The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.5.833
27(5), 833–842

Nonparametric Method for a Non-inferiority Test using Confidence Interval

Sujung Park[a] · Dongjae Kim[a,1]

[a]Department of Biomedicine · health science, The Catholic University of Korea

(Received September 29, 2014; Revised October 7, 2014; Accepted October 7, 2014)

Abstract

Non-inferiority trials indicate whether the effect of an experimental treatment is not worse than an active control. Chen *et al.* (2006) and Kang (2010) proposed a test method for non-inferiority trials using confidence intervals. In this paper, we suggest a new nonparametric method using a confidence interval based on Wilcoxon rank-sum test and Hodges-Lehmann estimator of active control. A Monte-Carlo simulation study compares the type I error and the power of the proposed method with previous methods.

Keywords: Non-inferiority trial, confidence interval, Wilcoxon rank-sum test, Hodges-Lehmann estimator

1. 서론

비열등성 시험(non-inferiority trial)이란 시험약(experimental treatment)의 약효가 활성대조약(active control)의 약효보다 열등하지 않음을 보이기 위한 임상시험으로 현재 국내외 많은 제약회사에서 주로 시행되고 있다. 비열등성 시험은 다음 두 가지 목적이 있다. 첫째, 시험약이 독성, 투약 방법, 비용 등의 면에서 장점이 있어 활성대조약보다 나쁘지 않음을 증명하기 위한 경우이다. 둘째, 윤리적으로 위약 사용이 어려워, 가상의 위약인 활성대조약으로 시험약의 유효성을 입증해야 하는 경우이다 (Kang, 2013). 이러한 비열등성 시험에서는 제1종 오류(type I error)의 확률을 명확하게 하는 것이 매우 중요하다. 제1종 오류의 확률이 유의수준(significance level) α보다 높게 나올 경우, 실제로 유의하지 않은 결과가 유의한 것으로 결론 나게 된다. 다시 말해 효과가 없는 약을 마치 효과가 있는 것처럼 결론 내릴 가능성이 유의수준보다 커지는 것이다 (Wang과 Hung, 2003). 따라서 목표한 유의수준을 잘 통제할 수 있는 방법을 이용하여 검정해야 한다.

두 군 모형에서 비열등성 검정을 위한 기존의 방법으로는 Hauschke 등 (1999)이 동등성 검정시 제안한 모수적 방법이 있다. 위약 처방에 윤리적 문제가 없는 경우, 위 방법을 세 군으로 확장한 Pigeot 등 (2003)이 제안한 방법이 있다. 추가로 Lee와 Kim (2008)은 두 군 모형에 대한 Wilcoxon 순위합 검정을 이용한 비모수적 방법과 세 군 모형에 대한 선형대비검정을 확장한 비모수적 방법을 제안하였다.

Chen 등 (2006)은 비열등성 검정에서 신뢰구간을 이용한 전통적 접근법(conventional approach)과 부트스트랩 방법(bootstrap method)을 모의실험을 통하여 비교하였다. 전통적 접근법은 두 모평균

[1]Corresponding author: Department of Biomedicine · health science, The Catholic University of Korea, 222, Banpo-daero, Seocho-gu, Seoul 137-701, Korea. E-mail: djkim@catholic.ac.kr

의 차이의 신뢰구간을 구하고 그 신뢰구간을 활성대조군의 모수적 추정치인 평균으로 나눠줌으로써 최종 신뢰구간을 구하는 방법이다. 부트스트랩 방법은 시험군과 활성대조군으로부터 부트스트랩 표본(bootstrap sample)을 k번 복원추출(sampling with replacement)하여 경험적 신뢰구간(empirical confidence interval)을 구하는 방법이다. Kang (2010)은 추가로 평균의 비율 추정치의 점근적 분포(asymptotic distribution)를 이용한 점근적 방법(asymptotic method)을 제시하였다.

Kang (2010)은 이 세 가지 방법을 모의실험을 통하여 비교하고 부트스트랩 방법이 가장 목표 유의수준을 잘 통제하고 있으며, 대체로 전통적 접근법에 비해 점근적 방법이 유의수준을 잘 통제하는 것을 보였다. 그러나 부트스트랩 방법은 과도한 연산 작업으로 인하여 결과를 도출하는데 많은 시간이 걸리며, 임상시험에서는 원자료만을 가지고 검정해야 하므로 실질적으로 잘 사용되지 않는다는 단점이 있다. 또한, 제시된 세 가지 방법 모두 구체적인 분포함수를 가정했을 때 사용하는 모수적 방법이다. 그러나 모수적 방법을 이용한 구간추정은 오차항이 정규분포를 따르며 두 모집단이 동일한 분산을 가진다는 가정하에 얻어진 방법으로서, 이 가정이 위배된 경우에는 신뢰구간의 길이가 길어지는 경향이 있다.

본 논문에서는 이러한 단점을 보완하기 위해 모평균의 차이의 신뢰구간을 비모수적 방법을 통해 구하고, 그 차이의 신뢰구간을 활성대조군의 비모수적 모평균 추정치로 나눈 비모수적 방법(nonparametric method)을 이용한 비열등성 검정 방법을 제시하였다. 또한, 비모수적 방법과 기존의 전통적, 부트스트랩, 점근적 방법을 모의실험을 통해 여러 조건에 따라 제1종 오류와 검정력(power)의 결과가 어떻게 달라지는지 비교하였다. 이를 통해 평균의 비율을 이용한 비열등성 시험에 있어 적절한 검정 방법을 알아보고 제시된 네 가지 방법들의 장 · 단점을 알아보았다.

