σ대신 그 추정치인 s를 사용하는데, 검정통계량은 자유도가 (N+rN-2)인 t-분포를 따른다. 2절에서 논의되었던 가설검정에 관한 통계적 개념을 도입하면 z의 분포로부터 표본크기를 구하는 공식을 쉽게 유도할 수 있다. 두 그룹에 동일한 수의 환자를 할당한다고 할 때(즉, r=1), 유의수준α와 검정력 1-β로 평균의 차이를 찾기위해서 필요한 N은

$$N = [2(Z_{n/2} + Z_\beta)^2 \sigma^2]/\sigma^2$$

과 같이 구할 수 있다.

위의 공식에서 보듯이, 관심있는 평균차이 σ가 작아지거나, 분산σ^2이 증가할수록 필요한 환자의 수가 증가한다. 또한 유의수준 α가 작을수록, 검정력(1-β)이 클수록 증가한다. 역시 단측 검정의 경우에는 $Z_{n/2}$대신 Z_α 를 사용하면 된다.

사례를 들어보자. 식이요법이 콜레스테롤 수준에 미치는 영향을 알아보기 위한 한 연구에서, 식이요법을 실시한 환자군이 대조군에 비해 콜레스테롤 수준에서 20mg/dl 만큼의 차이가 있음을 보이고자 할 때 필요한 환자수를 계산하여 보자. 선행연구로부터 각 그룹의 분산이 (50mg/dl)으로 추정되어졌다고 하자. 즉 d=20, $\sigma^2=50^2$이다. 5%의 유의수준과 80%의 검정력을 가지고 양측검정을 한다고 하면, 두그룹에 동일한 수의 환자를 할당한다고 했을 때

$$N = \frac{2(1.96+0.84)^2 (50)^2}{10^2} = 98.11$$

으로 계산된다. 즉, 한그룹당 99명씩 총 198명 정도의 환자가 필요하게 된다.

연구자가 평균의 차이를 보는 대신에 평균수준에 변화에 더욱 관심있는 경우가 종종있다. 가령 유방암 환자에 대해 한 그룹에는 새로운 신경안정요법을, 다른 그룹에는 기존의 신경안정요법을 행했을 때, 어떤 요법이 실험전과 후의 차이가 더 큰가에 관심이 있을 수 있다. 이 문제는 앞의 경우와 근본적으로는 같으나 처리를 하기 이전에 환자 개개인의 처음 혈압수준이 고려되어야 한다.

Δ_1와 Δ_2를 각각 대조군과 실험군에서의 평균수준의 변화(=연구종료 후의 수준-연구시작점에서의 수준)를 나타낸다고 하자. 이제 귀무가설과 대립가설은 각각

$$H_0: \delta = \Delta_1 - \Delta_2 = 0 \text{ vs } H_1: \delta = \Delta_1 - \Delta_2 \neq 0$$

이된다. 변화의 분산을 σ_Δ^2라고 하자. 일반적으로, 이는 각 관찰치의 분산 σ^2보다 작은 경향이 있다. 이런 식으로 σ와 σ_Δ^2을 정의한 다음, 앞의 공식을 적용할 수 있다. 앞의 예에서, 연구자가 두 그룹 사이에 10점 만큼의 신경 불안 점수의 변화수준 차이를 보이고 싶다고 하자, 변화의 분산 σ_Δ^2이 20^2이라면, 유의수준 5%, 검정력을 80%로 하고 각 그룹에 동일한 수의 환자를 할당한다고 했을 때, 각 그룹당 필요한 환자수는 63명이 된다.

4.4 두 집단에서의 비율 비교

두 개의 비율을 비교하기 위해 필요한 표본크기를 결정하는 문제를 고려해 보자. p_1과 p_2가 각각 그룹 1(예컨대 대조군)과 그룹 2(예컨데 실험군)에서 일정한 시간동안 발생한 사건의 비율을 나타낸다고 하고, N명의 환자가 그룹 1에, rN명의 환자가 그룹 2에 확률화를 통해 할당되었다고 하자.

