응용통계연구 (2009)
22(6), 1249–1263

단변량 및 이변량 순위변수의 비모수적 윌콕슨 검정법에 의한 표본수 결정방법

박해강[1] · 송혜향[2]

[1]가톨릭대학교 의과대학 의학통계학과, [2]가톨릭대학교 의과대학 의학통계학과

(2009년 8월 접수, 2009년 10월 채택)

요 약

표본수 결정에서 요구되는 검정력 함수는 연구가설에 상응하는 가장 적절한 검정방법에 의한 것이어야 한다. 의학연구의 논문에 자주 나타나는 순위자료 또는 범주형 빈도자료의 분석에는 비모수적 방법이 적절하며, 본 논문에서는 단변량 및 이변량 순위변수에 대한 윌콕슨-만-휘트니(Wilcoxon-Mann-Whitney; WMW) 검정법에 의한 표본수 결정방법을 제시한다. 단변량 순위변수의 윌콕슨 검정에서는 귀무가설과 대립가설 하의 분산을 이용한 표본수 공식이 귀무가설 하의 분산만 이용한 표본수 공식보다 정확하지만, 대립가설 하의 분산식에 나타나는 확률값이 일반적으로 알려져 있지 않으므로 이 확률값의 추정이 문제가 된다. 모의실험으로 두 방법에 대한 장, 단점을 알아본다. 효능과 안전성의 이변량 순위변수에서는 이변량 WMW 검정법에 의한 표본수 결정방법이 모수적 검정법에 의한 표본수 결정방법보다 더욱 바람직하다.

주요용어: 표본수 계산, 순위변수, 윌콕슨 검정통계량.

1. 서론

임상시험의 주요 반응변수가 순위변수 또는 범주형으로 측정되어 치료효과를 비모수 윌콕슨-만-휘트니(Wilcoxon-Mann-Whitney; WMW, 간략히 윌콕슨) 검정에 의해 결정하는 경우를 논문에서 빈번하게 볼 수 있다. 한편, 임상시험을 다룬 다수의 교재에서는 연구계획 단계에서 표본수를 정할 때, 실제 자료분석에 사용될 분석법에 의해 결정되어야 한다고 강조하고 있다. 따라서 순위변수를 비모수검정법으로 분석하는 경우에는 비모수검정법에 의해 결정되는 연구대상자수로 장차 연구가 계획되어야 한다. 그러나 일부 연구자는 이러한 원칙을 지키지 않고 쉽게 순위변수를 연속변수 취급하여 정규분포 가정 하에서 모수적 분석법에 해당하는 연구대상자수를 정하는데, 이와 같은 손쉬운 방법을 적용할 수 없는 경우가 있다. 순위변수를 연속변수로 취급하고자 한다면 순위변수의 각 분류항목에 대응하는 수치(score)가 구체적으로 제시될 때에 가능하며 만약 극단의 항목, 즉 가장 큰 분류항목이 어떤 정해진 수치 이상을 나타내는 열린 항목(open-ended category)으로 표현되는 경우에는 이 항목을 구체적으로 어떤 수치로 대치하는가에 따라 두 군의 t 검정결과가 크게 달라지므로 (Graubard와 Korn, 1987), 연속변수 검정법의 적용이 부적절하다. 이와 같이 열린 항목으로 표현된 순위자료의 분석에는 비모수검정법이 매우 적절한데, 그 이유는 비모수 검정통계량은 항목의 수치를 구체적으로 알 필요가 없고 단지 순위에 기초하여 구한다는 장점을 가지고 있기 때문이다. 또한 반응변수 수치가 두 봉오리 이상을 가지는 다항분포로

[2]교신저자: (137-701) 서울시 서초구 반포동 505, 가톨릭대학교 의과대학 의학통계학과, 교수.
E-mail: hhsong@catholic.ac.kr

써 정규분포를 가정할 수 없는 경우에도 비모수검정법이 매우 적절하다. 본 논문에서는 이와 같은 경우에 해당하는 순위변수의 표본수 결정방법을 검토하고 계산한다.

임상에서 행해지는 약제 비교에 대한 대다수의 연구에서 연구자는 효능과 안전성 모두에 관심이 있음을 표현하고 있으며, 한 가지 예가 소아열성 치료에 대한 ibuprofen과 acetaminophen의 병합이 ibuprofen 단독약제보다 효능과 안전성이 탁월한지 알아보기 위해 이중맹검법 랜덤임상시험으로 실시한 연구이다 (Nabulsi 등, 2006). 이들 저자를 포함하여 임상의 여러 연구자들은 효능과 안전성 모두에 관심이 있으면서도 실제로 단변량 효능에 대한 검정법에 근거하여 표본수를 결정하고 있으며, 이러한 추세의 가장 큰 이유는 이변량 반응변수의 검정법에 근거한 표본수 결정방법이 널리 알려져 있지 않기 때문이며, 이변량 순위변수의 검정법에 대해서는 더욱 그러하다.

본 논문에서는 우선 단변량 순위변수의 윌콕슨검정법에 근거한 표본수 결정방법을 제시한 Noether (1987)의 방법과 이를 보완한 여러 방법들을 설명한다. 이변량 순위변수의 검정법에 근거한 표본수 결정방법으로서는 이현학과 송혜향 (2009)에 제시된 이변량 윌콕슨검정법에 근거한 방법을 간략히 설명한다. 이변량 이항변수의 비율차 검정의 표본수 결정방법인 Thall과 Cheng (1999)의 방법과 비교하기 위해 이현학과 송혜향 (2009)에서는 이변량 이항변수의 표본수만을 계산해 보였으며 이변량 순위변수의 표본수를 제시하지 않았으나, 본 논문에서는 구체적으로 이변량 윌콕슨검정법에 의한 표본수를 계산하여 제시한다.

2. 단변량 윌콕슨 검정법의 표본수 결정

서로 독립인 대조군과 치료군을 각각 C와 T로 나타낼 때, Y_{Ci}, $i = 1, \ldots, n_C$와 Y_{Tj}, $j = 1, \ldots, n_T$는 대조군의 분포함수 $F(x)$와 치료군의 분포함수 $G(x) = F(x - \theta)$에서 추출된 반응변수이다. 비모수 윌콕슨검정에서는 두 모집단으로부터 추출된 분포함수가 동일형태이면서 어떤 구체적인 분포함수를 가정하지 않으며, 두 군 분포함수간에 위치모수의 차이(location shift)만을 가정한다. 이러한 대조군과 치료군의 위치모수의 차이 $\theta(> 0)$를 치료효과라 하며 치료군의 대상자들은 더욱 큰 반응수치를 가진다. 그리고 n_C와 n_T는 각각 대조군과 치료군의 표본수이고 전체 표본수는 $N = n_C + n_T$이다. 본 논문에서는 단측검정으로서 유의수준 α와 검정력 $1 - \beta$로 두 군에서 요구되는 표본수를 구하며, 이 때 귀무가설은 다음과 같다.

$$H_0 : \theta = 0 \quad vs. \quad H_a : \theta > 0. \tag{2.1}$$

2.1. Noether (1987)의 표본수 결정

Noether (1987)는 윌콕슨 검정통계량인 $U = \sum_{j=1}^{n_T} \sum_{i=1}^{n_C} \psi\{Y_{Ci},\ Y_{Tj}\}$에 근거한 표본수 공식을 제시하였다. 여기서

$$\psi(Y_{Ci},\ Y_{Tj}) = \begin{cases} 1, & \text{if } Y_{Ci} < Y_{Tj}, \\ 0, & \text{if } Y_{Ci} > Y_{Tj}. \end{cases}$$

검정통계량 U의 기대값 $\mu(U)$와 분산 $\sigma^2(U)$는 다음과 같다 (Sundrum, 1953).