2. 제안된 방법

확률 표본 $X_i (i = 1, \ldots, m)$와 $Y_j (j = 1, \ldots, n)$를 각각 활성대조군과 시험군의 모집단으로부터의 확률변수라고 하자. 주효과 변수가 연속형이고, 큰 값이 효과가 좋은 약이라고 가정했을 때, 시험군이 활성대조군보다 열등하지 않음을 보이는 비열등성 시험의 가설은 다음과 같다.

$$\mathrm{H}_0 : \frac{\mu_Y}{\mu_X} \leq M \quad \text{vs.} \quad \mathrm{H}_1 : \frac{\mu_Y}{\mu_X} > M, \quad M \in (0, 1).$$

여기서, μ_X는 활성대조군의 평균, μ_Y는 시험군의 평균이며, M은 비열등성 한계(non-inferiority margin)이다. 이를 평균의 차이의 추정치를 이용한 모평균 비율의 신뢰구간을 구하기 위해 위 가설의 양변에 각각 1을 빼주면

$$\mathrm{H}_0 : \frac{\mu_Y}{\mu_X} - 1 \leq M - 1 \quad \text{vs.} \quad \mathrm{H}_1 : \frac{\mu_Y}{\mu_X} - 1 > M - 1$$

이 된다. 이것을 정리하면 다음과 같이 표현할 수 있다.

$$\mathrm{H}_0 : \frac{\mu_Y - \mu_X}{\mu_X} \leq M - 1 \quad \text{vs.} \quad \mathrm{H}_1 : \frac{\mu_Y - \mu_X}{\mu_X} > M - 1.$$

이때, $\lambda = M - 1$이라 정의하면

$$\mathrm{H}_0 : \frac{\mu_Y - \mu_X}{\mu_X} \leq \lambda \quad \text{vs.} \quad \mathrm{H}_1 : \frac{\mu_Y - \mu_X}{\mu_X} > \lambda, \quad \lambda \in (-1, 0)$$

이다. 여기서 λ는 대조군 치료 효과에 대하여 허용할 수 있는 정도의 감소율을 말한다. 즉, 대조군 효능의 어느 정도 감소율까지 시험군 효능과 차이가 없다고 할 수 있는가를 표현하고 있다. 일반적으로 20%

감소는 임상적으로 중요하게 생각하지 않는다. 따라서 주효과 변수의 값이 클수록 효과가 좋은 약이라고 했을 경우 가설은 다음과 같다.

$$\mathrm{H}_0 : \frac{\mu_Y - \mu_X}{\mu_X} \leq -0.2 \quad \text{vs.} \quad H_1 : \frac{\mu_Y - \mu_X}{\mu_X} > -0.2. \tag{2.1}$$

반대로 작은 값이 효과가 좋은 약이라고 가정했을 경우의 가설은

$$\mathrm{H}_0 : \frac{\mu_Y - \mu_X}{\mu_X} \geq 0.2 \quad \text{vs.} \quad \mathrm{H}_1 : \frac{\mu_Y - \mu_X}{\mu_X} < 0.2 \tag{2.2}$$

으로 표현할 수 있다.

제안하는 비모수적 방법은 전통적 접근법과 마찬가지로 두 모평균의 차이의 신뢰구간을 구하고 그 신뢰구간을 활성대조군의 모평균 추정치로 나누어 비율의 신뢰구간으로 변형해 비열등성 여부를 검정하는 방법이다. 그러나 전통적 접근법과는 달리 두 모평균의 차이의 신뢰구간과 활성대조군의 모평균 추정치를 비모수적 방법 중 하나인 Wilcoxon 순위합 검정에 기초한 신뢰구간과 활성대조군의 Hodges-Lehmann 추정량을 이용하여 추정한다. 여기서 Wilcoxon 순위합 검정에 기초한 신뢰구간은 순위합 통계량을 이용하여 구한다. 모든 $i, j (i = 1, \ldots, m \;\; j = 1, \ldots, n)$에 대해 mn개의 U_{ij}를

$$U_{ij} = Y_j - X_i$$

이라 정의하고, 그 순서통계량을

$$U_{(1)}, U_{(2)}, \ldots, U_{(mn)}$$

이라 하면, 순위합 통계량은 U_{ij}의 중앙값으로 정의된다. $\mu_Y < \mu_X$의 조건에서 Wilcoxon 순위합 검정에 기초한 $100 \times (1-\alpha)\%$ 신뢰구간의 하한과 $\mu_Y > \mu_X$의 조건에서 Wilcoxon 순위합 검정에 기초한 $100 \times (1-\alpha)\%$ 신뢰구간의 상한은 다음과 같다.

$$\mathrm{LL_W} = U_{(C_\alpha)},$$
$$\mathrm{UL_W} = U_{(mn+1-C_\alpha)},$$

여기서 C_α는

$$C_\alpha = \frac{n(2m+n+1)}{2} + 1 - w_\alpha$$

이며, w_α는 Wilcoxon 순위합 통계량의 분포의 상위 $100 \times \alpha$ 백분위수이다. 이때 m, n이 모두 충분히 크다면 C_α는 다음 값에 근사 될 수 있다.

$$C_\alpha \approx \frac{mn}{2} - z_\alpha \sqrt{\frac{mn(m+n+1)}{12}}.$$

일반적으로 위의 두 식에 의해 구해진 C_α는 정수가 아니므로 가장 가까운 정숫값을 선택하여 사용한다 (Moses, 1965; Hollander와 Wolfe, 1973; Song 등, 2003). Hodges-Lehmann 추정량은 Walsh 평균을 이용하며, Walsh 평균은 $\frac{m(m+1)}{2}$개의 모든 $i \leq i' (i, i' = 1, \ldots, m)$에 대해 다음과 같이 정의된다.