귀무가설과 대립가설이 각각

$$H_0: P_1 = P_2 \ (\delta = p_1 - p_2 = 0) \text{ vs } H_1: p_1 > p_2 \ (\delta = p_1 - p_2 > 0)$$

일 때 유의 수준을 α, 검정력을 $1-\beta$로 그리고 비율의 차이를 α라고 하면, 연속보정(continuity)이 없는 표본크기공식은

$$N = \frac{\{Z_{\alpha}\sqrt{(r+1)\bar{p}(1-\bar{p})} + Z_{\beta}\sqrt{rp_1(1-p_1)+p_2(1-p_2)}\}^2}{r\delta^2}$$

와 같이 된다(cf. Fleiss, 1973). 여기서 $\bar{p}=(p_1+rp_2)/(r+1)$이고, Z_{α}는 유의 수준 α에 해당하는 표준정규분포의 임계치이다. 만일 양측 검정이 실시되면(즉, $H_1: p_1 \neq p_2$일 때), Z_{α} 대신 $Z_{\alpha/2}$를 대입한다.

이제 연구대상자 수를 계산하는 예를 들어보기로 하자. 여아에 대해 어머니의 유아영양(infant feeding)습관을 조사하기 위해 대조군을 남아로 하고 실험군을 여아로 하여 인공영양(artifical feeding)률을 조사한다고 하자. 대조군의 반응률이 $0.5(=p_1)$이고, 실험군에서의 반응율이 $0.3(=p_2)$일 때 실험군과 대조군의 반응률이 동일한지를 검정하려 한다. 유의수준 α를 5%, 검정력 $1-\beta$를 80%로 하고 양측 검정을 실시한다고 하자. 이때 귀무가설 H_0과 대립가설 H_1는 각각

$$H0: p_1 = p_2 \text{ vs } H_1: p_1 \neq p_2$$

이 된다. 즉, 각 그룹에 93명의 연구대상자가 필요하게 되며, 따라서 필요한 총 연구대상자수는 186명이 된다.

일반적으로 위의 공식은 실제크기에 가까운 근사값을 주지만, 특히 표본크기가 작을 때에는 연속보정(continuity correction)이 있다는 다음과 같은 공식

$$N' = \frac{N}{4}\left(1+\sqrt{1+\frac{2(r+1)}{rN\delta}}\right)^2$$

이 더욱 좋은 근사값을 제공한다(cd fleiss et al., 1980). 두 그룹에 같은 수의 환자가 할당되는 경우에는 (즉, r=1일 때), 이 공식은 Casagrande et al.(1978a, 1978b)의 공식과 같게된다. 또한 각 그룹으로부터 100P%의 환자가 다른 곳으로 이동하는 등의 이유로 추적 불가능하다면(loss to follow up), 같은 검정력을 얻기위해서 필요한 환자의 수는 N/(1-P)가 된다.

이제 환자그룹이 성별, 나이, 건강 상태 등의 인자들에 근거해서 J개의 서로 다른 위험층(risk stratum)으로 구분되는 경우를 생각해 보자. 전체 환자중의 $100f_j$%의 환자가 j번째 층에 할당되어 있다고 하자. jqjsWo(j=1,…,J)층에서 환자들이 두 그룹에 같은 확률로 임의 할당된다고 하고, p_{1j}와 p_{2j}가 각각 그룹 1과 그룹 2에서 발생하는 사건비율(evebt rate)이라고 가정한다. 또한 각층에서 오즈비(odds rate) Δ가 일정하다고 하자, 즉 모든 j에 대해서,

$$\Delta = p_{1j}(1-p_{2j}) / p_{2j}(1-p_{1j})$$

를 만족한다고 하자. 귀무가설과 대립가설을 각각

$$H0: \Delta = 1 \text{ vs } H1: \Delta > 1$$

으로 했을 때, 유의 수준을 α, 검정력을 $1-\beta$로 하면, 각 그룹당 필요한 표본크기는

$$N = \frac{2(Z_{\alpha}+Z_{\beta})^2}{\sum_{j=1}^{J} g_j f_j}$$

와 같이(cf. Gail, 1973). 여기서

$$g_i = \frac{(\log \Delta)^2}{1/[p_{1j}(1-p_{1j}] + 1/[p_{2j}(1-p_{2j}]}$$

이다. 이 공식은 두 그룹에 같은 수의 환자를 할당할 때만 유용하다. Munoz와 Rosner(1984)는 다른 수의 환자를 할당할 때 적용할 수 있는 공식을 유도하였다. 위의 공식보다 훨씬 복잡하므로, 여기에서는 논의하지 않겠다.