$$\begin{aligned} \mu(U) &= n_C n_T p_1, \\ \sigma^2(U) &= n_C n_T p_1(1 - p_1) + n_C n_T (n_C - 1)\left(p_2 - p_1^2\right) + n_C n_T (n_T - 1)\left(p_3 - p_1^2\right) \end{aligned} \tag{2.2}$$

이 때 p_1, p_2, p_3는 다음과 같이 정의된다.

$$p_1 = \Pr(Y_{Ci} < Y_{Tj}),$$

$$p_2 = \Pr(Y_{Ci} < Y_{Tk} \text{ and } Y_{Cj} < Y_{Tk}),$$
$$p_3 = \Pr(Y_{Ck} < Y_{Ti} \text{ and } Y_{Ck} < Y_{Tj}),$$

여기서 확률 p_1의 의미는 치료군에서 임의로 선택한 수치가 이와 독립인 대조군에서 임의로 선택한 수치보다 클 확률이며, 확률 p_2의 의미는 치료군에서 임의로 선택한 수치가 이와 독립인 대조군에서 임의로 선택한 두 다른 수치보다 클 확률이다. 그리고 확률 p_3의 의미는 대조군에서 임의로 선택한 수치가 이와 독립인 치료군에서 임의로 선택한 두 다른 수치보다 작을 확률이다.

귀무가설 $H_0 : F(x) = G(x)$ 하에서 이 확률들은 각각 $p_1 = 1/2$, $p_2 = 1/3$, $p_3 = 1/3$의 값을 가지며 (Sundrum, 1953), 이 확률값을 기대값과 분산 공식 (2.2)에 대입하면 다음과 같은 귀무가설 하에서의 U 통계량의 기대값 $\mu_0(U)$와 분산 $\sigma_0^2(U)$이 구해진다.

$$\mu_0(U) = \frac{n_C n_T}{2}$$
$$\sigma_0^2(U) = \frac{n_C n_T (N+1)}{12}$$

치료효과가 있다는 대립가설 하에서는 이 확률들은 각각 $p_1 = 1/2$, $p_2 = 1/3$, $p_3 = 1/3$보다 큰 값을 가지게 된다.

이제 Noether (1987)과 여러 연구자들이 제시한 표본수 공식을 구체적으로 살펴본다. 윌콕슨 검정통계량 U가 귀무가설 하에서 점근적(asymptotically)으로 정규분포를 따름이 Mann과 Whitney (1947)와 Haldane과 Smith (1948)에 의해 증명되었으며, p_1의 값이 0 또는 1에 가까운 값을 가지지 않는다면 대립가설 하에서도 점근적으로 정규분포에 따른다 (Sundrum, 1953). 따라서 정규분포를 이용한 검정력 함수는 다음과 같다.

$$\begin{aligned} \text{Power} &= \Pr\left(U > \mu_0(U) + z_\alpha \sigma_0(U) \mid H_a\right) \\ &\cong \Pr\left(Z > \frac{\mu_0(U) - \mu(U) + z_a \sigma_0(U)}{\sigma(U)}\right), \end{aligned} \tag{2.3}$$

여기서 Noether (1987)는 두 가설 하에서의 U의 분산이 크게 다르지 않다는 가정 하에서, 즉 $\sigma_0^2(U) = \sigma^2(U)$의 가정 하에서, 또한 전체 군에서의 치료군의 표본수 비율이 $k_1 = n_T/N (0 < k_1 < 1)$일 때 Noether (1987)는 다음과 같은 표본수 공식을 제시하였으며, 여기서의 표본수 N은 전체 군에서 요구되는 표본수이다.

$$N = \frac{(z_\alpha + z_\beta)^2}{12 k_1 (1-k_1)(p_1 - 0.5)^2}. \tag{2.4}$$

표본수 공식 (2.4)는 유의수준 α와 검정력 $1-\beta$에 해당하는 z값과 또한 $p_1 = P(Y_{Ci} < Y_{Tj})$으로서 표현되었는데, 이는 식 (2.2)에 제시된 $\mu(U) = n_C n_T p_1$에 근거한다. 비모수 검정에서는 분포함수의 어떤 형태도 가정하지 않으므로 위치모수의 차이인 θ의 구체적인 값을 정하기는 어려운 일이며, 따라서 윌콕슨 검정법의 표본수 결정에서는 θ보다도 확률 p_1이 더욱 자연스러운 모수라 하겠다. 그러므로 귀무가설은 다시 다음과 같이 표현된다.

$$H_0 : p_1 = 0.5 \quad vs. \quad H_a : p_1 > 0.5. \tag{2.5}$$

Vollandt와 Horn (1997)은 Noether (1987)와 다르게 윌콕슨 검정통계량의 귀무가설과 대립가설 하에서의 분산이 다름을 감안한 표본수 공식을 제안하였다. 저자들은 Van Dantzig (1951)이 제시한 윌콕슨 검정통계량의 대립가설 하에서의 분산의 상한값(upper bounds), 즉 모든 대립가설의 경우에 대해 성립되는 분산값을 이용하여 다음을 제안하였다. 여기서 $n = n_C = n_T$는 각 군에서 요구되는 표본수이다.

표 **2.1.** Probabilities of two groups in contingency tables

Group	1	2	$\cdots$	K	Total
Control Group	p_{C1}	p_{C2}	$\cdots$	p_{CK}	1
Treatment Group	p_{T1}	p_{T2}	$\cdots$	p_{TK}	1

$$\frac{1/2 - p_1 - z_{1-\alpha}\sqrt{(2n+1)/(12n^2)}}{\sqrt{(17n^2-20n+6)/[12(2n-1)^3]}} \geq z_{1-\beta}, \quad \text{for} \quad \frac{3}{8} < p_1 < \frac{1}{2}, \tag{2.6}$$

$$\frac{1/2 - p_1 - z_{1-\alpha}\sqrt{(2n+1)/(12n^2)}}{\sqrt{(n\Delta_1+\Delta_2)/(3n^2)}} \geq z_{1-\beta}, \quad \text{for} \quad 0 < p_1 < \frac{3}{8}, \tag{2.7}$$

여기서 $\Delta_1 = 1/2 - 6\Delta^2 + (2\Delta)^{3/2}$, $\Delta_2 = 1/4 + 3\Delta^2 - (2\Delta)^{3/2}$이고 $\Delta = 1/2 - p_1$이다. Vollandt와 Horn (1997)은 공식 (2.6)과 (2.7)에 의한 표본수와 Noether (1987)의 공식 (2.4)에 의한 표본수를 비교함으로써 Noether (1987)의 공식이 충분히 신뢰할만하다고 결론을 내렸다. 그러나 Vollandt와 Horn (1997)의 공식 (2.6)과 (2.7)은 단순치 않아서 표본수의 즉각적인 계산에 사용하기에는 한계가 있다하겠다.

한편, Zhao 등 (2008)은 순위자료가 $2 \times K$ 분할표 형태로 제시되어 동점 순위가 있는 경우에 대한 윌콕슨 검정통계량의 표본수 공식을 Noether (1987)의 공식에 근거하여 제시하였다.