$$\mathrm{W}_{ii'} = \frac{X_i + X_{i'}}{2}.$$

이때, Walsh 평균의 중앙값 med($W_{ii'}$)는 활성대조군의 Hodges-Lehmann 추정량이 된다 (Hodges와 Lehmann, 1963). 다음과 같이 구해진 Wilcoxon 순위합 검정에 기초한 신뢰구간의 하한과 상한을 활성대조군의 Hodges-Lehmann 추정량으로 나눈 것을 각각

$$\mathrm{LL}_N = \frac{U_{(C_\alpha)}}{\mathrm{med}(W_{ii'})}, \quad i \le i'(i, i' = 1, \ldots, m),$$
$$\mathrm{UL}_N = \frac{U_{(mn+1-C_\alpha)}}{\mathrm{med}(W_{ii'})}, \quad i \le i'(i, i' = 1, \ldots, m)$$

라고 한다면, LL_N와 UL_N는 $\frac{\mu_Y - \mu_X}{\mu_X}$의 비모수적 $100 \times (1-\alpha)\%$ 단측 신뢰구간의 하한과 상한의 추정치가 될 것이다.

주효과 변수가 큰 값이 효과가 좋은 약이라고 가정한 가설 (2.1)의 검정 방법은 모평균 비율의 $100 \times (1-\alpha)\%$ 단측 신뢰구간이 $(-0.2, \infty)$에 속하면, 즉 신뢰구간의 하한이 -0.2보다 크면 귀무가설을 기각하게 된다. 따라서 귀무가설은 $\mathrm{LL}_N > -0.2$일 경우 기각되며, 활성대조군보다 시험군이 비열등하다고 결론 내린다.

반대로 주효과 변수가 작은 값이 효과가 좋은 약이라고 가정한 가설 (2.2)의 검정 방법은 모평균 비율의 $100 \times (1-\alpha)\%$ 단측 신뢰구간이 $(-\infty, 0.2)$에 속하면, 즉 신뢰구간의 상한이 0.2보다 작으면 귀무가설을 기각하게 된다. 따라서 귀무가설은 $\mathrm{UL}_N < 0.2$일 경우 기각되며, 마찬가지로 활성대조군보다 시험군이 비열등하다고 결론 내린다.

3. 모의실험의 계획 및 결과

본 논문에서는 비열등성 시험의 검정에 대한 Wilcoxon 순위합 검정에 기초한 신뢰구간과 활성대조군의 Hodges-Lehmann 추정량을 이용한 비모수적 방법을 제안하였다. 모의실험은 귀무가설 $H_0 : (\mu_Y - \mu_X)/\mu_X \le -0.2$에 대한 신뢰구간을 이용한 비열등성 검정을 실시하기 위해, 기존에 제안된 전통적, 부트스트랩, 점근적 방법과 본 논문에서 제안한 비모수적 방법의 제1종 오류와 검정력을 비교하였다. 모집단의 분포로는 정규분포, 이중지수분포, 코시분포, 카이제곱분포를 채택하였으며, 정규분포의 난수는 RANNOR 함수, 코시분포의 난수는 RANCAU 함수, 카이제곱분포의 난수는 RANGAM 함수를 이용해 생성하였으며, 이중지수분포의 난수는 RANUNI 함수를 이용하여 역변환 방법을 통해 생성하였다. 유의수준 α는 0.05로 하였으며, 자료의 크기, 평균, 표준편차 등의 조건들을 변화시켜 가면서 몬테카를로 모의실험(Monte-Carlo simulation)을 실시하였다. 프로그램은 SAS 9.3을 이용하여 1000번 독립적으로 반복 실시하였다.

제1종 오류는 모집단이 정규분포와 이중지수분포의 경우 활성대조군과 시험군의 평균이 각각 100, 80일 때 등분산과 이분산의 경우로 나눠 실시하였다. 코시분포는 분산이 존재하지 않기 때문에 활성대조군과 시험군의 평균이 각각 100, 80일 경우에만 실시하였으며, 카이제곱분포의 경우 표준편차 값이 평균에 의해 결정되므로 마찬가지로 활성대조군과 시험군의 평균이 각각 100, 80일 경우에만 실시하였다. 표본의 수는 활성대조군과 시험군이 같고, 25, 100일 때를 고려하였다.

검정력은 모집단이 정규분포와 이중지수분포의 경우 활성대조군의 평균을 100으로 고정하고, 시험군의 평균을 82, 85, 87, 90으로 변화를 주어 등분산과 이분산의 경우로 나눠 실시하였다. 코시분포와 카이제곱분포의 경우 활성대조군의 평균을 100으로 고정하고, 시험군의 평균만 82, 85, 87, 90으로 변화를 주어 실시하였다. 표본의 수는 마찬가지로 활성대조군과 시험군이 같고, 25, 100일 때를 고려하였다.

Table 3.1은 각각의 분포에서 활성대조군의 평균이 100, 시험군의 평균이 80일 때의 각 조건에서의

Table 3.1. Simulation results for type I error of conventional, bootstrap, asymptotic, nonparametric method, with $\mu_X = 100$, $\mu_Y = 80$