4.5. 상관계수

두 확률변수 사이의 선형관계를 측정하는 단위로 상관계수(correlation coefficient)를 들 수 있는데, 만약 두 확률 변수, X, Y의 모 집단 상관계수가 0인지에 대해 알아보기 위해 귀무가설로 상관계수가 ρ가 0으로 하고, 대립가설을 0이 아닌 것으로 하면 가설은

$$H_0 : \rho_0 = 0 \text{ vs } H_1 : \rho = \rho_1(\neq 0)$$

로 표현할 수 있고, 표본 상관계수 r에 대한 함수인 검정 통계량

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r}}$$

는 자유도가 n-2인 t-분포를 따르고, 위의 식을 r에 대해 정리하면

$$r = \sqrt{\frac{T^2}{T^2 + (n-2)}}$$

이 된다. 위의 표본 상관계수를 피셔(Fisher)의 Z-변환을 통해서 다시 쓰면(cf. Pearcon & Hartley, 1972),

$$W = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

은 평균이 $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$이고 분산이 $\frac{1}{(n-3)}$인 정규분포를 따르게 된다.

대립가설하에서의 모집단 상관계수를 $\rho_{ES}(=\rho_1)$라고 할 때, 위의 변환 공식에 의한 결과를 W_{ES}라 하고, 2절에서의 가설 검정 개념을 적용하면 표본크기는

$$n = \left(\frac{Z_\beta + Z_{\frac{\alpha}{2}}}{W_{ES}}\right)^2 + 3$$

이 된다. 만약 단측검정일 경우 $Z_{\frac{\alpha}{2}}$를 Z_α로 대치시키면 될 것이다.

한 예를 들어보자. 한 연구자는 외향성(exteraversion)에 대한 신경생리학적 측정기준과 질문지의 응답 점수간에 연관성을 알아보기 위한 실험을 하려고 한다. 연구자가 유효크기를 0.3으로 유의수준을 0.05로 그리고 80%의 검정력을 기대한다면, 이에 근거해서 $Z_{\frac{\alpha}{2}}$는 1.96이고, Z_α는 0.84이며,

$$W_{ES} = \frac{1}{2}\ln\left(\frac{1+0.3}{1-0.3}\right) = 0.31$$ 이다. 따라서

$$n = \left(\frac{1.96+0.84}{0.31}\right)^2 + 3 = 84.58$$

로 약 85명이 필요하다.

만일 가설이

$$H_0 : \rho = \rho_0 \text{ vs } H_1 : \rho = \rho_1(\neq \rho_0)$$

처럼 귀무가설의 모상관계가 0이 아닌 다른 값일 경우에는 앞의 표본크기 결정과 조금 달라진다. 두 상관계수의 차이를 $q = |\rho_0 - \rho_1|$로 정의하고, ρ_0, ρ_1에 대한 피셔의 Z-변환 수치 W_0, W_1를 계산하여, 그차이를 $Q = W_0 - W_1$이라 하면 표본크기는

$$n = 2\left(\frac{Z_{\frac{\alpha}{2}} + Z_\beta}{Q}\right)^2 + 3$$

와 같이 결정된다.

4.6. 분산분석

두 집단의 평균을 비교하기 위한 경우와 다르게 집단의 수가 $k(\geq 3)$개 이상일 때 평균을 비교하는 경

우에는 표본크기의 결정문제가 훨씬 복잡해진다. 각 집단별로 n개의 개체로 이루어진 k개의 집단 집단을 고려할 때, $\mu_1, \cdots, \mu_k$는 각 집단별 평균이고, σ^2은 공통분산이며, 각각의 집단별 개체들은 독립적으로 정규분포를 따른다고 하자. $\mu_1, \cdots, \mu_k$ 가 모두 같은지에 대해 검정하기 위해

$H_0: =\mu_1=\cdots=\mu_k$ vs H_1: 평균이 모두 같지는 않다. 와 같은 가설을 세운다. 이 때 이를 검정하기 위한 검정통계량

$$F = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{i*} - x_{**})^2/k-1}{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - x_{i*})^2/(n-1)}$$

은 귀무가설하에서 중심 F-분포(central F-distributtion)를 따르게 된다. 중심 F-분포는 위의 식의 분자항인 집단간 평균제곱합(between sum of mean square)에 대한 자유도 $v_1=k-1$과 분모항인 집단내 평균제곱합(within sum of mean square)에 대한 자유도 $v_2=k(n-1)$을 모수로 갖는다.