여기서 표본수 결정방법의 가설은 위의 (2.5)와 같고 검정통계량은

$$\widetilde{\psi}(Y_{Ci},\ Y_{Tj}) = \begin{cases} 1, & \text{if} \quad Y_{Ci} < Y_{Tj}, \\ 0.5, & \text{if} \quad Y_{Ci} = Y_{Tj}, \\ 0, & \text{if} \quad Y_{Ci} > Y_{Tj} \end{cases}$$

을 이용하여 $U/(n_C n_T) = \sum_{j=1}^{n_T}\sum_{i=1}^{n_C}\widetilde{\psi}\{Y_{Ci},\ Y_{Tj}\}/(n_C n_T) \equiv \widetilde{p_1}$ 일 때, 평균과 분산의 추정량은

$$\widehat{\mu}(\widetilde{p_1}) = \sum_{i=1}^{K} p_{Ci}\left(\sum_{l=1}^{K} p_{Tl} - \frac{p_{Ti}}{2}\right), \tag{2.8}$$

$$\widehat{\sigma}_0^2(\widetilde{p_1}) = \frac{1}{12Nk_1(1-k_1)}\left[1 - \sum_{i=1}^{K}\{(1-k_1)p_{Ci} + k_1 p_{Ti}\}^3\right] \tag{2.9}$$

이고, 이를 이용한 표본수 공식은 다음과 같다 (Zhao 등, 2008).

$$N = \frac{(z_\alpha + z_\beta)^2\left[1 - \sum_{i=1}^{k}\{(1-k_1)p_{Ci} + k_1 p_{Ti}\}^3\right]}{12k_1(1-k_1)\left\{\sum_{i=1}^{K} p_{Ci}\left(\sum_{l=1}^{K} p_{Tl} - \frac{p_{Ti}}{2}\right) - 0.5\right\}^2}. \tag{2.10}$$

2.2. Wang 등 (2003)의 표본수 결정

Wang 등 (2003)은 귀무가설 하에서와 대립가설 하에서의 분산의 다름을 그대로 반영한 표본수 공식을 제안하며, 한편 저자들의 논문에서는 Noether (1987)의 표본수 결정방법에 대한 언급이나 비교를 찾아볼 수 없다. 또 다른 차이로 Noether (1987)는 U 통계량에 근거하여 공식을 유도하였고, Wang 등

(2003)은 다음과 같은 두 군의 통합순위에서 치료군의 순위합으로 정의되는 윌콕슨 순위합 검정통계량 W, 즉

$$W = \sum_{j=1}^{n_T} \left(\sum_{i=1}^{n_T} \psi\{Y_{Ti}, Y_{Tj}\} + \sum_{i=1}^{n_C} \psi\{Y_{Ci}, Y_{Tj}\} \right) \tag{2.11}$$

에 근거한 표본수 공식을 제시하며, 검정통계량 U와 W는 다음과 같이 상수차이의 관계를 가져서 분산이 동일하다.

$$\begin{aligned} W &= \frac{n_T(n_T+1)}{2} + \sum_{j=1}^{n_T}\sum_{i=1}^{n_C} \psi\{Y_{Ci}, Y_{Tj}\} \\ &= \frac{n_T(n_T+1)}{2} + U \end{aligned}$$

따라서 검정력 (2.3)의 분자는

$$\mu_0(W) - \mu(W) + z_a\sigma_0(W) = \mu_0(U) - \mu(U) + z_a\sigma_0(U) \tag{2.12}$$

에 의해 서로 동일하므로 결과적으로 Noether (1987)와 Wang 등 (2003)의 표본수 공식의 차이는 분모에 나타난 대립가설 하에서의 분산의 제곱근, 즉 Noether (1987)가 사용한 $\sigma_0(W) = \sigma_0(U)$이 Wang 등 (2003)이 사용한 $\sigma(W) = \sigma(U)$과 다름에 있다. 구체적으로 두 군의 표본수 비율이 $k_2 = n_C/n_T (k_2 > 0)$일 때 표본수는 다음과 같다(식 (2.4)에 사용한 상수 k_1을 사용하게 되면 표본수공식이 매우 복잡하게 된다).

$$n_C = k_2 n_T, \quad n_T = \frac{\left\{ z_\alpha \sqrt{k_2(k_2+1)/12} + z_\beta \sqrt{k_2^2\,(p_2 - p_1^2) + k_2\,(p_3 - p_1^2)} \right\}^2}{k_2^2 (p_1 - 0.5)^2}. \tag{2.13}$$

Noether (1987)의 표본수 공식은 확률 p_1만의 함수인 반면에 Wang 등 (2003)의 표본수 공식은 확률 p_1에 덧붙여 식 (2.2)에 나타난 바와 같이 p_2와 p_3의 함수로 표현된다. 따라서 Wang 등 (2003)의 표본수 공식에서는 이 세가지 확률이 주어지지 않는 경우에 어떻게 이 확률들을 추정하여 표본수를 구하는가에 따라 Wang 등 (2003)의 표본수 공식의 유용성이 결정된다 하겠다.

2.3. 표본수 결정에 이용되는 확률의 추정

Noether (1987)는 확률 p_1이 주어지는 경우를 생각하여 추정방법에 대해 언급하지 않았으나, Wang 등 (2003)은 세가지 확률에 대한 구체적인 정보가 없는 경우에 이들의 추정에 대해 설명하고 있다. 우선 Wang 등 (2003)이 제시한 확률의 추정량은 다음과 같다.

$$\widehat{p_1} = \frac{1}{n_C n_T} \sum_{j=1}^{n_T}\sum_{i=1}^{n_C} \psi\{Y_{Ci}, Y_{Tj}\}, \tag{2.14}$$

$$\widehat{p_2} = \frac{1}{n_C n_T (n_C - 1)} \sum_{k=1}^{n_T} \sum_{i \neq j} \psi\{Y_{Ci}, Y_{Tk} \ \text{ and } \ Y_{Cj}, Y_{Tk}\}, \tag{2.15}$$

$$\widehat{p_3} = \frac{1}{n_C n_T (n_T - 1)} \sum_{i \neq j} \sum_{i=k}^{n_C} \psi\{Y_{Ck}, Y_{Ti} \ \text{ and } \ Y_{Ck}, Y_{Tj}\}, \tag{2.16}$$

표 **2.2.** Effect of subject sizes on the estimated probabilities and sample sizes

Pilot study subject size	MSE			Range		No. of missing cases	
	$\widehat{p_1}$	$\widehat{p_2}$	$\widehat{p_3}$	$\mathbf{N}_{Noether}$	$\mathbf{N}_{Wang}$	$\mathbf{N}_{Noether}$	$\mathbf{N}_{Wang}$
5	0.02083	0.03016	0.0299	1349	1196	0	99
10	0.00899	0.01307	0.01284	458	432	95	96
20	0.00487	0.00692	0.00702	312	302	15	15
30	0.00266	0.00384	0.00378	178	174	4	4

$$\text{여기서,}\quad \psi\{Y_{Ci},\, Y_{Tk}\ \text{ and }\ Y_{Cj},\, Y_{Tk}\} = \begin{cases} 1, & \text{if } Y_{Ci} < Y_{Tk} \text{ and } Y_{Cj} < Y_{Tk}, \\ 0, & \text{if elsewhere,} \end{cases}$$
$$\text{단, } 1 \le k \le n_T,\ 1 \le i, j \le n_C,\ i \ne j,$$

$$\psi\{Y_{Ck},\, Y_{Ti}\ \text{ and }\ Y_{Ck},\, Y_{Tj}\} = \begin{cases} 1, & \text{if } Y_{Ck} < Y_{Ti} \text{ and } Y_{Ck} < Y_{Tj}, \\ 0, & \text{if elsewhere,} \end{cases}$$
$$\text{단, } 1 \le k \le n_C,\ 1 \le i, j \le n_T,\ i \ne j,$$

이 확률들은 전단계(pilot)로 계획된 소규모 연구로부터 또는 과거 연구의 자료로부터 추정되어야 하며 Wang 등 (2003)은 각 군에 개체수가 5명인 전단계 연구에서 $\widehat{p_1}$, $\widehat{p_2}$, $\widehat{p_3}$를 구하여 표본수를 계산해 보였다. 확률추정에 있어서 이는 매우 작은 개체수이다. 이러한 확률추정에 필요한 적절한 개체수를 알아보기 위해 모의실험을 실시하였다.