	σ_X	σ_Y	$n = m$							
			25				100			
			C	B	A	N	C	B	A	N
Normal	12	12	0.040	0.052	0.047	0.046	0.036	0.053	0.050	0.044
	16	16	0.040	0.052	0.045	0.046	0.036	0.053	0.049	0.044
	20	20	0.040	0.052	0.045	0.046	0.036	0.053	0.049	0.044
	24	24	0.040	0.052	0.045	0.046	0.036	0.053	0.048	0.044
	28	28	0.040	0.052	0.045	0.046	0.036	0.053	0.047	0.044
	20	12	0.027	0.052	0.039	0.038	0.033	0.051	0.047	0.036
	20	14	0.034	0.050	0.041	0.039	0.032	0.051	0.046	0.037
	20	16	0.036	0.046	0.043	0.043	0.035	0.048	0.044	0.040
	20	18	0.039	0.053	0.045	0.045	0.036	0.050	0.047	0.042
	20	22	0.041	0.053	0.049	0.048	0.037	0.053	0.047	0.046
	20	24	0.044	0.055	0.048	0.049	0.041	0.055	0.048	0.047
	20	26	0.046	0.057	0.048	0.050	0.041	0.055	0.050	0.050
	20	28	0.046	0.057	0.050	0.052	0.041	0.057	0.050	0.054
Double exponential	12	12	0.036	0.058	0.046	0.040	0.029	0.043	0.042	0.031
	16	16	0.036	0.058	0.045	0.040	0.029	0.043	0.041	0.031
	20	20	0.036	0.058	0.045	0.040	0.029	0.043	0.040	0.031
	24	24	0.036	0.058	0.044	0.040	0.029	0.043	0.040	0.031
	28	28	0.036	0.058	0.041	0.040	0.029	0.043	0.039	0.031
	20	12	0.028	0.056	0.038	0.035	0.024	0.047	0.037	0.029
	20	14	0.029	0.057	0.038	0.040	0.024	0.044	0.041	0.028
	20	16	0.031	0.057	0.040	0.039	0.025	0.045	0.040	0.030
	20	18	0.033	0.057	0.043	0.037	0.030	0.044	0.040	0.032
	20	22	0.040	0.056	0.046	0.042	0.030	0.044	0.037	0.032
	20	24	0.042	0.056	0.047	0.043	0.031	0.040	0.037	0.033
	20	26	0.043	0.059	0.047	0.043	0.031	0.040	0.038	0.033
	20	28	0.042	0.059	0.046	0.041	0.031	0.041	0.038	0.037
Cauchy	-	-	0.020	0.074	0.025	0.044	0.027	0.077	0.034	0.038
Chi-square	-	-	0.041	0.056	0.047	0.046	0.041	0.059	0.052	0.038

+ C: Conventional approach, B: Bootstrap method, A: Asymptotic method, N: nonparametric method

제1종 오류의 모의실험 결과이다. 표본수가 25인 정규분포에서 등분산일 경우 표준편차가 12일 때를 제외하고 전통적, 점근적, 비모수적 방법 순으로 높은 값을 얻었다. 이분산일 경우에는 시험군의 표준편차가 커질수록 값이 커지는 경향을 보이며, 등분산과 마찬가지로 전통적 접근법이 네 가지 방법 중에 가장 작은 값을 얻었다. 점근적 방법과 비모수적 방법은 0.038에서 0.052사이의 값으로 서로 비슷한 값을 얻었다. 부트스트랩 방법은 활성대조군의 표준편차가 20이고 시험군의 표준편차가 14, 16일 때를 제외하고는 모두 0.05보다 큰 값을 얻어 제1종 오류가 통제되지 않는 결과를 얻는다. 표본수가 100으로 커지고 등분산일 경우 전통적, 비모수적, 점근적 방법 순으로 높았으나, 표준편차가 커짐에 따라 점근적 방법과 비모수적 방법이 비슷한 값을 얻게 되었다. 이분산일 때 역시 전통적 접근법의 값이 가장 작고, 비모수적, 점근적 방법 순으로 좋았다. 부트스트랩 방법의 경우 대부분 0.050보다 큰 값을 얻어 제1종 오류가 통제되지 않았다. 이중지수분포에서 등분산이고 표본이 25일 때는 전통적 접근법과 비모수적 방법이 표준편차의 변화와 상관없이 각각 0.036, 0.040 값을 얻었고, 점근적 방법은 표준편차가 커짐에 따라 비모수적 방법과 비슷한 값을 가졌다. 이분산일 경우 대체로 전통적, 비모수적, 점근적 방법 순으로 큰 값을 얻었으며, 시험군의 표준편차가 활성대조군의 표준편차보다 커질수록 전통적, 비모수적 방법의 값들이 비슷해지는 추세를 보인다. 부트스트랩 방법의 경우 모두 0.056보다 큰 값을 얻었다. 표본수가 100으로 커졌을 경우 등분산과 이분산 모두에서 전통적, 비모수적, 점근적, 부트스트랩 방법 순으로 높은 결과를 보였으며, 특히 이분산일 경우 시험군의 표준편차가 커짐에 따라 비모수적 방법과 점근적 방법의 차이가 줄어드는 경향을 보였다. 코시분포일 경우 표본수가 25와 100일 때 모두 전통적, 점근적, 비모수적 방법 순으로 높은 값을 얻었으며, 부트스트랩 방법의 경우 0.07보다 큰 값을 얻어 제1종 오류

Table 3.2. Simulation results for power of conventional, bootstrap, asymptotic, nonparametric method, with $\mu_X = 100$ and $n = m = 25$

