

4

치의학 연구에서 노모그램을 이용한 표본수 계산

연세대학교 치과대학 BK21 플러스 통합구강생명과학 사업단, 치의학 교육개발실
김기열

ABSTRACT

Sample size estimation using nomogram in dental research

BK21 PLUS Project, Institute for Dental Education, Yonsei University College of Dentistry
Ki-Yeol Kim, Ph.D.

The appropriate sample size calculation in dental research is important to achieve the study purpose at the first step in study design. However, it cannot be easy to calculate sample size using standard formulas, because the several factors must be considered for calculation. This study introduced the graphic method for sample size calculation, which is called nomogram. The purpose of this study is to show the effectiveness of the nomogram using examples, expecting the researchers can easily use nomogram for sample size determination.

Key words : sample size; dental research; nomogram

Corresponding Author

Ki-Yeol Kim, Ph.D.

BK21 PLUS Project, Institute for Dental Education, Yonsei University College of Dentistry

250 Seongsanno, Seodaemun-gu, Seoul 120-752, Korea

Tel : +82 2 2228 3039, Fax : +82 2 392 2959, E-mail : kky1004@yuhs.ac

본 연구는 연세대학교 치과대학 2015년도 교수연구비에 의하여 이루어졌음(6-2015-0003).

I. 서론

모든 임상연구는 전체 대상의 일부인 표본으로부터 얻은 결과를 이용하여 전체 모집단에 대하여 일반화한

결론을 내리게 되며, 표본으로부터 발견한 결과가 진실인지 혹은 우연히 그렇게 나타난 것인지는 통계적 추론에 의해 결정하게 된다.

연구진행시 randomization, blinding 등의 기법

을 사용하여 bias 를 감소시킴으로써 연구의 결과가 우연히 나타난 것일 가능성을 줄일 수 있다. 그러나, 연구결과가 틀릴 수도 있는 가능성에 영향을 주는 또 다른 요인이 있는데, 그것이 바로 연구에 포함되는 환자의 수이다.

과학적 연구의 결론은 표본크기에 많이 영향을 받는다. 너무 적은 표본은 의미있는 발견을 할 기회와 연구 결과의 일반화 할 가능성을 감소시키며, 너무 큰 표본은 임상적으로 의미가 없는 작은 차이를 통계적으로 큰 의미가 있다고 결론을 내리기도 한다.

연구에 포함되는 환자의 수, 즉 표본의 크기가 클수록 전체 모집단에 대한 더 정확한 답을 얻을 수 있을 거라는 것은 누구나 예상할 수 있을 것이다. 그러나 너무 큰 표본수를 사용한 연구는 불필요한 추가적인 표본과 비용의 증가에 의해 비도덕적으로 간주될 수도 있다¹⁾. 반대로 너무 작은 표본수를 사용한 연구는 임상적으로 중요한 효과를 찾을 수 없을 것이다. 이러한 연구도 과학적으로 무익한 연구, 비도덕적인 연구가 되는 것이다. 그렇다면 옳은 결과를 얻기 위해서는 얼마나 많은 표본수가 필요한 것일까? 표본크기의 결정은 연구설계를 하는데 있어서 중요한 단계이다. 최소한의 필요한 표본수를 결정하는 것은 도덕적, 경제적뿐만 아니라 과학적이고 통계적으로 의미있는 결과를 얻는데 절대적으로 중요하다. 표본크기의 결정은 연구를 시작하기 전에 결정되어야 한다.

본 연구는 치의학분야에서 연구를 계획할 때 그래픽

방법 - 노모그램(nomogram) - 을 이용하여 필요한 표본의 크기를 간편하게 계산하는 방법에 대해 소개한다. 표본수 계산을 위한 노모그램은 Altman DG (1982) 에 의해 소개되었으며, 그림을 이용하여 편리하게 표본수를 계산할 수 있다. 이것은 두 집단간 비교와 각 집단의 표본수가 동일한 경우에 검정력을 계산하기 적절한 형태로 작성되었다²⁾. 노모그램 활용의 예로 사용할 통계적 분석방법은 두 집단간 비율비교, 두 집단간 평균 비교, 대응표본 비교이다.