Wetherill (1960)은 대립가설 하에서, 즉 대조군과 치료군의 분포가 각각 $N(0, \sigma_C^2)$, $N(\theta, \sigma_T^2)$일 때 이 확률값을 제시하였다. 이제 $z_e = \theta/(\sigma_C + \sigma_T)$, $\rho_1 = \sigma_T^2/(\sigma_C^2 + \sigma_T^2)$, $\rho_2 = \sigma_C^2/(\sigma_C^2 + \sigma_T^2)$이라 두면, p_1은 표준 정규분포의 누적 확률, p_2와 p_3는 표준 이변량 정규분포의 누적 확률로 다음과 같이 표현된다.

$$p_1 = \Phi(z_e) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_e} \exp\left[-\frac{u^2}{2}\right] du,$$
$$p_2 = \mathrm{BVN}\Phi(z_e, z_e, \rho_1) = \frac{1}{2\pi\sqrt{1-\rho_1^2}} \int_{-\infty}^{z_e}\int_{-\infty}^{z_e} \exp\left[-\frac{u^2 - 2\rho_1 uv + v^2}{2(1-\rho_1^2)}\right] dvdu,$$
$$p_3 = \mathrm{BVN}\Phi(z_e, z_e, \rho_2) = \frac{1}{2\pi\sqrt{1-\rho_2^2}} \int_{-\infty}^{z_e}\int_{-\infty}^{z_e} \exp\left[-\frac{u^2 - 2\rho_2 uv + v^2}{2(1-\rho_2^2)}\right] dvdu.$$

표 2.2는 대조군과 치료군의 분포가 각각 $N(0,1)$, $N(0.5,1)$일 때 전단계 연구의 각 군의 표본수를 각각 5, 10, 20, 30으로 정한 후, 각 경우에서 세가지 확률추정값을 구한 후 이를 10,000번 반복하여 MSE(mean square error)와 이 확률의 $E(\widehat{p_i}) \pm \sqrt{\mathrm{var}(\widehat{p_i})}$에 의한 표본수 범위를 계산한 결과이다. 여기서 정확한 확률값으로 구한 두 방법의 표본수는 $N_{Noether} = 108$이며 $N_{Wang} = 106$이다.

모든 확률을 추정하여 표본수를 계산할 때 정확히 $\widehat{p}_1 = 0.5$가 구해지면 Noether (1987)의 표본수 공식 (2.4)와 Wang 등 (2003)의 표본수 공식 (2.15)에서 표본수를 구할 수 없고, 0.5에 매우 근접한 추정값이 구해지면 표본수가 무한대에 가깝게 된다. 특이할 점은 추정에 사용되는 개체수가 홀수이면 $\widehat{p}_1 = 0.5$가 결코 될 수 없다. 또한 Wang 등 (2003)의 표본수 공식 (2.15)에서 제곱근안의 수치 $k_2^2\left(p_2 - p_1^2\right) + k_2\left(p_3 - p_1^2\right)$가 음수가 되면 표본수를 구할 수 없다. 따라서 Wang 등 (2003)의 표본수 공식에서는 $\widehat{p}_1 = 0.5$이 되거나 제곱근안이 음수인 경우 모두가 발생한다. 표 2.2의 결측의 경우수는 이러한 결과를 뜻하며, 확률추정에 사용되는 개체수가 증가할수록 결측의 경우수는 줄어드는 것을 알 수

표 **2.3.** Sample sizes of Noether (1987) and Wang *et al.* (2003) methods under equal variances

					Power 80%		Power 90%	
θ	σ_C^2=σ_T^2	p_1	p_2	p_3	$\mathbf{N}_{Noether}$	$\mathbf{N}_{Wang}$	$\mathbf{N}_{Noether}$	$\mathbf{N}_{Wang}$
	1	0.55690	0.39083	0.39330	637	634	882	876
0.2	2	0.52880	0.36307	0.36209	2485	2482	3442	3436
	3	0.51952	0.35713	0.34907	5412	5410	7496	7492
	1	0.61877	0.46048	0.46007	147	144	203	198
0.4	2	0.55387	0.38919	0.38849	711	708	984	978
	3	0.54297	0.37700	0.37771	1117	1114	1547	1540
	1	0.66412	0.51400	0.51242	77	74	106	100
0.6	2	0.58844	0.42645	0.42620	264	260	366	360
	3	0.55921	0.39488	0.39424	588	586	815	808
	1	0.70985	0.56808	0.57025	47	44	65	60
0.8	2	0.61414	0.45419	0.45568	159	156	220	214
	3	0.57784	0.41437	0.41497	341	338	472	466
	1	0.76330	0.63596	0.63936	30	28	42	36
1.0	2	0.63721	0.48224	0.48045	110	106	152	146
	3	0.59478	0.43227	0.43430	230	226	318	312

있다.

표 2.2에서 보면 확률추정에 사용된 개체수가 클수록 확률추정값의 MSE가 작아지며 따라서 더욱 신뢰할 수 있는 확률추정값이 구해진다. 모의실험의 각 경우에서 구해진 $\widehat{p}_1$과 이 추정량의 표준편차로부터 구한 최대 표본수와 최소 표본수의 범위는 개체수가 클수록 좁아지며 Wang 등 (2003)의 표본수 범위가 Noether (1987)의 표본수 범위보다 $\widehat{p}_2$와 $\widehat{p}_3$를 추가로 사용하더라도 더욱 좁은 것을 볼 수 있다. 따라서 확률추정시 사용되는 개체수는 홀수이어야 하고 개체수는 클수록 바람직하다.

2.4. Noether (1987)와 Wang 등 (2003)의 표본수 비교

두 가지 공식에 의한 표본수를 비교하기 위해 Wang 등 (2003)에서 채택한 방법을 그대로 적용하여 모의실험 자료로부터 확률을 추정한 후 모의실험의 각 경우마다 표본수를 계산한다. 다시 말하면 대조군은 $N(0,\sigma_C^2)$을 따르는 분포에서 10,000명 개체의 수치를 생성하고 치료군은 $N(\theta,\sigma_T^2)$을 따르는 분포에서 10,000명 개체의 수치를 생성하여 p_1, p_2, p_3를 추정한다. 여기서 채택한 모의실험의 경우는 다음과 같다. 대조군과 치료군에 동일 표본수를 할당할 때 검정력이 가장 커지므로 $k_1 = 0.5, k_2 = 1$로 두었으며, θ는 0.2에서 1까지 0.2씩 변화시킨다. 표 2.3에는 두 군의 분산이 같은 경우의 σ_C^2와 σ_T^2를 1에서 3까지 변화시킨 표본수를, 표 2.4에는 분산이 다른 경우의 σ_C^2와 σ_T^2를 1에서 2까지 변화시킨 표본수를 제시하였다.

표본수 공식의 일반적인 현상으로서 전체 표본수는 θ가 작을수록 $\widehat{p}_1$이 작아져서 대립가설이 귀무가설에 근접한 경우가 되어 표본수가 커지며 또한 분산이 클수록 커진다.

