Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 2, pp. 411 ~ 418

Sample Size Comparison for Non-Inferiority Trials

Dong-Wook Kim[1] · Dongjae Kim[2]

Abstract

Sample size calculation is very important in clinical trials. In this paper, we propose sample size calculation method for non-inferiority trials using sample size calculation method suggested by Wang et al.(2003) based on Wilcoxon's rank sum test. Also, sample size comparison between parametric method and proposed method are presented.

Keywords : 비열등성 시험, 임상시험, 표본 수

1. 서론

신약이 개발되었을 때 기존의 약과 비교하는 임상시험을 통하여 그 유효성을 증명한다. 이러한 비교 임상시험은 제 3상 시험 단계에서 실시하며, 우월성 시험, 동등성 시험(또는 치료적 동등성 시험), 비열등성 시험 등이 있다. 현재 국내 많은 제약회사들이 신약개발이나 새로운 치료법의 효과를 증명하는 방법으로 비열등성 시험(non-inferiority trials)을 주로 시행하고 있으며, International Conference on Harmonization Guideline(ICH Guideline) E9와 E10(ICH E9, 1998; ICH E10, 1999)에도 비열등성 시험에 대해서 언급 되어있다. 비열등성 시험은 새로운 약의 치료 효과가 기존 약의 치료 효과보다 못 하지 않음을 보이는 것이 목적인 시험이다.

예를 들면, 당뇨병 치료를 위한 새로운 혈당강하제가 개발되었을 때 기존의 표준 치료제보다 통계적, 임상적으로 효과가 못하지 않음을 증명하고자할 때 비열등성 시험을 수행한다. 이 경우는 두 당뇨병 치료제의 평균 반응 값의 비교 문제가 되며, μ_S와 μ_T가 각각 기존 약제와 새 약제의 평균 반응 값이라면, 가설은 다음과 같이 주어지게 된다.

$$H_0 : \mu_S - \mu_T \geq \delta \text{ vs. } H_1 : \mu_S - \mu_T < \delta$$

1) 서울시 서초구 반포동 505번지, 가톨릭대학교 의학통계학교실, 대학원생
E-mail: dwkim0504@catholic.ac.kr
2) 교신저자 : 서울시 서초구 반포동 505번지, 가톨릭대학교 의학통계학교실, 교수
E-mail: djkim@catholic.ac.kr

여기서 δ는 실험 전에 임상적인 정보를 이용하여 결정되어지는 양의 실수이며, 비열등성의 한계라고 한다.

임상시험에 있어서 가장 중요한 사항중의 하나가 연구에 필요한 환자수(표본의 수)를 결정하는 문제이다. 아무리 완벽하고 정확한 방법을 이용하여 연구를 수행하여도 표본의 크기가 너무 작으면 그 연구의 가설을 증명하기 힘들게 된다. 반면에 표본의 크기가 필요 이상으로 크면 연구의 수행이 어렵고 과다한 경비가 들며, 또한 윤리적인 문제가 발생할 수 있다. 그러므로 적당한 크기의 표본 수를 산정하는 것은 통계적으로 유의한 연구결과의 도출을 위해서, 그리고 동시에 연구수행의 효율적인 비용을 산출하기 위해서 중요하다.

연구대상 환자수는 연구의 설계단계에서 미리 고려되어야 하는 것임에도 불구하고 많은 연구에서 이러한 원칙이 무시되고 일정기간동안에 확보할 수 있는 정도를 대상환자수로 결정해 버리는 경우가 있다. 사전에 대상환자수를 고려하지 않은 임상시험 중에는 임상적으로 의미 있는 차이가 있는데도 검정력이 떨어져서 이를 찾아내지 못하는 경우가 많다. 검정력이 떨어지는 연구의 위험성은 유효한 신약 등이 충분히 검증되지 못하고 기각되어 두 번 다시 고려될 수 없다는 데 있다.