II. 연구가설의 설정과 오류의 종류

연구를 진행하기 위해서는 연구가설을 설정하게 되며 가설은 다음의 두가지 형태로 정의된다.

귀무가설(H_0): 두 집단간 평균의 차이가 없다.

대립가설(H_1): 두 집단간 평균의 차이가 있다.

이때, 실제로 귀무가설이 참인 경우 분석결과도 참으로 결론이 내려지는 경우와 그렇지 않은 경우가 발생하게 된다. 실제와 다르게 결론이 내려지는 경우에는 오류가 발생하게 되며, 두 가지의 오류가 존재하게 된다. 제 1종 오류와 제 2종 오류가 그것이다. 오류의 종류는 표 1 로 요약될 수 있다.

표 1. 통계적 오류의 종류

가설선택 \ 진실	H_0 (귀무가설) 참	H_1 (대립가설) 참
	H_0 (귀무가설) 선택	옳은결정
H_1 (대립가설) 선택	제 1종 오류 (α)*	옳은 결정 ($1-\beta$) :검정력 (power)

*제 1종 오류 (α) : 실제로 차이가 없는데 차이가 있다고 결론을 내릴 확률. H_0 가 참인데, H_1 으로 잘못 선택하는 오류 (H_0 를 기각하는 오류). 유의수준이라고도 함.

**제 2종 오류 (β) : 실제로 차이가 있는데 차이가 없다고 결론을 내릴 확률. H_1 이 참인데, H_0 로 잘못 선택하는 오류 (H_1 을 기각하는 오류)

Ⅲ. 표본의 크기를 계산할 때 고려해야 할 요인

표본의 크기를 계산하는 공식을 사용하기 위해서는 집단간에 임상적으로 의미있는 효과의 차이(difference), 데이터의 변동(표준편차, standard deviation), 유의수준(significance level, α), 검정력(power)이 결정되어야 한다. 다음은 각 요인에 대한 설명이다.

(1) 효과차이

이것은 두 군간에 임상적으로 의미있는 차이이며, 이 값은 표본크기와 반비례관계이다. 이 값이 작을수록 작은 효과의 차이를 발견하고자 하는 것이므로 표본의 크기는 커지게 된다. 효과차이는 기존에 알려진 값이나 전임상연구(preclinical study)로부터 추정하기도 한다.

(2) 표준편차(standard deviation)

표준편차는 기존의 연구 결과나 사전연구(pilot study)로부터 구할 수 있다. 표준편차가 커질수록 표본의 크기는 증가하게 된다.

(3) 유의수준(significance level)

유의수준은 연구결과 오류의 허용범위를 의미하며, 실제로는 두 그룹간에 차이가 없는데 차이가 있다는 결론을 내릴 확률이다. 대부분의 연구에서 주로 사용되는 값은 0.05 또는 0.01이다. critical result를 얻기 위해서는 유의수준은 작게 하는 것이 좋다. 유의수준이 작아질수록 표본크기는 증가한다.

(4) 검정력(power)

검정력이란 통계적 검정력(statistical power)을 말하며, 실제 그룹간에 차이가 있을 때 분석결과도 그 차이를 발견할 확률을 말한다. 예를 들면, 신약개발을

하는 연구에서 실제로 신약이 약효가 있는데 분석결과도 약효가 있다고 결론을 내리게 되는 확률이다. 검정력은 다음과 같이 정의한다.

$$\text{power} = P(\text{reject } H_0 \mid H_1 \text{ true}),$$

여기서 H_0 와 H_1 는 각각 귀무가설과 대립가설을 의미한다. P 는 확률(probability)를 의미하며, 검정력은 대립가설이 참일 때 귀무가설을 기각할 확률로 정의된다.