표 2.3에서는 Noether (1987)의 표본수가 Wang 등 (2003)의 표본수보다 약간 크거나 비슷하고, 두 군의 분산이 이질적인 경우에 해당하는 표 2.4에서는 오히려 Wang 등 (2003)의 표본수가 Noether (1987)의 표본수보다 크게 나왔다. 그러한 이유는 그림 2.1로서 이해할 수 있다. 그림 2.1은 각 군의 표본수가 $n_C = n_T = 10$일 때 2.3절에 설명한 정규분포 $N(0,1)$과 $N(\theta,\sigma_T^2)$ 하에서, 즉 두 군의 참분포가 알려진 경우에 정확한 p_1, p_2, p_3를 알 수 있으며 이를 이용하여 대립가설 하에서의 분산을 구해 p_1의

표 **2.4.** Sample sizes of Noether (1987) and Wang *et al.* (2003) methods under unequal variances

						Power 80%		Power 90%	
θ	σ_C^2	σ_T^2	p_1	p_2	p_3	$\mathbf{N}_{Noether}$	$\mathbf{N}_{Wang}$	$\mathbf{N}_{Noether}$	$\mathbf{N}_{Wang}$
0.2	1	2	0.52863	0.42647	0.31127	2514	2576	3482	3592
	2	1	0.53714	0.32027	0.43528	1495	1530	2071	2134
0.4	1	2	0.57059	0.46975	0.35682	414	422	573	588
	2	1	0.56814	0.35460	0.46610	444	452	615	630
0.6	1	2	0.60373	0.50445	0.39485	192	194	266	268
	2	1	0.61283	0.40516	0.51480	162	164	225	226
0.8	1	2	0.63665	0.54028	0.43404	111	110	153	152
	2	1	0.63624	0.43336	0.54036	112	112	154	154
1	1	2	0.67058	0.57809	0.47638	71	70	99	96
	2	1	0.66983	0.47575	0.57676	72	70	99	96

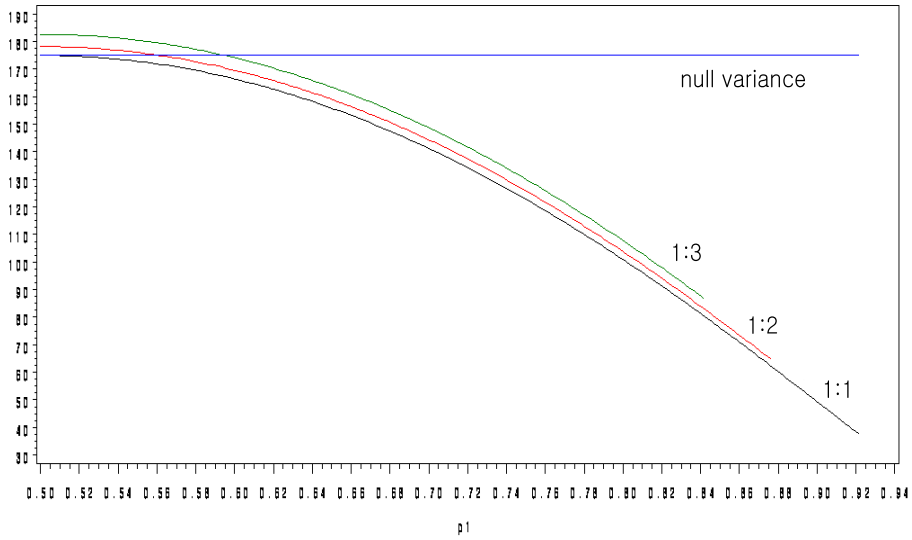

그림 **2.1.** null and alternative variance with different ratios of population variances for two groups

변화에 따라 어떻게 변하는지를 그린 것이다. 대조군과 치료군 분포의 분산이 각각 1 : 1, 1 : 2 그리고 1 : 3로 변경됨에 따라 귀무가설에 가까운 p_1값에서 차이를 보인다 (여기서 p_1은 θ가 커질수록 정비례로 커지므로 p_1을 θ로 대체시켜도 같은 결과가 나온다). 대조군과 치료군의 분포의 분산비가 1 : 1인 경우, p_1값이 커질수록 대립가설 하에서의 분산이 귀무가설 하에서의 분산보다 점점 작아지므로 Noether (1987)의 표본수가 Wang 등 (2003)의 표본수보다 더 커진다. 그러나 대조군과 치료군의 분포의 분산 비율이 1 : 2인 경우에는 대략 $p_1 < 0.565$일 때 대립가설 하에서의 분산이 귀무가설 하에서의 분산보다 커져서 오히려 Wang 등 (2003)의 표본수가 Noether (1987)의 표본수보다 크다. 그러나 $p_1 > 0.565$일 때에는 반대로 Noether (1987)의 표본수가 Wang 등 (2003)의 표본수보다 더 크다. 두 군의 분포의 분산 비율이 1 : 3인 경우에 두 가설 하에서의 분산의 크기가 바뀌는 p_1값이 대략 0.595이며, 분산비가 1 : 3의 표본수는 표에 제시되지 않았으나 두 군 분포의 분산비가 크게 차이가 날수록 더 큰 p_1값에서 두 방법의 표본수의 크고 작음이 바뀐다.

표 **3.1.** The 3×2 and 3×3 contingency tables of efficacy and safety probabilities ($i = T, C$)

(a) 3×2

Efficacy \ Safety	1	2	Total
1	$p_{i,11}$	$p_{i,12}$	$p_{i,1\cdot}$
2	$p_{i,21}$	$p_{i,22}$	$p_{i,2\cdot}$
3	$p_{i,31}$	$p_{i,32}$	$p_{i,3\cdot}$
Total	$p_{i,\cdot 1}$	$p_{i,\cdot 2}$	1

(b) 3×3

Efficacy \ Safety	1	2	3	Total
1	$p_{i,11}$	$p_{i,12}$	$p_{i,13}$	$p_{i,1\cdot}$
2	$p_{i,21}$	$p_{i,22}$	$p_{i,23}$	$p_{i,2\cdot}$
3	$p_{i,31}$	$p_{i,32}$	$p_{i,33}$	$p_{i,3\cdot}$
Total	$p_{i,\cdot 1}$	$p_{i,\cdot 2}$	$p_{i,\cdot 3}$	1

3. 이변량 윌콕슨 검정법의 표본수 결정

본 논문에서 다루는 이변량 순위변수의 윌콕슨 검정통계량에 근거한 표본수 계산방법이 이현학과 송혜향 (2009)에 제시되었다. 따라서 이현학과 송혜향 (2009)에 제시된 기호를 그대로 사용하여 앞 절에서 $p_1 = \Pr(Y_C < Y_T)$로 제시된 것과 다르게 범주형 자료의 경우를 고려해 $\Delta = \Pr(Y_C \le Y_T)$로 표현한다.

효능(efficacy; E)과 안전성(safety; S)을 나타내는 순위 반응변수를 Y_k^E과 $Y_k^S (k = C,\ T)$로 나타내며, 각각 1부터 R까지와 1부터 J까지의 값을 가지므로 각 군의 효능과 안전성의 순위자료는 $R \times J$ 분할표로 표현된다. 표 3.1에는 3×2와 3×3인 자료형태가 제시되었으며, 임상시험의 현장에서는 효능은 순위변수로 측정되고 안전성은 이항변수로 측정되고 경우를 자주 볼 수 있어서 3×2 자료형태를 제시하였다. 따라서 이변량 효능은 $K \times K (K > 2)$ 형태의 자료로 표현되고 이변량 안전성은 2×2 형태의 자료로 표현된다. 3×2인 자료형태는 3×3인 자료형태에서 안전성의 분류항목 1과 2를 합한 것이다.