	σ_X	σ_Y	μ_Y															
			82				85				87				90			
			C	B	A	N	C	B	A	N	C	B	A	N	C	B	A	N
Normal	12	12	0.138	0.180	0.159	0.157	0.439	0.516	0.486	0.454	0.677	0.747	0.716	0.688	0.917	0.946	0.938	0.924
	16	16	0.105	0.144	0.132	0.125	0.290	0.354	0.322	0.312	0.466	0.549	0.516	0.483	0.731	0.792	0.771	0.741
	20	20	0.081	0.128	0.101	0.102	0.214	0.269	0.244	0.238	0.335	0.412	0.366	0.349	0.561	0.622	0.585	0.573
	24	24	0.072	0.111	0.085	0.089	0.172	0.220	0.191	0.194	0.260	0.330	0.286	0.283	0.439	0.516	0.465	0.454
	28	28	0.061	0.100	0.075	0.082	0.147	0.193	0.160	0.170	0.214	0.269	0.235	0.238	0.344	0.419	0.366	0.360
	20	12	0.087	0.131	0.121	0.111	0.270	0.379	0.327	0.299	0.450	0.559	0.506	0.470	0.714	0.816	0.780	0.734
	20	14	0.086	0.132	0.117	0.103	0.252	0.345	0.304	0.282	0.420	0.517	0.473	0.435	0.668	0.777	0.732	0.687
	20	16	0.087	0.130	0.113	0.105	0.235	0.311	0.276	0.265	0.390	0.490	0.434	0.417	0.620	0.726	0.684	0.644
	20	18	0.086	0.126	0.108	0.108	0.229	0.288	0.256	0.247	0.361	0.452	0.400	0.380	0.583	0.678	0.626	0.610
	20	22	0.081	0.118	0.096	0.101	0.200	0.245	0.227	0.227	0.311	0.378	0.342	0.337	0.524	0.588	0.556	0.542
	20	24	0.082	0.113	0.093	0.101	0.196	0.231	0.211	0.223	0.296	0.356	0.319	0.311	0.486	0.550	0.511	0.511
	20	26	0.082	0.109	0.090	0.104	0.185	0.225	0.199	0.208	0.277	0.336	0.299	0.295	0.450	0.518	0.476	0.473
	20	28	0.080	0.107	0.087	0.103	0.175	0.215	0.186	0.197	0.262	0.311	0.276	0.277	0.421	0.485	0.437	0.444
Double exponential	12	12	0.108	0.162	0.137	0.167	0.449	0.536	0.495	0.587	0.682	0.746	0.723	0.832	0.913	0.937	0.930	0.975
	16	16	0.080	0.122	0.101	0.114	0.279	0.365	0.327	0.402	0.477	0.569	0.530	0.616	0.734	0.791	0.769	0.869
	20	20	0.066	0.101	0.082	0.095	0.197	0.271	0.237	0.297	0.344	0.417	0.384	0.477	0.588	0.646	0.614	0.730
	24	24	0.057	0.089	0.067	0.082	0.147	0.214	0.178	0.233	0.255	0.336	0.278	0.362	0.449	0.536	0.481	0.587
	28	28	0.050	0.082	0.059	0.074	0.121	0.179	0.134	0.185	0.197	0.271	0.221	0.297	0.356	0.438	0.380	0.487
	20	12	0.066	0.129	0.101	0.116	0.264	0.376	0.331	0.389	0.465	0.563	0.529	0.633	0.722	0.808	0.774	0.874
	20	14	0.067	0.123	0.091	0.106	0.250	0.343	0.299	0.362	0.429	0.530	0.491	0.586	0.686	0.764	0.729	0.835
	20	16	0.066	0.112	0.086	0.094	0.238	0.321	0.274	0.335	0.402	0.503	0.453	0.544	0.650	0.726	0.691	0.798
	20	18	0.063	0.110	0.081	0.096	0.218	0.302	0.252	0.316	0.375	0.457	0.415	0.517	0.620	0.684	0.653	0.767
	20	22	0.066	0.097	0.080	0.091	0.186	0.250	0.212	0.284	0.320	0.392	0.351	0.436	0.546	0.616	0.584	0.694
	20	24	0.066	0.095	0.081	0.089	0.171	0.229	0.192	0.270	0.298	0.366	0.325	0.411	0.497	0.584	0.535	0.659
	20	26	0.065	0.095	0.078	0.089	0.165	0.207	0.178	0.262	0.276	0.341	0.302	0.386	0.468	0.541	0.489	0.624
	20	28	0.066	0.093	0.075	0.087	0.157	0.195	0.168	0.248	0.263	0.318	0.282	0.374	0.444	0.494	0.464	0.599
Cauchy	-	-	0.296	0.436	0.333	0.963	0.649	0.713	0.668	1.000	0.749	0.789	0.757	1.000	0.823	0.853	0.825	1.000
Chi-square	-	-	0.113	0.156	0.139	0.127	0.341	0.425	0.388	0.357	0.547	0.645	0.607	0.580	0.842	0.888	0.884	0.851

+ C: Conventional approach, B: Bootstrap method, A: Asymptotic method, N: nonparametric method

가 통제되지 않음을 볼 수 있다. 카이제곱분포일 경우 표본수가 25일 때는 전통적 접근법이 가장 그 값이 작았으며, 비모수적, 점근적 방법은 비슷한 값을 얻었다. 부트스트랩 방법의 경우 0.056으로 큰 값을 얻었다. 표본수가 100으로 커질 경우에는 비모수적, 전통적 접근법 순으로 좋았고, 부트스트랩과 점근적 방법의 경우 0.05보다 값이 커 제1종 오류가 통제되지 않음을 볼 수 있다.