검정력은 옳은 결정을 내릴 확률이므로 가능한 한 높은 것이 좋다. 검정력은 표본수가 증가하면 높아지게 되나 표본수를 무한히 크게 할 수 없는 상황에서 최소한의 검정력은 주로 80%로 한다.

(5) 표준화된 효과크기(effect size)

노모그램을 사용하기 위해서는 효과차이의 표준화된 값이 필요하다. 이 값은 효과차이를 표준편차로 나눈 값이다. 이 값은 두 집단의 비율비교 시에는 $\frac{(P_1 - P_2)}{\sqrt{P(1-P)}}$, 두 집단간 평균비교 시에는 $\frac{\bar{X} - \bar{Y}}{s}$ 이다. 여기서 p_1, p_2 는 두 집단에서 관심있는 사건의 발생률, P 는 이 두 확률의 평균이며, \bar{X}, \bar{Y}, s 는 두 집단의 각 평균과 표준편차이다.

Ⅳ. 노모그램 소개와 예제를 이용한 표본수 계산

표본수는 다양한 방법에 의해 구할 수 있다. G power³⁾ 처럼 표본수를 계산해주는 소프트웨어를 사용할 수도 있고, 특정 조건에 대해 표본수를 계산해 놓은 테이블⁴⁾을 사용할 수도 있다. 이러한 방법들과 비교하여 상대적으로 간단한 방법이 그래픽 방법을 이용하는 방법이다²⁾. 그림 1은 표본수를 계산하는 노모그램을 보여준다.

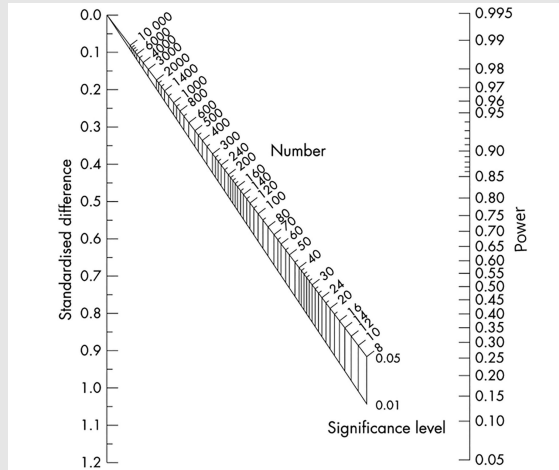


그림 1. 표본수를 계산하는 노모그램²⁾.

그림 1에서 왼편의 수직축은 비교하고자 하는 두 집단의 효과차이를 표준화한 값을 나타낸다.

이것은 두 집단간 효과의 차이를 표준편차로 나누어서 계산한다. 오른편의 수직축은 0.05 부터 0.995 까지 다양한 검정력을 보여준다. 노모그램 내에 두 개의 축은 유의수준이 0.05, 0.01 인 경우의 표본수를 나타낸다. 노모그램은 표준화된 효과차이만 계산되면 다양한 검정력과 유의수준에 대하여 표본수를 구할 수 있다. 노모그램은 비교하고자 하는 두개의 집단이 동일한 표본수를 갖는다는 가정하에 적용된다.

본 연구에서는 공식에 의한 표본수와 Altman 의 노모그램을 이용하여 구한 표본수를 비교하여 노모그램의 유용성을 보여주었다. 이를 위해 치의학 연구에서 많이 사용되는 두 집단의 비율비교, 두 집단의 평균 비교, 대응표본 비교에 대해 살펴본다.

4-1. 두 집단간의 비율 비교

[예1] 실험대조군 연구를 계획하고 있다. 교정치료를 받은 환자군과 교정치료를 받지 않은 대조군에서 잇몸이 내려앉는 증상(gingival recession)의 발생이 연관성이 있는지 평가하는 연구이다. 신뢰구간은 95% 이며, 그룹당 8명씩 할당할 사전연구로 부터 다음의 결과를 얻었다⁵⁾.