3.1. 검정가설

Thall과 Cheng (1999)에 의해 임상시험의 목적은 어떤 구체적인 효능의 증가를 얻기 위해 '어느 정도의 부작용을 감수해야 하는가' 또는 부작용을 줄이기 위해서는 '어느 정도까지 효능을 떨어뜨릴 수 있는가'라는 trade-off의 문제로 표현되었다. 이러한 trade-off는 $\boldsymbol{\xi} = (\xi^E, \xi^S)$가 목표모수(target parameter) 벡타이고 효능과 안전성의 차이가 없음에 해당하는 귀무가설 하의 모수벡타가 (0.5, 0.5)일 때, 다음과 같은 경우로 제시된다. 효능과 안전성의 목표모수가 둘 다 증가하는 경우($\xi^E > 0.5$와 $\xi^S > 0.5$)는 임상에서 불가능하며, 효능이 떨어지지 않으면서 안전성이 증가는 경우($\xi^E = 0.5 < \xi^S$)나 안전성이 떨어지지 않으면서 효능이 증가하는 경우($\xi^E > 0.5 = \xi^S$)가 가장 최적이고, 일반적인 경우로는 안전성의 ξ^S만큼 감소에 효능이 ξ^E만큼 증가하거나 효능의 ξ^E만큼 감소에 안전성의 ξ^S만큼 증가하는 것이다. 따라서 효능과 안전성의 이변량 검정에서 대립가설은 여러 trade-off 목표모수의 최소 볼록집합(convex hull)으로 표현되며, 여기서 대립가설에 해당하는 목표모수는 임상의가 제시한다.

3.2. 윌콕슨 검정통계량

윌콕슨 검정통계량은 $\Delta = \Pr(Y_C \le Y_T)$를 각 반응변수의 처리효과 모수로 사용하며, 다변량 다표본 U-통계량(multivariate multi-sample U-statistics)의 공분산을 이용하여 검정하고 표본수를 결정한다. 따라서 구체적인 윌콕슨검정 통계량은 $\boldsymbol{\Delta}$의 일치추정량 $\hat{\boldsymbol{\Delta}}$의 분포, 즉 $\sqrt{n}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})$가 점근적으로(asymptotically) 평균이 $(0, 0)$이고 공분산행렬이 $\boldsymbol{\Sigma}_\Delta$인 이변량 정규분포함을 이용한다.

$$\sqrt{n_C}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \approx \mathrm{BVN}(\mathbf{0}, \boldsymbol{\Sigma}_\Delta), \quad \text{여기서} \quad \boldsymbol{\Sigma}_\Delta = \left[\begin{pmatrix} \sigma_{10}^{EE} & \sigma_{10}^{ES} \\ \sigma_{10}^{ES} & \sigma_{10}^{SS} \end{pmatrix} + k_2 \begin{pmatrix} \sigma_{01}^{EE} & \sigma_{01}^{ES} \\ \sigma_{01}^{ES} & \sigma_{01}^{SS} \end{pmatrix} \right] \tag{3.1}$$

식 (3.1)의 $\mathbf{\Sigma}_\Delta$는 일치추정량인 $\hat{\mathbf{\Sigma}}_\Delta$로 대체한다. 본 논문의 표본수 계산에서는 두 군에 동일 표본수를 고려하여 $k_2 = 1$로 둔다.

이변량 순위변수의 표본수 계산에서 필요한 정보는 효능과 안전성의 이변량 처리효과 벡터 $\mathbf{\Delta} = (\Delta^E, \Delta^S)$와 이에 대한 공분산 행렬 $\mathbf{\Sigma}_\Delta$ 그리고 효능과 안전성 사이의 상관성이다. 이 상관성은 대조군 및 목표모수점마다 여러 범주의 결합분포가 제시되거나, 아니면 여러 범주의 주변확률과 대조군의 효능과 안전성의 global 오즈비 ψ가 제시되는 것이다. 이변량 윌콕슨검정 통계량의 공분산 행렬은 결합확률, 주변확률 그리고 표본수 n_C와 n_T를 이용하여 추정하는데 (Delong 등, 1988), 결합확률은 $R \times J$ 분할표 순위 반응변수의 연관성 척도로서 global 오즈비를 이용하여 결합확률을 구한다 (Agresti, 1984, pp. 15–22). 만약 각 목표모수점에서의 global 오즈비가 제시되지 않는다면 대조군의 오즈비와 동일하다고 가정하고서 대조군의 오즈비를 사용할 수 있다. 임상에서 과거에 진행된 연구로부터 대조군의 결합분포를 알 수 있다. 결합확률을 구하는 자세한 과정은 이현학과 송혜향 (2009)에 제시되었다. 단지 출판과정에서 수정되지 않은 다음 두 결합확률만을 제시하며 이는 음수 결합확률이 발생치 않는 공식이다.

$$p_{C,13} = \begin{cases} \dfrac{\omega_{12} + \sqrt{\omega_{12}^2 + 4(\psi_{C,12} - 1)p_{C,1\cdot}p_{C,\cdot 3}}}{2(\psi_{C,12} - 1)}, & \text{for } \psi_{C,12} \neq 1, \\ p_{C,1\cdot}p_{C,\cdot 3} & \text{for } \psi_{C,12} = 1, \end{cases}$$
$$\text{단, } \omega_{12} = [(\psi_{C,12} - 1)(p_{C,1\cdot} + p_{C,\cdot 3}) - \psi_{C,12}],$$

$$p_{C,31} = \begin{cases} \dfrac{\omega_{21} + \sqrt{\omega_{21}^2 + 4(\psi_{C,21} - 1)p_{C,3\cdot}p_{C,\cdot 1}}}{2(\psi_{C,21} - 1)}, & \text{for } \psi_{C,21} \neq 1, \\ p_{C,3\cdot}p_{C,\cdot 1}, & \text{for } \psi_{C,21} = 1, \end{cases}$$
$$\text{단, } \omega_{21} = [(\psi_{C,21} - 1)(p_{C,3\cdot} + p_{C,\cdot 1}) - \psi_{C,21}].$$

3.3. 표본수 결정

표본수를 결정하기 위해서는 검정력 함수 $\phi(\mathbf{\Delta}) = \Pr\{\hat{\mathbf{\Delta}} \in R(c_\alpha) | \mathbf{\Delta} = (\Delta^E, \Delta^S)\}$가 요구된다. 또한 임상의가 제시한 여러 개의 목표모수 $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$로서 기각영역 $\Omega_a (\subset R^2)$가 정해진다. 그리고 α와 β를 각각 제 1종과 제 2종 오류라 할 때 기각영역과 검정력은 다음을 만족한다.

$$\Pr\left\{\hat{\mathbf{\Delta}} \in R(c_\alpha) | \mathbf{\Delta} = (0.5, 0.5)\right\} = \alpha,$$
$$\Pr\left\{\hat{\mathbf{\Delta}} \in R(c_\alpha) | \boldsymbol{\xi}_i = \left(\xi_i^E, \xi_i^S\right)\right\} = 1 - \beta \quad (i = 1, \ldots, K)$$

따라서 기각영역 Ω_a를 45도 $\left\{\mathbf{\Delta} : \Delta^E = \Delta^S\right\}$로 귀무가설인 (0.5, 0.5) 방향으로 이동하면서 식 (3.1)에 제시된 이변량 정규분포의 이중적분(double integration)으로 제 1종 오류를 만족할 때 멈추면서 구체적인 기각영역과 표본크기 $n(= n_C = n_T)$을 구한다. 효능과 안전성에 동일 중요성을 둔다면 기각영역을 45도 선을 따라 움직이며 만약 중요성을 달리 두고자 한다면 45도가 아닌 각도로 움직일 수 있다.

이변량 정규분포에서 검정력 함수는 두 모수 Δ^E, Δ^S 각각에 대해 모두 증가하므로 여러 $\phi(\boldsymbol{\xi}_1), \ldots, \phi(\boldsymbol{\xi}_K)$에 대해서 표본크기 n은

$$\min_{1 \leq k \leq K} \phi(\boldsymbol{\xi}_K) \geq 1 - \beta$$

를 만족하는 표본수로 결정한다. 이는 각 목표모수에서의 최소 검정력이 정해진 검정력보다 크게 된다면 모든 목표모수에서의 검정력은 정해진 검정력보다 크기 때문이다.