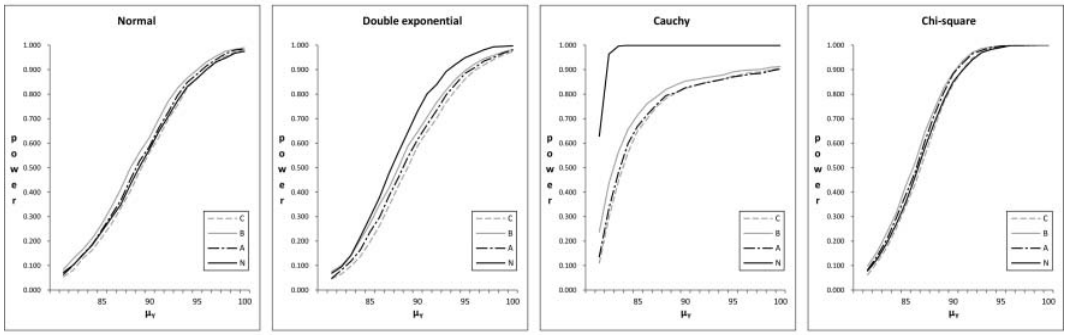
Table 3.2는 표본수가 25일 때 활성대조군의 평균을 100으로 고정하고 시험군의 평균과 표준편차의 변화에 따른 각 분포의 검정력 모의실험의 결과이다. 모든 분포에서 공통으로 시험군의 평균이 같을 경우 표준편차가 커짐에 따라 검정력이 작아지고, 시험군의 평균이 82에서 90으로 커짐에 따라 검정력이 점차 증가한다. 정규분포에서 등분산일 경우 전통적 접근법이 모든 경우에서 가장 작은 검정력 값을 얻었다. 점근적 방법은 비모수적 방법과 비슷한 값을 가지나 표준편차가 커짐에 따라 비모수적 방법보다 약간 낮은 검정력을 보였다. 부트스트랩 방법의 경우 네 가지 방법 중 가장 높은 검정력을 얻었으나, 이는 제1종 오류가 잘 통제되지 않았기 때문에 나타난 결과이다. 이분산일 경우 등분산과 그 경향은 비슷하나 네 가지 방법 간의 값의 차이가 더 큰 것을 볼 수 있다. 이중지수분포이고 시험군의 평균이 82일 경우 등분산과 이분산 모두 전통적, 점근적, 비모수적, 부트스트랩 방법 순으로 검정력이 높게 나왔다. 시험군의 평균이 커져 85, 87, 90이 될 경우에는 전통적, 점근적, 부트스트랩, 비모수적 방법 순으로 검정력이 높게 나왔으며 부트스트랩 방법의 경우 제1종 오류가 통제되지 않아 높은 검정력을 얻은 것으로 보인다. 코시분포의 경우 전통적, 점근적, 부트스트랩, 비모수적 방법 순으로 검정력이 높으며, 특히 비모수적 방법의 검정력이 매우 큰 것을 볼 수 있다. 카이제곱분포의 검정력은 전통적, 비모수적, 점근적, 부트스트랩 방법 순으로 높았으나 부트스트랩 방법은 제1종 오류가 통제되지 않았기 때문에 검정력이 높게 나온 것이다.

Table 3.3은 표본수가 100일 때 활성대조군의 평균을 100으로 고정하고 시험군의 평균과 표준편차의 변화에 따른 각 분포의 검정력 모의실험의 결과를 보여준다. 정규분포에서 등분산과 이분산 모두 전통적

Table 3.3. Simulation results for power of conventional, bootstrap, asymptotic, nonparametric method, with $\mu_X = 100$ and $n = m = 100$

	σ_X	σ_Y	μ_Y															
			82				85				87				90			
			C	B	A	N	C	B	A	N	C	B	A	N	C	B	A	N
Normal	12	12	0.305	0.365	0.354	0.304	0.912	0.944	0.933	0.908	0.996	0.997	0.997	0.996	1.000	1.000	1.000	1.000
	16	16	0.216	0.265	0.252	0.210	0.730	0.783	0.776	0.732	0.935	0.959	0.954	0.933	0.997	0.999	0.998	0.998
	20	20	0.162	0.206	0.192	0.165	0.556	0.621	0.611	0.561	0.826	0.857	0.853	0.817	0.978	0.985	0.986	0.982
	24	24	0.135	0.172	0.158	0.139	0.403	0.490	0.458	0.412	0.685	0.731	0.723	0.684	0.912	0.944	0.929	0.908
	28	28	0.127	0.150	0.140	0.126	0.334	0.395	0.373	0.336	0.556	0.621	0.602	0.561	0.837	0.863	0.858	0.828
	20	12	0.201	0.278	0.266	0.209	0.724	0.807	0.797	0.715	0.929	0.961	0.954	0.930	0.998	0.999	0.999	1.000
	20	14	0.184	0.259	0.241	0.196	0.680	0.767	0.752	0.676	0.907	0.941	0.935	0.902	0.996	0.998	0.997	0.999
	20	16	0.175	0.236	0.228	0.179	0.644	0.713	0.701	0.642	0.881	0.908	0.907	0.875	0.994	0.997	0.997	0.994
	20	18	0.168	0.222	0.213	0.167	0.601	0.673	0.656	0.604	0.850	0.885	0.879	0.849	0.989	0.994	0.993	0.988
	20	22	0.157	0.194	0.178	0.161	0.507	0.573	0.562	0.520	0.775	0.825	0.815	0.779	0.965	0.978	0.975	0.966
	20	24	0.153	0.183	0.170	0.155	0.466	0.530	0.508	0.484	0.736	0.775	0.768	0.736	0.950	0.966	0.960	0.947
	20	26	0.153	0.176	0.165	0.151	0.443	0.487	0.468	0.448	0.697	0.732	0.727	0.692	0.931	0.946	0.941	0.927
	20	28	0.145	0.169	0.161	0.146	0.415	0.458	0.443	0.422	0.657	0.693	0.688	0.639	0.905	0.928	0.923	0.901
Double exponential	12	12	0.312	0.378	0.366	0.429	0.910	0.928	0.928	0.977	0.996	0.997	0.998	1.000	1.000	1.000	1.000	1.000
	16	16	0.193	0.258	0.239	0.287	0.724	0.781	0.773	0.874	0.928	0.948	0.947	0.986	0.998	0.998	0.998	1.000
	20	20	0.145	0.190	0.182	0.209	0.546	0.611	0.597	0.723	0.811	0.847	0.838	0.926	0.975	0.982	0.983	0.996
	24	24	0.122	0.154	0.148	0.174	0.425	0.489	0.469	0.567	0.660	0.735	0.712	0.823	0.910	0.928	0.926	0.977
	28	28	0.101	0.138	0.128	0.138	0.347	0.405	0.388	0.461	0.546	0.611	0.590	0.723	0.822	0.858	0.846	0.934
	20	12	0.169	0.256	0.239	0.270	0.709	0.791	0.776	0.871	0.931	0.960	0.954	0.987	0.997	0.999	0.999	1.000
	20	14	0.166	0.246	0.221	0.248	0.669	0.752	0.737	0.835	0.906	0.934	0.934	0.982	0.997	0.997	0.997	1.000
	20	16	0.157	0.230	0.207	0.228	0.625	0.706	0.691	0.795	0.874	0.908	0.909	0.964	0.992	0.997	0.996	0.999
	20	18	0.154	0.207	0.193	0.215	0.581	0.652	0.638	0.763	0.841	0.876	0.874	0.943	0.988	0.992	0.990	0.998
	20	22	0.144	0.180	0.168	0.199	0.507	0.570	0.548	0.675	0.777	0.812	0.803	0.905	0.958	0.973	0.968	0.992
	20	24	0.141	0.174	0.164	0.190	0.468	0.532	0.514	0.638	0.731	0.775	0.765	0.884	0.940	0.951	0.949	0.988
	20	26	0.139	0.166	0.159	0.185	0.442	0.489	0.471	0.602	0.686	0.730	0.723	0.852	0.922	0.937	0.935	0.979
	20	28	0.134	0.160	0.153	0.178	0.417	0.456	0.451	0.570	0.651	0.687	0.674	0.813	0.893	0.914	0.909	0.969
Cauchy	-	-	0.293	0.410	0.325	1.000	0.602	0.683	0.623	1.000	0.707	0.763	0.721	1.000	0.807	0.832	0.807	1.000
Chi-square	-	-	0.265	0.317	0.308	0.281	0.863	0.900	0.887	0.852	0.981	0.991	0.986	0.985	1.000	1.000	1.000	1.000