기존의 연구나 시험에서 비열등성 시험의 표본 수를 구하기 위해 전 임상시험에서 얻은 자료들의 값을 이용하여 정규분포 한다는 가정 하에 모수적 방법을 이용하여 표본 수 계산을 하였다. 하지만 정규분포는 이론적인 분포로써, 실제로 어떤 데이터도 정밀한 정규모집단에서 얻어지는 경우는 없다. 다만 모집단의 분포가 정규분포에 가까운 경우가 많을 뿐이다. 한편 표본 통계량의 분포가 정규분포에 수렴하는 속도가 아주 느리거나, 특히 데이터들 가운데 이상점(outlier)이 있는 경우에는 표본평균에 기초한 추론은 신뢰수준이나 유의수준을 제어하기 힘들고 검정력도 급격히 떨어지게 된다.

이와 같이 모집단에 대하여 구체적인 분포함수를 가정하는 것이 무리일 때에는 모집단분포에 대한 가정을 약화시키는 것이 오류의 가능성을 줄이고 효율도 높일 수 있는 대안이 된다. 본 논문에서는 이러한 문제점을 보완하기 위해 Wang등(2003)이 제안한 비모수적 방법의 Wilcoxon 순위 합 검정을 이용한 표본 수 공식을 이용하여 비열등성 시험을 위한 표본 수 계산방법을 제안한다.

비열등성 시험에서의 비모수적 표본 수 계산방법을 2절에서 제안하고 모수적 계산방법을 설명한다. 3절에서는 모수적 방법과 제안된 비모수적 방법을 비교하기 위한 상황을 설정한다. 마지막으로 4절에서는 모수적 방법과 비모수적 방법의 표본 수 계산 결과를 비교하고 결론을 제시한다.

2. 비열등성 시험에서의 표본 수 계산방법

확률표본 $X_1, X_2, \cdots, X_{n_1}$과 $Y_1, Y_2, \cdots, Y_{n_2}$를 각각 대조군의 누적분포함수 $F(x)$와 치료군의 누적분포함수 $G(x) = F(x-\theta)$에서 추출된 확률변수라고 하자.

이때 대조군과 치료군의 평균차이는 θ이며, θ를 치료효과라고 한다. 그러므로 비열등성 시험에서의 가설은 다음과 같다.

$$H_0 : \mu_S - \mu_T \geq \delta \text{ vs. } H_1 : \mu_S - \mu_T < \delta$$

여기서 δ는 실험 전에 임상적인 정보를 이용하여 결정되어지는 양의 실수이며, 비열등성의 한계라고 한다.

2.1. 비모수적 방법의 표본 수

비모수적 방법에서의 표본 수를 구하는 방법은 Wang등(2003)이 제안한 Wilcoxon 순위합 검정(Wilcoxon, 1945)을 이용한 표본 수 공식을 비열등성 시험에 확장시켜 적용하여 표본 수를 구하는 방법이다.

Wilcoxon 순위합 검정의 검정통계량 W는 통합순위에서 치료군의 순위의 합으로 정의하며 분산은 다음과 같다.

$$var(W) = n_1 n_2 p_1 (1-p_1) + n_1 n_2 (n_1 - 1)(p_2 - p_1^2) + n_1 n_2 (n_2 - 1)(p_3 - p_1^2)$$

여기서 p_1, p_2, p_3는 다음과 같이 정의된다.

$$\begin{aligned} p_1 &= P(Y_j \geq X_i) \\ p_2 &= P(Y_j \geq X_{i_1} \text{ and } Y_j \geq X_{i_2}) \\ p_3 &= P(Y_{j_1} \geq X_i \text{ and } Y_{j_2} \geq X_i) \end{aligned}$$

그리고 W의 점근분포(asymptotic distribution)는 다음의 평균 μ_W와 분산 σ_W^2을 갖는 정규분포이다.

$$\begin{aligned} \mu_W &= \frac{n_2(n_2+1)}{2} + n_1 n_2 p_1 \\ \sigma_W^2 &= n_1 n_2 p_1 (1-p_1) + n_1 n_2 (n_1 - 1)(p_2 - p_1^2) + n_1 n_2 (n_2 - 1)(p_3 - p_1^2) \end{aligned} \quad (2.1)$$

두 군의 표본 수가 동일할 때 검정력이 가장 크므로 $n_1 = n_2 = n$으로 놓으면 검정력은 다음과 같이 쓸 수 있고

$$Power = 1 - \Phi\left(\frac{z_\alpha \sqrt{1/6} - \sqrt{n}(1/2 - p_1 - \delta)}{\sqrt{p_2 + p_3 - 2p_1^2}} \right)$$

한 그룹당 표본 수는 다음과 같이 계산되어 진다.

$$n = \frac{(z_\alpha \sqrt{1/6} + z_\beta \sqrt{p_2 + p_3 - 2p_1^2})^2}{(1/2 - p_1 - \delta)^2}$$

2.2. 모수적 방법의 표본 수

모수적 방법에서도 대조군과 치료군에 동일한 표본 수를 할당할 때 검정력이 가장 크므로 $n_1 = n_2 = n$이라 가정한다.