교정치료 여부에 따라 gingival recession 의 발생률이 동일한지 알아보려고 하는 연구에 적절한 표본수는 다음의 공식에 의해 계산된다.

$$n = \frac{2\bar{P} \times (1-\bar{P})(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$$

여기서, Z_{α} , Z_{β} 는 유의수준(α)와 검정력($1-\beta$)에 의해 결정된다. n 는 교정치료를 받은 집단과 받지 않은 집단

표 2. 교정치료를 받은 환자군과 교정치료를 받지 않은 대조군에서 gingival recession 발생

Exposure	Gingival recession(%)	No gingival recession	Total patients
Ortho t/t	6(75%)	2(25%)	8
No Ortho t/t	3(37%)	5(63%)	8

에서 gingival recession 이 발생한 확률의 평균이다. d 는 두 집단 gingival recession 발생률의 차이이다.

이 계산식에서 Z_{α} , Z_{β} 는 표 3을 참고하여 α , β 값에 따라 구할 수 있다.

공식에 의하여 연구에 필요한 표본수는 각 군당 36명이 필요하므로 전체 72명의 환자가 필요하게 된다. 만약 탈락율을 고려한다면 표본수는 더 늘어날 수 있다.

$$P_1=0.75, P_2=0.37$$

$$\bar{P} = \frac{(P_1+P_2)}{2} = \frac{(0.75+0.37)}{2} = 0.56$$

$$1-\bar{P}=1-0.56=0.44$$

$$Z_{0.05}=1.96, Z_{0.10}=1.282$$

$$d=0.75-0.37=0.38$$

$$\text{표본수}=(1.96+1.282)^2 \times 0.56 \times 0.44 \times 2 / (0.38)^2 = 36$$

Altman 노모그램을 이용하여 표본수를 구하는 경우에는 Z_{α} , Z_{β} 값을 필요로 하지 않는다. 왜냐하면 이 값들이 이미 노모그램에 포함되어 있기 때문이다. 다만, Altman 노모그램을 사용할 때에는 표준화된 유

효크기(effect size, standardized difference)를 사용한다. 이 값은 두 집단에서 gingival recession 발생률만 알면 계산할 수 있다.

$$\text{유효크기} : \frac{(P_1-P_2)}{\sqrt{\bar{P}(1-\bar{P})}} = 0.7656$$

유효크기가 계산되면 Altman 노모그램에서 대응되는 표본수를 찾는다.

유의수준이 5%인 경우 총 표본수는 72명이 필요하며 그룹당 36명에 해당된다. 유의수준을 1%로 하는 경우에는 총 100명, 그룹당 50명이 필요하다. 노모그램은 항상 총 표본수를 보여준다. 유의수준을 표시하는 축과 만나는 점을 표본수로 결정할 때에는 넉넉한 값을 취하는 것이 좋다.

노모그램을 이용하면 일반적인 공식에서 유의수준과 검정력에 대응되는 값을 알아야 하는 번거로움을 줄일 수 있게 된다. 노모그램 사용시 알아야 하는 유효크기에 대하여 Cohen(1988)⁴⁾는 0.2, 0.5, 0.8을 대략적으로 작은, 중간 정도, 큰 유효크기라고 정의하기도 했다.