표 **3.2.** Sample sizes of WMW methods with $\psi = 3$ and $\boldsymbol{\theta}_C = (0.75, 0.20)$

Design	Cases	$\boldsymbol{\theta}_T$	$\boldsymbol{\delta}_k$	Power 80%		Power 90%	
				WMW 3×2	WMW 3×3	WMW 3×2	WMW 3×3
One-target	1	(0.95, 0.15)	(0.20, −0.05)	136	158	172	210
	2	(0.90, 0.15)	(0.15, −0.05)	250	290	316	382
	3	(0.95, 0.10)	(0.20, −0.10)	126	158	156	208
Two-target	4	(0.95, 0.15) (0.75, 0.40)	(0.20, −0.05) (0.00, 0.20)	454	610	636	876
	5	(0.90, 0.15) (0.75, 0.40)	(0.15, −0.05) (0.00, 0.20)	642	872	906	1236
	6	(0.95, 0.10) (0.75, 0.40)	(0.20, −0.10) (0.00, 0.20)	562	746	792	1066
Three-target	7	(0.95, 0.15) (0.85, 0.20) (0.75, 0.40)	(0.20, −0.05) (0.10, 0.00) (0.00, 0.20)	824	1008	1162	1446
	8	(0.95, 0.15) (0.80, 0.20) (0.75, 0.40)	(0.20, −0.05) (0.05, 0.00) (0.00, 0.20)	2696	2996	3822	4396
	9	(0.95, 0.15) (0.85, 0.20) (0.75, 0.35)	(0.20, −0.05) (0.10, 0.00) (0.00, 0.15)	992	1294	1402	1898

구체적인 계산과정은 이현학과 송혜향 (2009)에서와 같고, 본 논문의 표본수 결정에서도 프로그램 Mathematica의 FINDROOT 함수를 이용하여 이변량 정규분포의 이중적분을 계속 시행해 가면서 뉴튼-랩슨(Newton-Raphson) 방법으로 표본수를 구한다. 이 과정에서 각 목표변수에 따라 공분산 행렬이 다르게 추정되며 이를 사용하여 검정력을 계산한다.

3.4. 표본수 결과

본 논문에서는 세 목표모수계획까지를 고려하는 여러 연구계획에 대한 표본수를 계산하게 된다. 우선 대조군의 확률 $\boldsymbol{\theta}_C = (\theta_{C,E}, \theta_{C,S})$가 제시되고 k번째 목표모수의 확률 $\boldsymbol{\theta}_T = (\theta_{T,E}, \theta_{T,S})$가 제시되며 이 두 확률의 차이가 $\boldsymbol{\delta}_k = (\delta_{k,E}, \delta_{k,S})$이다.

표 3.2에서는 효능과 안전성의 차이에 따라 표본수가 어떻게 변하는지를 알아보기 위해 대조군의 효능과 안전성의 확률이 $\boldsymbol{\theta}_C = (0.75, 0.20)$이고 global 오즈비가 $\psi = 3$일 때 표본수를 계산하였다. 우선 한 목표모수계획에서 경우 1은 효능이 0.2 증가할 때 안전성이 0.05로 낮아지는 것을 목표로 두었고, 경우 2는 경우 1에서 효능만 0.05만큼 축소한 것이고, 경우 3은 경우 1에서 안전성만 0.05 축소한 경우로 동일한 크기 0.05만큼의 축소이다. 경우 2의 표본수 250은 경우 1의 표본수 136보다 크고, 경우 3의 표본수 126은 경우 1보다 적게 계산되었다. 그 이유는 귀무가설인 $(0.5, 0.5)$와의 거리가 가까울수록 동일 검정력으로 더욱 작은 차이를 검정하기 위해서는 표본수는 커지기 때문이다. 두 목표모수계획이나 세 목표모수계획에서도 비슷한 경향을 보인다.

표 3.3에서는 $\boldsymbol{\delta}_k$가 같은 경우에 $\boldsymbol{\theta}_C$와 $\boldsymbol{\theta}_T$가 각각 0과 1의 극한값에 가까운 경우와 0.5에 가까운 경우에 표본수가 어떻게 변하는지를 알아보기 위해 $\psi = 3$일 때 표본수를 계산하였다. 목표모수와 대조군 사이의 확률의 차이가 같더라도 $\boldsymbol{\theta}_C = (0.40, 0.60)$인 경우가 표본수가 가장 크게 나오는데 이러한 이유는 치료군의 확률이 $\boldsymbol{\theta}_T = (0.60, 0.55)$로 귀무가설 $(0.5, 0.5)$에 가까워져서 분산이 크기 때문이다. 두 목표모

표 **3.3.** Sample sizes of WMW methods with $\psi = 3$

Design	Cases	$\boldsymbol{\theta}_C$	$\boldsymbol{\theta}_T$	$\boldsymbol{\delta}_k$	Power 80%		Power 90%	
					WMW 3×2	WMW 3×3	WMW 3×2	WMW 3×3
One-target	1	(0.75, 0.20)	(0.95, 0.15)	(0.20, −0.05)	136	158	172	210
	2	(0.60, 0.40)	(0.80, 0.35)		198	208	254	270
	3	(0.50, 0.50)	(0.70, 0.45)		214	224	280	290
	4	(0.40, 0.60)	(0.60, 0.55)		228	230	292	298
	5	(0.20, 0.75)	(0.40, 0.70)		220	220	282	282
Two-target	6	(0.75, 0.20)	(0.95, 0.15) (0.75, 0.40)	(0.20, −0.05) (0.00, 0.20)	454	610	636	876
	7	(0.60, 0.40)	(0.80, 0.35) (0.60, 0.60)		644	730	898	1016
	8	(0.50, 0.50)	(0.70, 0.45) (0.50, 0.70)		700	756	976	1040
	9	(0.40, 0.60)	(0.60, 0.55) (0.40, 0.80)		706	740	994	1012
	10	(0.20, 0.75)	(0.40, 0.70) (0.20, 0.95)		630	636	906	910
Three-target	11	(0.75, 0.20)	(0.95, 0.15) (0.85, 0.20) (0.75, 0.40)	(0.20, −0.05) (0.10, 0.00) (0.00, 0.20)	824	1008	1162	1466
	12	(0.60, 0.40)	(0.80, 0.35) (0.7, 0.4) (0.60, 0.60)		1156	1260	1624	1784
	13	(0.50, 0.50)	(0.70, 0.45) (0.60, 0.50) (0.50, 0.70)		1264	1308	1778	1832
	14	(0.40, 0.60)	(0.60, 0.55) (0.50, 0.60) (0.40, 0.80)		1292	1310	1826	1848
	15	(0.20, 0.75)	(0.40, 0.70) (0.30, 0.75) (0.20, 0.95)		1184	1188	1704	1710

수계획이나 세 목표모수계획에서도 비슷한 경향을 보인다.

표 3.4에서는 연관성의 척도인 오즈비 $\psi = \psi_C = \psi_T$에 따라서 표본수의 크기가 어떻게 변하는지 알아보기 위해서 대조군의 효능과 안전성의 확률이 $\boldsymbol{\theta}_C = (0.75, 0.20)$일 때 여러 계획에 대해서 오즈비를 여러가지로 변화시켜 표본수를 계산하였다. 전체적으로 오즈비가 증가할수록 표본수가 커지는데 오즈비가 무한대로 커져도 표본수의 크기는 한 목표모수계획인 경우에서는 변화폭이 거의 없고 두 목표모수계획인 경우와 세 목표모수계획인 경우에서도 변화폭이 크지 않지만 세 목표모수계획인 경우가 두 목표모수계획인 경우보다 표본수의 변화폭이 더욱 크다.