대조군의 분산을 σ_S^2, 치료군의 분산을 σ_T^2이라 하면, 이표본 정규검정에서의 검정력은

$$Power = \Phi\left(\frac{\delta-\theta}{\sqrt{\frac{\sigma_S^2+\sigma_T^2}{n}}} - Z_\alpha\right).$$

따라서 대조군과 치료군의 분산이 같을 경우($\sigma_S^2 = \sigma_T^2$)의 각 군의 표본 수는

$$n = \frac{2\sigma^2(Z_\alpha + Z_\beta)^2}{(\theta-\delta)^2}$$

이고 대조군과 치료군의 분산이 다를 경우($\sigma_S^2 \neq \sigma_T^2$)의 각 군의 표본 수는

$$n = \frac{(\sigma_S^2+\sigma_T^2)(Z_\alpha + Z_\beta)^2}{(\theta-\delta)^2}$$

이다.

3. 두 방법의 표본 수 비교

본 논문에서는 비열등성 시험을 위한 두 가지 방법에 대한 표본 수를 비교하기 위해 다음과 같은 상황을 설정하고 그 각각의 경우에서 표본 수를 계산하였다. 모수적 방법에서는 표본 수를 계산하기 위해서 두 그룹의 분산을 가정하여 표본 수를 계산한다. 하지만 비모수적 방법에서는 순위합의 분산을 이용한다. 순위합의 분산은 식(2.1)에서와 같이 확률 p_1, p_2, p_3의 값이 계산되어야 한다.

먼저, 유의수준 $\alpha = 0.05$로, 검정력 $1-\beta = 0.8$로 놓는다. 그리고 δ는 연구자가 무시해도 좋은 정도의 차이가 되므로(이재원 등, 2005) 두 그룹의 평균차이($\mu_S - \mu_T$)에서 비열등성의 한계(δ)와의 차이를 Δ로 정의한다. 이 Δ의 범위를 0~2까지, 분산의 범위를 모수적 방법에서는 1~4까지로 변화하고, 비모수적 방법에서는 모수적 방법에서의 분산에 각각 대응되는 확률 p_1, p_2, p_3값을 계산하여 그 각각의 경우에 표본 수를 계산하였다. 확률 계산에 필요한 모집단의 분포로는 여러 가지 분포를 고려할 수 있으나 본 논문에서는 이용가능성이 많은 정규분포와 이중지수분포를 채택하여 아래와 같이 계산하였으며, 두 그룹의 분산이 같을 때와 다를 때 표본 수 계산 결과를 표 3.1과

표 3.2로 정리하였다.

대조군과 치료군의 분포가 각각 $N(\mu_X,\sigma_X^2)$, $N(\mu_Y,\sigma_Y^2)$을 따른다면 p_1,p_2,p_3는 다음과 같이 계산된다(Simonoff등, 1986).

$$p_1 = P(X_i \le Y_j) = \Phi\left(\frac{\mu_Y-\mu_X}{(\sigma_X^2+\sigma_Y^2)^{1/2}}\right)$$

$$p_2 = \Pr(X_{i_1} \le Y_j \text{ and } X_{i_2} \le Y_j) = \frac{p_1^2}{p_1^2+p_1(1-p_1)+(1-p_1)^2} = \frac{p_1^2}{p_1^2-p_1+1}$$

$$p_3 = \Pr(X_i \le Y_{j_1} \text{ and } X_i \le Y_{j_2}) = \frac{p_1^2}{p_1^2-p_1+1}$$

또한 대조군과 치료군의 분포가 각각 평균 μ_X, μ_Y이고 분산 σ_X^2, σ_Y^2인 이중지수 분포를 따른다면 p_1,p_2,p_3는 다음과 같이 계산된다.