[예2] 새로운 치료법의 성공율을 기존치료법의 성공율과 비교하기 위하여 평행설계 임상시험을 수행하는 연구이다. 기존 치료법(대조군)의 성공율=0.25, 새로

표 3. 유의수준과 검정력에 따른 Z_{α} , Z_{β} 값을 정리한 표

제1종 오류 (α)	$Z_{\alpha/2}$		제2종 오류 (β)	검정력 (power)	Z_{β}
	단측	양측			
0.005	2.576	2.813			
0.010	2.326	2.576			
0.025	1.960	2.248			
0.050	1.645	1.960	0.05	0.95	1.645
0.100	1.282	1.645	0.10	0.90	1.282
0.200	0.842	1.282	0.20	0.80	0.842
0.300	0.524	1.036			
0.400	0.253	0.842			
0.500	0.000	0.674			

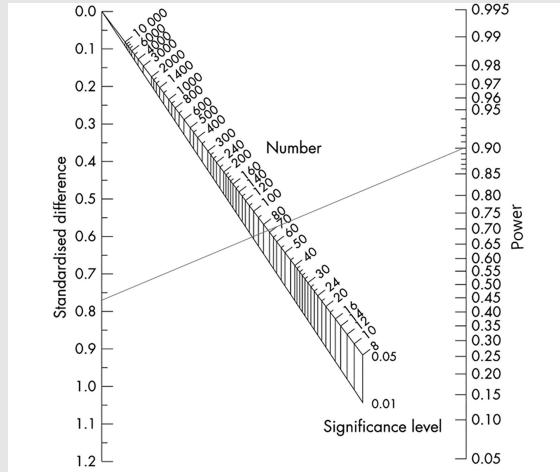


그림 2. 표본수 계산을 위한 노모그램. 표준화된 효과차이가 0.7656, 검정력이 90% 인 경우 표본수 72명이 구해진다.

은 치료법(시험군)의 성공율=0.35, 제 1종 오류 =0.05, 검정력=0.9, 군간 배정비는 1:1인 경우 표본 수는 다음과 같이 계산된다.

$$P_1=0.25, P_2=0.35$$

$$\bar{P} = \frac{(P_1+P_2)}{2} = \frac{(0.25+0.35)}{2} = 0.3$$

$$1-\bar{P}=1-0.3=0.7$$

$$Z_{0.05}=1.96, Z_{0.10}=1.282$$

$$d=0.35-0.25=0.1$$

$$\text{표본수}=(1.96+1.282)^2 \times 0.3 \times 0.7 \times 2 / (0.1)^2 = 441$$

계산된 표본수는 441명으로 총 882명의 표본이 필요하다.

이 문제에서 유효크기는 $\frac{(P_1-P_2)}{\sqrt{\bar{P}(1-\bar{P})}}$ 이며, 노모그램에서 표본수를 찾을 수 있다.

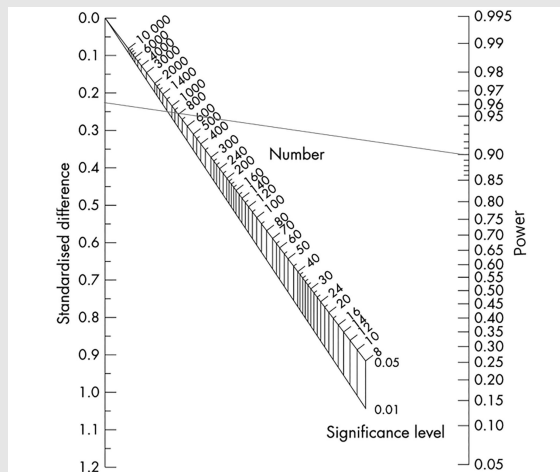


그림 3. 표본수 계산을 위한 노모그램. 표준화된 효과차이가 0.218, 검정력이 90% 인 경우 표본수 882명이 구해진다.