만약 귀무가설이 옳지 않을 경우, 위의 검정통계량은 비중심 F-분포(noncentral F-distributton)를 따르는데, 비중심 F-분포는 집단내 평균제곱합에 대한 자유도와 집단간 평균제곱합에 대한 자유도 외 비중심모수(noncentrality parameter)

$$\delta = \frac{\sqrt{n\sum_{i=1}^{k}(\mu 1 - \mu 2)^2}}{\delta}$$

를 갖는다. 여기서 $\overline{\mu} = \sum_{i=1}^{k}\mu_i/k$이다. 만약 각 집단별 평균이 모두 같으면 비중심 모수가 0이 되므로 중심 F-분포가 된다.

만약 검정력(1-β)만큼의 통계적 유의성을 가질 정도로 각 집단별로 평균들의 차이가 정해지고, 공통분산 σ^2과 유의수준 α가 정해졌다고 하자. $F_{v1,v2,\alpha}$ 가 유의 수준α일 때의 중심 F-분포의 임계치라고 하고 , $F^*_{v1,v2,\delta}$ 를 비중심 모수가 δ이고, 자유도가 각각 v1,v2인 비중심 F-분포의 임계치라고 하면, 검정력은

$$\Pr\ [\ F^*_{v1,v2,\delta} > F^*_{v1,v2,\alpha}\ ] = 1 - \beta$$

으로 표현할 수 있다(cf. Fleiss, 1986).

여기서 자유도 v_2와 비중심모수 δ가 표본크기 n의 함수이기 때문에, 자유도 v_1, 유의수준 α 그리고 제 2종 오류 β가 주어지면, 표본크기 n을 결정할 수 있다.

표본크기 n의 결정에 있어서 계산의 편의를 위하여 Laubscher(1960)이 제안한 비중심 F-분포로부터 정규분포의 근사를 이용하면,

$$Z = \frac{\sqrt{v_1(2v_2-1)F^*_{v1,v2,v3}} + \sqrt{2(v_1+\delta^2) - \frac{v_1+2\delta^2}{v_1+\delta^2}}}{\sqrt{\frac{v_1F^*_{v1,v2,\delta}}{v_2} + \frac{v_1+2\delta^2}{v_1+\delta^2}}}$$

는 근사적으로 표준정규분포를 다르게 된다. 이 근사분포를 이용하여 β가 주어졌을 때 표준정규분포의 단측상한의 임계치는

$$Z_\beta = \frac{\sqrt{v_2[2(v_1+\delta^2)^2 - (v_1+2\delta^2)]} - \sqrt{v_1(v_1+\delta^2(2v_2-1)F^*_{v1,v2,\alpha}}}{\sqrt{v_1(v_1+\delta^2)F^*_{v1,v2,\alpha} + v_2(v_1+2\delta^2}}$$

와 같이 계산된다. 집단이 k개 있다고 하자. 집단간과 집단내의 분산비를

$$f = \frac{\sigma^2 m}{\sigma^2}$$

으로 정의하자. Cohen(1977)은 f를 유효크기를 정의하였다. 여기서

$\sigma^2{}_m = \sum_{i=1}^{k} (\mu i - \mu)^2/(k-1)$이 라고 할 때(Flesiss, 1986), 분산비 f와 비중심모수 δ 간에는

$$\delta^2 = n(k-1)F$$

의 관계가 성립되고, 이 관계식을 Z_β 공식에 대입시키면

$$Z_\beta = \frac{1}{\sqrt{(k-1)(1+nf)F_{v1,v2,\alpha} + k(n-1)(1+2nf)}}$$
$$\times \sqrt{k(n-1)[2(k-1)(1+nf)^2 - (1+2nf)]}$$
$$- \sqrt{F_{v1,v2,\alpha}(k-1)(1+nf)(2k(n-1)-1)}$$

으로 표현된다.

위의 공식은 컴퓨터 프로그램을 통해 구해질 수 있는데. 이를 위해서는 $F^*_{v1,v2,\alpha}$의 수치를 알아야 한다. $F^*_{v1,v2,\alpha}$의 수치는 Paulsong(1942)에 의해 제안된 근사공식

$$F^*_{v1,v2,\alpha} = \frac{k^3(n-1)^3}{(k-1)^3([9k(n-1)-2]^2 - 18Z^2_\alpha k(n-1)^3)}$$
$$\times \{(9k-11)(9k(n-1)+3\sqrt{2}Z_\alpha$$
$$\times \sqrt{(k-1)(9k(n-1)-2)^2 + k(n-1)(9k-11)^2 - 18Z^2_\alpha k(k-1)(n-1)}\}^3$$

으로 계산된다.