$$p_1 = \begin{cases} 1+\left(\frac{\sigma_X^2}{2(\sigma_X^2-\sigma_Y^2)}-\frac{1}{2}\right)e^{\frac{\mu_X-\mu_Y}{\sigma_Y}} - \frac{\sigma_X^2}{2(\sigma_X^2-\sigma_Y^2)}e^{\frac{\mu_X-\mu_Y}{\sigma_X}} & ,\sigma_X^2 \neq \sigma_Y^2 \\ 1+\left(\frac{\sigma_X}{4(\sigma_X+\sigma_Y)}-\frac{1}{2}\right)e^{\frac{\mu_X-\mu_Y}{\sigma_Y}} + \left(\frac{\mu_X-\mu_Y}{4\sigma_Y}\right)e^{\frac{\mu_X\sigma_Y-\mu_Y\sigma_X}{\sigma_X\sigma_Y}} - \frac{\sigma_X}{4(\sigma_X+\sigma_Y)}e^{\frac{\mu_X-\mu_Y}{\sigma_X}} & ,\sigma_X^2 = \sigma_Y^2 \end{cases}$$

$$p_2 = \frac{p_1^2}{p_1^2-p_1+1}$$

$$p_3 = \frac{p_1^2}{p_1^2-p_1+1}$$

4. 비교결과 및 결론

비열등성 시험에서 모수적 방법을 사용하여 표본 수 계산을 하는 경우 검정력이 떨어지거나 유의수준을 제어하지 못하는 문제점이 발생할 수 있다. 본 논문에서는 이러한 문제점을 보완하기 위해 비모수적 방법을 이용한 표본 수 계산방법을 제안하고 그에 따른 계산 결과를 비교하였다. 표 3.1과 표 3.2를 보면 모든 표본 수는 위에서 다룬 3가지 방법에서 Δ가 작을수록, 분산이 클수록 표본 수가 커지는 경향을 보였다. 또한 분포만을 생각했을 때에는 모수적 방법보다는 비모수적 방법으로 계산했을 때 표본 수가 증가하는 것을 알 수 있었고, 같은 비모수적 방법으로 계산했을 때에는 정규분포를 가정했을 때 보다 이중지수분포를 가정했을 때 표본 수가 증가하는 경향을 나타냈다. 이것은 표본 수를 계산할 때 일반적으로 보이는 성질이다.

표 3.2에서 나타나는 결과와 표 3.1에서 나타나는 결과와의 차이점은 표 3.2에서는 두 그룹의 평균과 분산이 모두 다르므로 앞에서 시행했던 분산이 같은 경우의 모의시험보다 모두 요구되어지는 표본 수가 많음을 알 수 있다. 분산의 차이에 따라서도 요구되어지는 표본 수가 많아지는 것을 알 수 있다.

실제로 사용되는 모수적 방법으로 계산된 표본 수와 비모수적 방법으로 계산된 표본 수를 비교하면 차이가 많음을 알 수 있다. 환자의 모집과 관리, 측정에 소요되는 기간과 경제적 측면에서 보면 환자의 수가 많을수록 연구수행에 어려움이 많을 수 있다. 하지만 표본 수가 적다고 해서 모수적 방법이 더 좋다고는 말할 수 없다. 모수적 방법에서의 가정을 만족시키지 못한다면 검정력이 떨어지거나 유의수준을 제어하지 못하는 문제점이 발생할 수도 있기 때문이다. 이로 인해 올바른 결과를 유도하지 못하여 임상적으로 의미 있는 차이가 있는데도 신약의 유효성을 찾아내지 못하고 기각되어 다시 고려될 수 없는 위험성이 나타날 수 있다.

임상시험에서는 표본 수가 많음으로 인해 연구의 검정력 증가, 유의수준 제어 등 올바른 결과를 도출할 수 있는 장점이 있다. 본 논문에서 제시한 비모수적 방법은 모수적 방법과 비교하여 어떤 것이 좋고 나쁘다고 말하기는 어렵다. 하지만 제안한 방법은 임상시험의 특징과 모집단의 분포를 고려했을 때 모수적 방법을 이용한 표본 수 계산에서 발생할 수 있는 문제를 보완할 수 있는 방법의 한 가지가 될 수 있을 것이며 비용이나 소요되는 시간의 단축보다는 올바른 방향으로 시험의 결과를 유도하여 신약의 유효성을 찾아내는 기회를 제공하는데 좋은 방법이 될 수 있을 것이다.