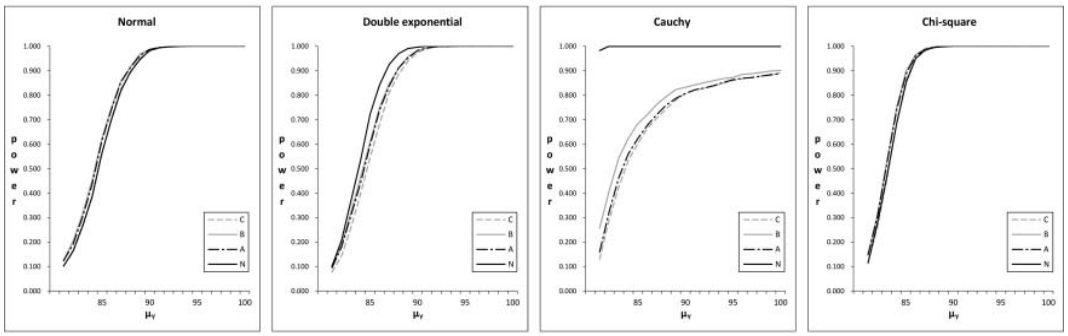
+ C: Conventional approach, B: Bootstrap method, A: Asymptotic method, N: nonparametric method



+ C: Conventional approach, B: Bootstrap method, A: Asymptotic method, N: nonparametric method

Figure 3.1. Simulation results for power of conventional, bootstrap, asymptotic, nonparametric method, with $\mu_X = 100$ and $n = m = 25$

접근법과 비모수적 방법의 검정력이 비슷하게 나왔으며, 그다음으로 점근적 방법이 높은 값을 얻었다. 같은 조건하에 부트스트랩 방법의 검정력이 가장 좋은 것으로 나왔으나 이는 제1종 오류가 0.05보다 크게 나왔기 때문이다. 이중지수분포일 경우 등분산과 이분산 모두 전통적, 점근적, 부트스트랩, 비모수적 방법 순으로 검정력이 높게 나왔다. 코시분포의 경우도 모두 전통적, 점근적, 부트스트랩, 비모수적 방법 순으로 검정력이 높으며, 시험군의 평균의 변화와 상관없이 비모수적 방법의 검정력은 1이다. 카이제곱분포에서 시험군의 평균의 변화에 따른 검정력의 모의실험의 결과는 전통적, 비모수적, 점근적, 부트스트랩 방법 순으로 검정력이 크게 나왔으나, 부트스트랩 방법의 검정력이 큰 이유는 제1종 오류의 값



+ C: Conventional approach, B: Bootstrap method, A: Asymptotic method, N: nonparametric method

Figure 3.2. Simulation results for power of conventional, bootstrap, asymptotic, nonparametric method, with $\mu_X = 100$ and $n = m = 100$

이 통제되지 않았기 때문이다.

Figure 3.1과 Figure 3.2는 표본수가 각각 25와 100일 때의 시험군과 활성대조군의 표준편차가 20인 정규분포와 이중지수분포에서의 활성대조군의 평균 변화에 따른 검정력 변화와 코시분포와 카이제곱분포에서의 활성대조군의 평균 변화에 따른 검정력 변화를 확인할 수 있다. 표본수가 25일 때 정규분포의 경우 전통적, 비모수적, 점근적, 부트스트랩 방법 순으로 높게 나왔으나 부트스트랩 방법의 경우 제1종 오류를 통제하지 못해서 나온 결과이다. 이중지수분포와 코시분포의 경우 전통적, 점근적, 부트스트랩, 비모수적 방법 순으로 큰 값을 얻었으며, 특히 코시분포는 활성대조군의 평균이 커짐에 따라 비모수적 방법의 검정력이 급격하게 커지는 것을 볼 수 있다. 카이제곱분포의 경우 정규분포와 마찬가지로 전통적, 비모수적, 점근적, 부트스트랩 방법 순으로 높게 나왔다. 표본수가 100으로 커질 때 정규분포의 경우 각각 전통적 접근법과 비모수적 방법, 점근적 방법과 부트스트랩 방법이 비슷한 검정력을 보였다. 이중지수분포일 경우 표본수가 25일 때와 마찬가지로 전통적, 점근적, 부트스트랩, 비모수적 방법 순으로 큰 값을 얻었으며, 코시분포의 경우 모두 전통적, 점근적, 부트스트랩, 비모수적 방법 순으로 검정력이 높은 것으로 나왔다. 특히 비모수적 방법의 검정력은 활성대조군의 평균이 81일 때를 제외하고는 모든 조건에서 1을 얻었다. 카이제곱분포에서는 전통적 접근법과 비모수적 방법이 비슷한 값을 얻었고, 그다음으로 점근적, 부트스트랩 방법의 검정력이 높은 것을 볼 수 있다.