집단수가 4개인 경우의 예를 들어보자. 유의수준 0.05, 검정력이 80%일 때 각각의 집단별 평균이 9.775, 12.000, 12.000, 14.225이라 하자. 각 집단별에서 표준편차가 3이라고 하고, 80% 검정력을 고려한다면 분산비의 분자항은

$$\delta^2 m = (9.775 - 12.000)^2 + \cdots + (14.225 - 12.000)^2/(4-1) = 3.300$$

으로 계산되고, 따라서 분산비는

$$f = \frac{3.300}{3^2} = 0.367$$

이고, F-분포의 분산표를 찾아보면 표본크기 n에 따른 $F_{v1,v2,\alpha}$와 근사한 Z_β를 얻을 수 있을 것이다. 이를 정리하면 <표4>와 같다.

<표1> 표본크기에 따른 $F_{v1,v2,\alpha}$와 Z_β 값

표본크기 n	$F_{v1,v2,\alpha}$	Z_β
10	$F_{3,36,0.05} \approx 2.85$	0.712
11	$F_{3,40,0.05} \approx 2.83$	0.873
12	$F_{3,44,0.05} \approx 2.80$	1.027

위의 표에서 $Z_\beta = Z_{0.20} = 0.84$를 넘는 가장 가까운 Z_β의 값은 표본크기 n=11일 때이다. 따라서 각 집단별로 11명이 필요하고 전체 44명의 연구대상자가 필요하게 된다.

4.7. 기타연구

앞 절에서는 비교적 단순한 형태의 연구에서 표본크기를 결정하는 방법을 살펴보았다. 하지만 간호학 및 의약학 연구에서는 다중회귀분석, r×c 분할표(contingency table), 동등성 검정, 변화율 비교, 생존

분포의 비교와 같은 복잡한 임상시험을 실시하는 경우가 많다. 이 절에서는 이와 같은 형태의 연구에서 표본크기를 결정하는 방법을 제안한 연구에 대해서 살펴보기로 한다. 보다 자세한 내용은 박미라 · 이재원(1996, 2장)을 참조하기 바란다.

1) 회귀분석

하나의 반응변수에 영향을 미치는 독립변수들과의 관계가 선형식으로 이루어진 경우가 있다. 예를 들어, 국소좌반구(Focal Left Hemoshere)장애 환자의 지능에 실어증과 운동신경장애가 어떠한 영향을 미치는 가를 알아보기 위해, 지능을 반응변수로 하고 구조적 운동신경장애 점수와 뇌손상정도 그리고 실어증의 심각한 정도를 독립변수에 의한 반응변수의 설명정도를 평가하는 측도 중의 하나가 결정계수(coefficient of determination)인데, 만약 결정계수가 높다면 주어진 독립변숙 반응변수를 잘 설명한다고 할 수 있다. 4.6절과 유사하게 F-분포를 이용한 분산비 검정을 통하여 연구자가 선택한 독립변수들을 사용하였을 때 원하는 정도의 결정계수 값을 얻을 수 있는지를 검정한다. Cohen(1988)은 이에 필요한 표본크기를 구하는 공식을 제안하였다.

2) r×c 분할표

간호학 및 의약학 분야게서의 연구에서 r개의 처리집단과 이에 따른 c가지 결과를 갖는 형태를 자료를 흔히 볼 수 있다. 예를 들어, 알츠아이머병(Alzheimer's disease)을 앓고 있는 환자를 대상으로 냉수욕 치료 효과에 관한 연구를 한다고 하자. 환자를 세 집단으로 나누어 한 집단에는 냉수욕 치료를 하지 않고, 한 집단은 하루 1회, 나머지 한집단은 하루 3회의 냉수욕을 한다고 하자. 일정한 시간이 지난 후에 각 환자들의 상태가 호전, 조금 호전, 효과 없음의 세가지 결과를 갖는다고 하자. 이런 경우에 세 개의 집단으로부터 세가지 결과를 얻으므로, 자료를 3×3 분할표(contingency table)형태로 표현할 수 있고, 카이제곱 검정별(chi-square test)을 사용해서 처리집단간에 결과의 차이가 있는 가를 검정할수 있게 된다. 이러한 형태의 연구에서 Lachin(1977)은 비중심 카이제곱 분포이론에 근거하여 필요한 표본크기를 구하는 방법을 제안하였다.