4. 결론

임상시험 또는 다른 분야의 연구에서 반응변수 자료에 이상점이 존재하거나 또는 치우친 분포형태를 보

표 **3.4.** Sample sizes of WMW methods with $\boldsymbol{\theta}_C = (0.75, 0.20)$

Design	$\boldsymbol{\theta}_T$	$\boldsymbol{\delta}_k$	ψ_C	Power 80%		Power 90%	
				WMW 3×2	WMW 3×3	WMW 3×2	WMW 3×3
One-target	(0.95, 0.15)	(0.20, −0.05)	0	118	138	164	198
			1	136	158	172	210
			3	136	158	172	210
			21	136	156	172	208
			∞	136	154	170	206
Two-target	(0.95, 0.15) (0.75, 0.40)	(0.20, −0.05) (0.00, 0.20)	0	172	274	294	476
			1	416	530	590	780
			3	454	610	636	876
			21	478	686	668	964
			∞	482	714	676	1004
Three-target	(0.95, 0.15) (0.85, 0.20) (0.75, 0.40)	(0.20, −0.05) (0.10, 0.00) (0.00, 0.20)	0	322	504	610	894
			1	756	932	1082	1400
			3	824	1008	1162	1466
			21	862	1172	1212	1686
			∞	870	1220	1222	1744

이는데 이러한 분포의 특성을 무시하고 연속변수로 취급하여 표본수를 구하는 것은 바람직하지 않다. 본 논문에서는 이러한 반응변수 자료의 분석에 적합한 윌콕슨 검정법의 표본수 결정공식을 다루었으며 이 표본수 공식은 자료의 순위로부터 구하게 되는 하나 또는 여러 확률에 의해 결정된다. 따라서 자료의 어떤 단조함수 변환에 의해서도 변함이 없는 표본수 결정방법이며, 이상점이 있거나 또는 치우친 분포에 대해서 로버스트한 표본수 결정방법이다.

단변량 순위변수의 표본수 결정방법에서는 Noether (1987)의 방법과 이를 보완한 여러 방법들을 설명하였고, 그 중에 Noether (1987)의 표본수 결정방법과 Wang 등 (2003) 표본수를 비교하였다. 두 방법의 가장 큰 차이점은 표본수 공식에 대립가설 하의 분산의 사용 여부에 있으며 대립가설 하의 분산은 위치모수가 커질수록 귀무가설 하의 분산보다 작아져 Noether (1987)의 표본수가 Wang 등 (2003)의 표본수보다 커졌으나 대조군과 치료군의 분산비가 서로 다르게 되면 오히려 Wang 등 (2003)의 표본수가 더욱 크다. 그러나 Wang 등 (2003)의 표본수 공식에서는 Noether (1987)의 표본수 보다 p_2와 p_3를 알아야 하기 때문에 확률 p_2와 p_3가 알려져 있지 않은 경우에 이 확률들을 어떻게 추정하여 표본수 공식에 사용하는가에 따라 유용성이 결정되며, 확률추정의 개체수가 작으면 이 확률들은 잘못 추정되어질 가능성이 높다하겠다. 따라서 확률추정이 쉽지 않은 경우에는 Noether (1987)의 표본수 공식을 사용하는 것이 바람직하다. 표본수가 더욱 많아짐은 연구수행에 어려움을 뜻할 수 있지만 올바른 검정력과, 유의수준을 확보할 수 있는 장점이 있다.

이변량 순위변수의 표본수 결정방법에서는 이현학과 송혜향 (2009)의 표본수 방법을 가지고 순위변수의 표본수를 여러 경우에 대해 계산하였다. 특히 표 3.3의 한 목표모수계획의 다섯가지 경우의 단변량 표본수를 계산해 보면 단변량 안전성에서 80% 검정력에 대해 2,000명 이상의 표본수가 요구됨을 알 수 있으며, 그러나 이변량의 경우는 158–230명 수준이다. 이변량 순위변수 자료를 단순히 단변량 효능에 대한 검정법에 근거하여 표본수를 결정하는 것은 옳지 않으며, 이변량 순위변수의 표본수 결정방법을 사용하는 것이 더욱 바람직하다.

참고문헌

0|현학, 송혜향 (2009). 이변량 효능과 안전성 이항변수의 표본수 결정방법, <응용통계 연구>, **22**, 341–353.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, John Wiley& Sons, New York.

Delong, E. R., Delong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, *Biometrics*, **44**, 837–845.

Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered 2 × K contingency tables, *Biometrics*, **43**, 471–476.

Haldane, J. B. S. and Smith, C. A. B. (1948). A simple exact test for birth-order effect, *Annals Eugen*, **14**, 117–124.

Mann, H. B. and Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics*, **18**, 50–60.

Nabulsi, N. N., Tamim, H., Mahfoud, Z., Itani, M., Sabra, R., Chamseddine, F. and Mikati, M. (2006). Alternating ibuprofen and acetaminophen in the treatment of febrile children: A pilot study, *BMC Medicine*, **4**, 4.

Noether, G. E. (1987). Sample size determination for some common nonparametric tests, *Journal of American Statistical Association*, **82**, 645–647.

Sundrum, R. M. (1953). The power of Wilcoxon's 2-sample test, *Journal of the Royal Statistical Society*, **15**, 246–252.

Thall, P. F. and Cheng, S. (1999). Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials, *Biometrics*, **55**, 746–753.

Van Dantzig, D. (1951). On the consistency and the power of Wilcoxon's two sample test, *Koninklijke Nederlandse Akademie Van Wetenschappen, Proceedings Serise A*, **54**, 1–9.

Vollandt, R. and Horn, M. (1997). Evaluation of Noether's method of sample size determination for the Wilcoxon-Mann-Whitney test, *Biometrical Journal*, **39**, 823–829.

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.

Wetherill, G. B. (1960). The Wilcoxon test and non-null hypotheses, *Journal of the Royal Statistical Society. Series B (Methodological)*, **22**, 402–418.

Zhao, Y. D., Rahardja, D. and Qu, Y. (2008). Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties, *Statistics in Medicine*, **27**, 462–468.

Sample Size Determination of Univariate and Bivariate Ordinal Outcomes by Nonparametric Wilcoxon Tests

Hae-Gang Park[1] · Hae-Hiang Song[2]

[1]Department of Biostatistics, Medical College, The Catholic University of Korea
[2]Department of Biostatistics, Medical College, The Catholic University of Korea

(Received August 2009; accepted October 2009)

Abstract

The power function in sample size determination has to be characterized by an appropriate statistical test for the hypothesis of interest. Nonparametric tests are suitable in the analysis of ordinal data or frequency data with ordered categories which appear frequently in the biomedical research literature. In this paper, we study sample size calculation methods for the Wilcoxon-Mann-Whitney test for one- and two-dimensional ordinal outcomes. While the sample size formula for the univariate outcome which is based on the variances of the test statistic under both null and alternative hypothesis perform well, this formula requires additional information on probability estimates that appear in the variance of the test statistic under alternative hypothesis, and the values of these probabilities are generally unknown. We study the advantages and disadvantages of different sample size formulas with simulations. Sample sizes are calculated for the two-dimensional ordinal outcomes of efficacy and safety, for which bivariate Wilcoxon-Mann-Whitney test is appropriate than the multivariate parametric test.

Keywords: Sample sizes, ordinal outcomes, univariate and bivariate WMW.

[2]Corresponding author: Professor, Department of Biostatistics, Medical College, The Catholic University of Korea, Seoul 137-701. E-mail: hhsong@catholic.ac.kr