4-2. 두 집단간의 평균 비교시 표본수 계산

[예3] 예1 에서 실험군의 gingival recession 평균과 표준편차가 0.36mm, 0.42mm 이고, 대조군에서는 이 값들이 각각 0.22mm, 0.32mm 라고 가정해 본다⁵⁾. 이런 타입의 연구에서 요구되는 표본의 크기는 다음의 공식에 의해 계산한다.

$$n = \frac{2 \times \sigma^2 (Z_{\alpha/2} + Z_{\beta})^2}{d^2}$$

여기서, σ 는 표준편차, d 는 두 군의 평균 차이다. 공식에 의해 계산된 표본수는 그룹당 147명이다. 총 표본수는 294명 이다.

$$\begin{aligned} Z_{0.05} &= 1.96, Z_{0.10} = 1.282 \\ \sigma &= (0.42 + 0.32) / 2 = 0.37 \\ d &= 0.36 - 0.22 = 0.14 \end{aligned}$$

$$\text{표본수} = (1.96 + 1.282)^2 \times (0.37)^2 \times 2 / (0.14)^2 = 147$$

이 연구에서 유효크기는 d/σ 로 0.14/0.37 = 0.3783 이다.

유의수준 5%, 1% 인 경우 약300명, 400명이 필요하다. 이것은 그룹당 약 150명, 200명을 의미한다.

이러한 결과는 논문이나 연구계획서에 표본수 계산에 대한 서술시 “유효한 차이가 0.14, 각 그룹의 편차는 0.37, 유의수준 5%, 검정력이 80% 일때 그룹당 150 명이 필요하다” 라고 서술하면 된다.

[예4] 두 가지 치약의 효과를 비교하는 연구에서 연구자는 특정 치약이 우식경험 영구치(DMFS:DMF surfaces index) 를 다른 치약보다 평균 0.5 만큼 감소시킬 것으로 예측하고 있다. 사전 연구로부터 표준편차는 1.25라는 것을 알고 있다고 가정한다. 유의수준 0.05, 검정력 90% 에 해당하는 표본수는 각 그룹당 132명, 총 264명이 필요하게 된다⁶⁾.

$$\begin{aligned} Z_{0.05} &= 1.96, Z_{0.10} = 1.282 \\ \sigma &= 1.25 \\ d &= 0.5 \end{aligned}$$

$$\text{표본수} = (1.96 + 1.282)^2 \times (1.25)^2 \times 2 / (0.5)^2 = 131.3$$

유효크기는 0.5/1.25=0.4 이므로 노모그램에 의한 표본수는 다음과 같다.

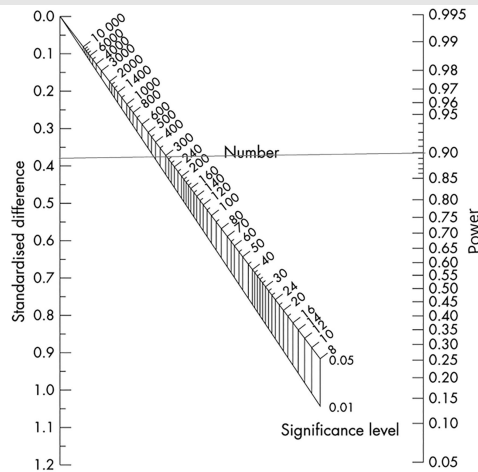


그림 4. 표본수 계산을 위한 노모그램. 표준화된 효과차이가 0.3783, 검정력이 90% 인 경우 표본수는 약 300명으로 구해진다.

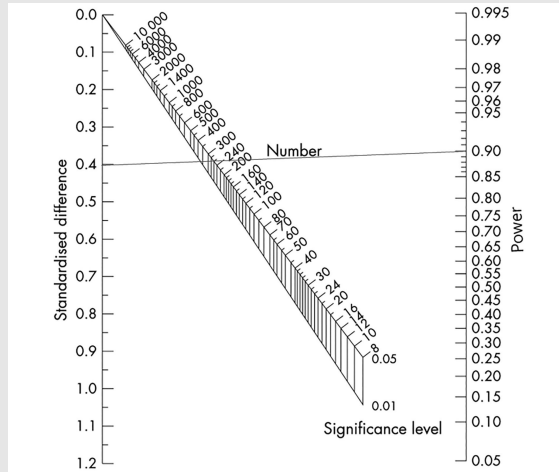


그림 5. 표본수 계산을 위한 노모그램. 표준화된 효과차이가 0.4, 검정력이 90% 인 경우 표본수 264명이 구해진다.