3) 동등성 비교

어느 질병에 대해서 좋은 효과가 있는 기존의 처리가 있다고 하자. 새로 개발된 처리방법이 기존의 처리(standard treatment)보다 훨씬 독성이 없고 부작용이 적거나, 또는 시행하기 쉽다면, 연구자는 이 새로운 처리의 효과가 기존의 처리의 효과에 못지 않다는 것을 보이고 싶을 것이다. p1과 p2를 각각 기존의 피료법과 새로운 치료법에서의 성공률이라고 하자. 두 효과의 차이가 δ 이하일 때, 두 치료법의 효과는 동등한(equivalent)것으로 본다고 하자. 이러한 경우에 두 치료법의 효과가 동등하다는 것을 대립가설로 놓고, 새로운 치료법의 효과가 기존 치료법보다 못하다는 것을 귀무가설로 놓게 된다. 즉,

$$H_0 : p_1 - p_2 > \delta \quad vs \quad H_1 : p_1 - p_2 \leq \delta$$

이 된다. 일반적으로 동등성 검정에는 단측검정을 사용하는데, 여기서는 새로운 치료법이 기존 치료법보다 더 좋다는 것을 보이는 것에는 관심이 없기 때문이다. 두 효과의 차이가 유의하지 않다는 것이 효과가 동등하다는 것을 의미하는 것은 아니기 때문에 앞절에서 유도된 공식은 동등성 검정에는 적절하지 않다. Donner(1984)는 처리효과에 동등성 검정을 위해 필요한 표본크기를 구하는 공식을 제안하였고, 이외에 축차적으로 동등성검정을 보이는 연구에서 표본크기를 구하는 방법도 개발되고 있다(cf. Durrlement and Simon, 1990).

4) 변화율 비교

4.3절에서는 평균수준의 변화를 정의할 때 단지 연구시작과 종료 두 시점만을 고려하였다. 예를 들면, 환자가 여러 차례 병원을 방문할 경우 가장 마지막 방문시 측정한 수준과 처음 방문시 측정한 수준과의 차이만을 고려한 것이다. 하지만 환자가 병원을 방문할 때마다 측정하는 경우가 많으며, 이러한 경우에 연구자는 두 집단간의 평균변화율에 차이가 있는가를 알고 싶을 것이다. 가능한 방법중의 하나는 측정

치가 시간에 대해 대략 선행적으로 변화한다고 가정하고, 변화율을 그 직선의 기울기로 표시하는 것이다. 각 환자로부터의 측정치에 이러한 회귀모형을 적용시키고 최소 제곱 추정법(least square method)에 의해 기울기를 추정한다. 이러한 인구를 계획하는데 있어서, 연구자는 환자가 얼마나 자주 또한 얼마나 오랫동안 병원을 방문해야 하는가를 고려해 보아야 한다. 이런 연구 형태에 대해 Schelesslman(1973a, 1973b)은 시간에 따른 집단간 평균변화율 차이를 알아보기 위한 표본크기를 구하는 공식을 제안하였다.

5) 생존분포의 비교

임상시험에서는 환자의 사망이나 질병의 재발(relapse)등과 같은 실패(gailure)가 발생할 때까지의 시간(즉, 생존시간 또는 실패시간 등)을 비교하는 것이 주요 관심사인 경우가 많다. 이러한 생존분포(survival distribution)를 비교하는 경우에 필요한 대상환자 수를 결정하는 방법에 관한 많은 연구가 있어 왔다. 박미라, 김선우, 이재원(1997)은 생존분포를 비교하기 위해서 각각의 방법에 대한 장단점을 비교, 정리하였다.

Ⅴ. 사례연구

서론에서 언급하였듯이 검정력 분석은 임상시험 연구의 성패를 가를 수 있는 중요한 사항으로, 연구시작전에 표본크기를 결정하기 위하여 검정력 분석과정을 수행할 수 있고, 그렇지 못한 경우 연구분석 후에 연구결과를 정당화하기 위해서 검정력 분석이 행해지는 경우가 있다. 이 절에서는 두 임상시험연구에 대해서 검정력 분석을 중심으로 살펴보기로 한다.

5.1. 사전 검정력분석

Cussion, Madonia와 Taekman(1997)은 신생아에 대한 정확한 체온 측정이 체온 불균형으로 인한 위험을 방지하는데 중요한 것으로 인식하고 63명의 신생아에 대해서 각 신체부위(고막(tympanic), 직장(rectal), 서혜부(inguinal), 겨드랑이(axillaray))의 체온에 미치는 주위환경(incubator, radiant warmer, bassinet)의 효과에 대해서 조사하였다. 이를 의해 연구자는 ① 양쪽 고막의 온도를 측정하기 위해 paired t-test를 수행하였고, ② 각 부위별 온도간의 차이를 알아보기 위해 상관관계를 조사하였으며, ③ 각 신체부위별, 집단(주위환경)별 반복측정 분산북석을 실시하였다.