4. 결론 및 고찰

본 논문에서는 비열등성 시험에서 신뢰구간을 이용한 검정에 대한 비모수적 방법을 제안하였다. 이 방법은 Wilcoxon 순위합 검정에 기초한 신뢰구간과 활성대조군의 Hodges-Lehmann 추정량을 사용하여 만들어졌으며, 또한 Chen 등 (2006)과 Kang (2010)이 제안한 기존의 방법과 본 논문에서 제안한 방법의 제1종 오류와 검정력을 모의시험을 통하여 비교하였다. 모의실험을 통하여 이 네 가지 방법을 정규분포, 이중지수분포, 코시분포, 그리고 카이제곱분포에서 비교한 결과 분포에 따라 각 방법의 제1종 오류와 검정력이 다름을 알 수 있었다. 제1종 오류의 결과를 보면 대부분의 경우에서 전통적 접근법은 네 가지 방법 중 가장 작은 제1종 오류 값을 나타냈다. 또 부트스트랩 방법의 경우 표본이 100인 이중지수분포를 제외하고는 0.05보다 큰 값을 얻어, 비열등성 시험의 검정에서는 지양해야 하는 통계 방법으로 생각된다. 비모수적 방법의 경우 코시분포를 제외하고는 점근적 방법과 비슷하거나 작은 값을 보였으며, 표본수가 작을수록 더욱 비슷해지는 경향을 보였다. 코시분포에서는 비모수적 방법이 점근적 방

법보다 더 제1종 오류를 잘 통제하였다. 또한, 일반적으로 정규분포일 때 비모수적 방법을 사용한다면 검정력의 감소를 가지고 오는데, 정규분포를 포함한 여러 분포에서도 다른 검정 방법의 검정력보다 비모수적 방법의 검정력이 크게 낮아지지 않는 것으로 나타났다. 그뿐만 아니라 본 논문에서 제안한 방법이 이중지수분포와 코시분포에서 Chen 등 (2006)과 Kang (2010)이 제안한 방법보다 검정력이 높았다. 특히 코시분포에서 비모수적 방법의 검정력은 활성대조군의 평균의 변화에 매우 민감한 것으로 나왔다. 따라서 모집단의 분포가 꼬리가 두텁고 넓게 퍼져있는 특성을 보이는 경우에는 본 논문에서 제안한 방법을 쓰는 게 더 효율적인 분석이 될 것으로 생각된다.

References

Chen, M., Kianifard, F. and Dhar, S. (2006). A bootstrap-based test for establishing noninferiority in clinical trials, *Biopharmaceutical Statistics*, **16**, 357–363.

Hauschke, D., Kieser, M., Diletti, E. and Burke, M. (1999). Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Statistics in Medicine*, **18**, 93–105.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests, *The Annals of Mathematical Statistics*, **34**, 598–611.

Hollander, M. and Wolfe, D. A. (1973). Nonparametric statistical methods, 2nd edition, *John Wiley*, New York.

Kang, H. (2010). A comparison of confidence interval methods in a non-inferiority test. *Master's Thesis*, Korea University.

Kang, S. (2013). Medical statistics needed for drug development, 2nd edition, *Free Academy*.

Lee, J. and Kim, D. (2008). Non-inferiority test in a two-arm trial and a three-arm trial including a placebo, *The Korean Journal of Applied Statistics*, **21**, 947–957.

Moses, L. (1965). Confidence limits from rank tests, *Technometrics*, **7**, 257–260.

Pigeot, I., Schafer, J., Rohmel, J. and Hauschke, D. (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo, *Statistics in Medicine*, **22**, 883–899.

Song, M., Park, C. and Lee, J. (2003). Non-parametric statistics using S-LINK, *Free Academy*.

Wang, S. and Hung, H. (2003). TACT method for non-inferiority testing in active controlled trials, *Statistics in Medicine*, **22**, 227–238.

신뢰구간을 이용한 비열등성 시험에서 비모수적 검정법

박수정[a] · 김동재[a,1]

[a]가톨릭대학교 의생명 · 건강과학과

(2014년 9월 29일 접수, 2014년 10월 7일 수정, 2014년 10월 7일 채택)

요 약

비열등성 시험이란 시험군이 활성대조군보다 열등하지 않음을 증명하는 임상시험이다. 이러한 비열등성 시험에서 신뢰구간을 이용한 검정 방법에는 Chen 등 (2006)과 Kang (2010)이 제안한 방법이 있다. 본 논문에서는 Wilcoxon 순위합 검정에 기초한 두 모평균 차이의 신뢰구간과 활성대조군의 Hodges-Lehmann 추정량을 이용하여 비모수적 방법을 제안하였다. 또한 몬테카를로 모의실험(Monte-Carlo simulation)을 통하여 기존의 방법과 제안한 방법의 제1종 오류와 검정력을 비교하였다.

주요용어: 비열등성 시험, 신뢰구간, Wilcoxon 순위합 검정, Hodges-Lehmann 추정량

[1]교신저자: 가톨릭대학교 의생명 · 건강과학과, 서울특별시 서초구 반포대로 222 137-701.
E-mail: djkim@catholic.ac.kr