4-3. 두 집단간의 대응되는 평균 비교시 표본수 계산 (paired data)

대응표본 t-검정은 ‘두 개의 종속적인 집단간에 평균의 차이는 없다’ 라는 귀무가설을 검정할 때 사용되며, 필요한 표본수는 다음과 같이 계산된다.

$$n = \frac{\sigma_d^2(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$$

대응표본에 대한 분석에서 표준화된 유효한 차이는 $2d/\sigma_d$, 여기서 σ_d 는 차이의 표준편차이다.

[예5] 특정 치약의 효과를 보기 위해 사용전과 사용 일주일 후에 치태지수를 측정하는 연구를 계획한다고 가정한다. 사용전과 사용후의 유효한 차이를 0.5, 차이의 표준편차를 1 이라고 하면 80% 검정력의 경우 32명의 환자가 필요하게 된다.

$$\text{표본수} = \frac{1^2(1.96+0.842)^2}{(0.5)^2} = 31.4$$

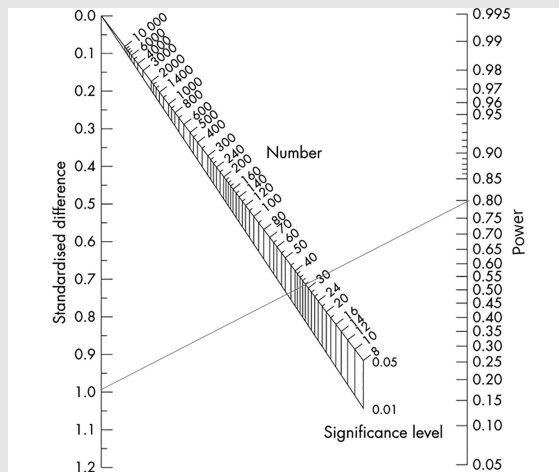


그림 6. 표본수 계산을 위한 노모그램. 표준화된 효과차이가 1, 검정력이 80% 인 경우 표본수 32명이 구해진다.

V. 결론

노모그램은 그래픽방법으로써 사용하기 쉽고 유용하기는 하지만, 비교하고자 하는 데이터의 타입과 분석방법에 대하여 확인할 필요가 있다^{1,7)}. 그러나, effect size 만 계산하면 서로 다른 검정력과 유의수준에 대한 표본수를 한번에 계산할 수 있고, 유의수준과 검정력에 대응되는 $Z_{\alpha/2}$, Z_{β} 값을 구하지 않아도 되

기 때문에 편리한 방법임에는 틀림없다.

연구자들이 표본수 계산에 필요한 요인들의 값을 정확히 알고 있는 경우가 거의 없고, 표본수 계산은 이러한 요인의 값에 아주 민감하기 때문에 표본크기 계산은 정확성이 많이 결여될 수 있다. 이런 경우에는 요인들의 변화에 표본수가 어떻게 변하는지 확인해 보고 가장 큰 표본수를 선택하는 것이 좋다⁸⁾.

참 고 문 헌

- Altman DG. Statistics and ethics in medical research. VII--Interpreting results. Br Med J 1980;281(6255):1612-4.
- Altman DG, Gore S. Statistics in practice : articles published in the British medical journal. London: British Medical Association; 1982.
- Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-91.
- Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988.
- Sandhya Jain AG, Deshraj Jain. Estimation of sample size in dental research. International Dental & Medical Journal of Advanced Research 2015;1:1-6.
- Petrie A, Bulman JS, Osborn JF, Bulman JS. Further statistics in dentistry. London: British Dental Association; 2002.
- Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20(5):453-8.
- Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant 2010;25(5):1388-93.