Cussion et al.(1997)은 연구개시전 연구대상자 수를 결정하기 위해 검정력 분석을 실시하였고 검정력 분석결과에 따라 60명의 연구대상자가 필요한 것을 알게 되었다. 다음은 논문 중에서 검정력분석을 논의한 부분을 발췌한 것이다.

‘ A power analysis was performed before data collection to aid in determining an appropriate sample size. A medium effect size was used in the computation of the power analysis indicating that 60 subjects were needed to reduce Type Ⅱ error(Lipsey, 1990). '

연구자는 유효크기를 중간(medium)으로 하여 표본크기를 결정한 것으로 보고하고 있으나 3절에서 언급했듯이 Polit and Sherman(1990)이 제안한 유효크기 결정 방법중에서 어떤 것인지에 대해서는 명확히 알 수 없었다. 유효크기를 중간으로 선택한 근거가 있었다면 더욱 좋았을 것이다.

5.2. 사후검정력분석

Pasacreta(1997)는 우울증의 특징과 범위에 대해서 그리고 Functional staus의 결과와 depressive symptom 및 physical symptom distress의 연관성에 대해 알아보기 위해 3-7개월전 유방암 진단을 받은 여성을 79명을 연구대상자로 하여 임상시험을 진행하였다. 이러한 연구목적을 위해 연구자는 ①인구통계학적 분석을 위해 빈도수 및 백분율을 조사하였고, ② depressive symptom과 functional status의 관계를 알아보기 위하여 상관계수를 구하였고, ③ 이 연구의 주요한 목적으로 우울증 정도에

따른 functional status의 차이를 알아보기 위해 분산분석을 실시하였으며, ④ functional status에 대한 depressive stmptom, symptom distress 및 임상적 변수와의 관계를 알아보기 위해 회귀분석 실시하였고, ⑤ 마지막으로 화학치료법 실시 유무에 따른 우울증 정도 비교를 위해 t-test를 실시하였다.

Pasacreta(1997)의 연구에서는 연구 개시전 연구대상자 수의 결정에 대한 언급이 없었으나 임상시험의 주목적인 우울증의 정도에 따른 functional status의 차이를 알아보기 위한 분석분석 후 연구결과의 정당성을 입증하기 위하여 사후 검정력 분석을 실시하였다. 연구자는 우울증 정도에 따라 세 집단(집단 1: depressive syndrome, 7명, 평균 20.1; 집단 2; depressive symptom, 14, 평균 20.1; 집단 3; 우울증 증상 없은, 58명, 평균 167.1)으로 나누고 분산분석 결과 $F(2.78)=8.7$, $p<0.001$를 얻었다. 즉 집단간 functional status에 유의한 차이가 있음을 알 수 있었고, 이를 정당화 하기 위해 검정력분석 결과를 다음과 같이 보고하고 있다.

‘For ANOVA, a power of 94 was reached to detect an effect size of .43 with an alpha set al .05. This errect size if considered large by Cohen(1988)'

유의 수준 5%에서의 검정력 94%는 4절 연구형태별 표본크기의 결정에서 표본크기의 결정에서 표본크기 공식을 Z_β로 정리하여 얻을 수 있는데, 위의 예의 경우 각 집단별 표본크기가 다르기 때문에 유효크기 계산시 이에 대한 보정이 필요하다. 보정 방법에 대한 자세한 내용은 Cohen(1988, 8장)을 참조하기 바란다.

또한 연구자는 functional status를 종속변수로 하고 depressive sysptom, symptom distress 및 임상적 변수를 독립변수로 하여 회귀분석을 실시하여, stepwise 변수선택법에 의해 독립변수를 depressive symptom와 symptom distress로 하는 최종모형으로 선택하여 $R^2=0.346$을 얻었다. 이에 대해서도 연구자는 검정력분석을 통해 99%의 검정력을 얻었음을 보고 하고 있다.

참 고 문 헌

1. 박미라. 김선우, 이재원(1997). 생존함수의 비교연구를 위한 표본수의 결정. 응용통계연구, 제개 예정

2. 박미라, 이재원(1992). 임상시험 연구를 위한 통계적 방법. 서울: 자유아카데미.