또한 표본 수 계산 공식에서는 두 그룹의 분산을 필요로 한다. 하지만 표본 수를 계산하기 전에 분산은 알 수 없으므로 기존의 유사한 연구에서 얻은 분산을 이용하여 표본 수를 추정한다. 제안한 방법의 표본 수 공식에서는 두 그룹의 분산 대신 두 그룹에서 추정할 수 있는 확률 p_1, p_2, p_3를 이용한다. 이 확률은 사전연구가 없을 때에도 모집단의 분포를 가정하여 구할 수 있는 장점이 있다.

참고 문헌

1. 이재원, 박미라, 유한나 (2005). 생명과학연구를 위한 통계적 방법, 자유아카데미.
2. ICH (1998). Statistical Principles for Clinical Trials. *International Conference on Harmonization Guideline*, E9.
3. ICH (1999). Choice of Control Group in Clinical Trials. *International Conference on Harmonization Guideline*, E10.
4. Simonoff, J. S., Hochberg, Y. and Reiser, B. (1986). Alternative estimation procedures for Pr(X<Y) in categorized data, *Biometrics*, 42, 895-907.
5. Wang, H., Chen, B. and Chow, S. C. (2003). Sample size determination based on rank tests in clinical trials, *Journal of Biopharmaceutical Statistics*, 13, 735-751.
6. Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics Bulletin*, 6, 80-83.

[2007년 3월 접수, 2007년 5월 채택]

<표 3.1> $\sigma_X^2 = \sigma_Y^2$인 경우의 표본 수 계산 결과 ; $\alpha = 0.05$, $power(1-\beta) = 0.8$

Δ $= \mu_S - \mu_T - \delta$	$\sigma_X^2 = \sigma_Y^2$	모수적 방법	비모수적 방법	
			정규 가정	이중지수 가정
0.5	1	50	62	81
	2	99	119	151
	3	149	174	221
	4	198	228	290
0.8	1	20	27	36
	2	39	50	65
	3	58	72	93
	4	78	94	121
1.0	1	13	18	25
	2	25	34	44
	3	38	48	62
	4	50	62	81
1.2	1	9	13	18
	2	18	24	32
	3	26	35	45
	4	35	45	58
1.5	1	6	9	13
	2	11	17	22
	3	17	23	31
	4	22	30	40
1.8	1	4	7	10
	2	8	12	17
	3	12	17	23
	4	16	22	29
2.0	1	4	6	9
	2	7	10	14
	3	10	14	20
	4	13	18	25

<표 3.2> $\sigma_X^2 \neq \sigma_Y^2$인 경우의 표본 수 계산 결과 ; $\alpha = 0.05$, $power(1-\beta) = 0.8$

Δ $= \mu_S - \mu_T - \delta$	σ_X^2	σ_Y^2	모수적 방법	비모수적 방법	
				정규 가정	이중지수 가정
0.5	1	2	75	91	113
		3	99	119	142
		4	124	146	169
	2	3	124	146	185
		4	149	174	215
	3	4	174	201	254
0.8	1	2	29	39	49
		3	39	50	61
		4	49	61	72
	2	3	49	61	78
		4	58	72	91
	3	4	68	83	106
1.0	1	2	19	26	34
		3	25	34	42
		4	31	41	49
	2	3	31	41	53
		4	38	48	61
	3	4	44	55	71
1.2	1	2	13	19	25
		3	18	24	31
		4	22	30	36
	2	3	22	30	39
		4	26	35	44
	3	4	31	40	52
1.5	1	2	9	13	18
		3	11	17	22
		4	14	20	25
	2	3	14	20	27
		4	17	23	31
	3	4	20	27	35
1.8	1	2	6	10	13
		3	8	12	16
		4	10	15	19
	2	3	10	15	20
		4	12	17	23
	3	4	14	20	26
2.0	1	2	5	8	11
		3	7	10	14
		4	8	12	16
	2	3	8	12	17
		4	10	14	19
	3	4	11	16	22