3. Casagrande, J. T., Pike, M. C., & Smith, P. G. (1978a). An improved approximate formula for calculating sample sixe for comparing two binomial distributions. Biometrices, 34, 483-486.

4. Cohen, J. (1977). Statistical power analysis for the behavioral sciences(rev ed). New York: Academic Press.

5. Cohen, J. (1988). Statistical power analysis for the behavioral sciences(2nd ed). New York: Academic Press.

6. Cussion, F. M., Madonia, J. A., & Taekman, J. B. (1997). The effect of environment on body site temperatures in full-term neonates. Nursing Research, 46(4), 202-207.

7. Donner, A. (1984). Approaches to sample size edtimation in the design of clinical trials-A review. Statistics in Medicin3, 3, 199-214.

8. Fleiss, J. L.(1973). Statistical Methods for Rates and Proportions. New York; John wiey and Sons.

9. Fleiss, J. L., Tytun, A., & Ury, H. K. (1980). A simple approximation for calculation sample size for comparing independent proportions. Biometrics, 36, 343-346.

10. Fleiss, J. L. (1986). The design and analysis of clinical experiments. New York: John Wiley and Sons.
11. Gail, M. (1973). The determination of the number of partients requored in a preliminary and a follow-up trial of a new chemotherapeutic agent. Journal of Chronic Disease, 13, 346-353.
12. Jennison, C., & Tunnbull, B. W. (1989). Interim analysis: the repeated confidence interval approach. Journal of the Royal Starisrical Society, Ser. B. 51, 305-361.
13. Kim, K., & Demets, D. L.(1987). Design and analysis of group sequential tests based on the Type Ⅰ error spending rate function. Biometrika, 74, 149-154.
14. Lachin, J. M.(1997). Sample size determinations for r×c comparative trials. Biometrics, 33, 315-324.
15. Laubscher, N. F.(1960). Normalizing the noncentral t and F-distributions. Ann. Math. Stat., 31. 1105-1112.
16. Mosteller, F., & Bush, R. R.(1954). Selected quantitative techniqyes In G. Lindzey(Ed.), In Handbook of social psychology(289-334). Cambridge: Addison-Wesley.
17. Munoz, A., & Rosner, B.(1984). Power and sample size for a collection of 2×2 tables. Biometrics, 40, 995-1004.
18. Pasacreta, J. V. (1997). Depressive Phenomena, Physical Symptom Dostressm and Functional Status Among Women With Breast Cancer. Nursing Research, 46(4), 214-221.
19. Paulson, E. (1942). An approximate normalization of the analysis of variance distribution. Ann. Math. Stat., 13, 233-235.
20. Pearson, E. S., & Hartley, H. O. (1972). Biometrika tables for statisticians 2. Lodon: Cambridege University Press.
21. Polity, D. F. and Sherman, R. E. (1990). Statistical power in nursing research. Nursing Research, 39, 365-369.
22. Schlesselman, J. J. (1973b). Planning a longitudinal study: Ⅱ. Frequency of measurment and study duration. Journal of Chronic Disease, 26, 561-570.

-Abstract-

Key concept: Power analysis, Sample size determination, Effect size

A Review on the Methods of Sample Size Determination in Nursing Reaserch

Lee, Jae Won · Park, Mi Ra
Lee, Jung Nok · Lee, sook Ja
Park, Eun, Sook · Park, Young Joo

Lee, Jae Won · Lee, Jung Nok:Department of Statics, Korea University Anam-Dong 5 Ga 1,Sungbuk-Gu, seoul, Korea
TEL:(02)3290-2237 FAX:(02)924-9895 E-mail:jael@kustat.korea,ac,kr
Park, Mi Ra:Department of Premedicine, Ulji Medical College
Lee, sook Ja · Park, Eun, Sook · Park, Young Joo:College of nursing, Korea University

In clinical trials of nursing research, the sample size determination is one of the most important factor. Although sample size must be considered at the design stage, it has been disregarded in most clinical trials. The power analysis is usually performed before study begins to compute sample size and the power can also be calculated at the end of study in order to justify study result. The power analysis is essential especially when the clinical trials can not show significant differences.
In this paper, we review the statistical methods for power analysis and sample size formulae in nursing research. Sample size formulae and the corresponding examples are discussed according to the six types of studies ; mean for one sample, proportion for one sample, means in two samples, proportions in two samples, correlation coefficient and ANOVA.