

New Method for Combining P -values in Meta-Analysis

Jeongyeon Seon^a · Dongjae Kim^{a,1}

^aDepartment of Biostatistics, The Catholic University of Korea

(Received August 13, 2013; Revised October 11, 2013; Accepted October 11, 2013)

Abstract

Meta-analysis is used in variety of areas to synthesize the results of previous studies. Among the methods for Meta-analysis, combining p -values is the simplest method; in addition Tippett (1931), Fisher (1932), Stouffer *et al.* (1949), proposed various methods to combine p -values. We propose a new method to combine p -values based on exponential distribution. A Monte Carlo simulation study compares the power of the proposed methods with previous methods.

Keywords: Meta-analysis, combining p -values, multiple comparisons.

1. 서론

특정한 주제에 대해 하나의 연구 논문을 가지고 결론을 내린 경우보다 여러 연구 논문의 각 결과를 하나로 통합하여 결론을 내린 경우가 더 객관적이고 일반화된 결론을 얻을 수 있다. 이러한 경우 메타분석법이 요구되는데, 메타분석법이란 수년간에 걸쳐 축적된 연구 논문들을 요약하고 ‘분석’하는 분석법을 말한다 (Song, 1998). 특히 메타분석법은 상반되는 결과를 제시하는 수많은 연구결과들이 계속 누적되어 갈 때 이 논문들을 객관적으로 평가하고 종합하는 모둠작업을 훌륭히 해낼 수 있는 통계적 방법이다.

대표적인 메타분석법인 p -value의 통합은 여러 연구결과들을 가장 단순히 통합하는 방법으로 각 연구의 측정변수나 검정통계량이 다르다는 사실은 문제가 되지 않고, 정확한 p -value만 있으면 통합이 가능하다는 장점이 있다. 그러나 p -value의 통합은 여러 연구의 p -value를 평균하는 것과 흡사하여 서로 방향이 다른 연구 결과도 그대로 통합하고, 각 연구의 질이나 표본크기로서 가중시켜 통합하지 않고 단순한 방법으로 통합한다는 단점이 있다.

여러 개의 p -value를 통합하는 방법은 메타분석이라고 정의되기 이전에 Tippett (1931)이나 Fisher (1932)에 의해 이미 제시되었다. Tippett은 최소 p -value를 이용하여 종합적인 유의성 검정을 하였고, Fisher는 p -value를 변환하여 카이제곱분포를 이용한 p -value 통합방법을 제시하였다. Stouffer 등 (1949)은 각 연구의 검정통계량이 정규분포한다는 가정 하에 각 p -value를 정규점수로 변환하여 정규분포 p -value 통합 방법을 제시하였고, Zaykin 등 (2002)은 p -value를 모두 이용하지 않고 어떤 기준값 이하의 값만을 이용하여 통합 p -value를 구하는 절삭곱 방법(Truncated Product Method; TPM)을 제시하였다. 이 외에도 Wilkinson (1951), Edgington (1972), Simes (1986) 등이 다양한 p -value의 통합 방법을 제시하였다.

¹Corresponding author: Professor, Department of Biostatistics, The Catholic University of Korea, Banpo-Dong, Seocho-Gu, Seoul 137-701, Korea. E-mail: djkim@catholic.ac.kr

본 연구에서는 귀무가설 하에서 p -value의 값들이 $[0, 1]$ 사이의 균일분포함을 이용하여 지수분포로 근사시켜 p -value를 통합하는 방법을 제시하여 기존에 제시된 p -value 통합 방법들과 비교하였다. 2절에서 기존의 p -value 통합 방법과 지수분포에 기초한 p -value의 통합 방법에 대해 자세히 설명하고, 3절에서는 모의시험을 통해 Fisher의 역카이제곱법, Tippett의 최소 p 방법, Stouffer의 역정규법, Zaykin 등의 절삭곱방법, Simes의 순서통계량 p 방법과 본 연구에서 제시한 지수분포에 기초한 p -value의 통합 방법의 검정력을 비교하였다. 4절의 결론 및 고찰에서는 모의시험의 결과를 요약하여 각 방법들의 장단점을 제시하였다.

2. p -value의 통합방법

통합하고자 하는 연구의 수가 총 L 편이라 하면 p -value 통합에 대한 귀무가설은 ‘ L 개의 연구에서 치료군과 대조군의 효과가 동일하다.’이고, 대립가설은 다음과 같이 3가지가 존재할 수 있다. 여기서, $\theta_1, \theta_2, \dots, \theta_L$ 은 각 연구에서의 치료효과를 나타낸다.

$$H_1 : \text{일부연구에서 } \theta_i > 0,$$

$$H_2 : \text{일부연구에서 } \theta_i < 0,$$

$$H_3 : \text{일부연구에서 } \theta_i \neq 0.$$

본 연구에서는 L 개의 연구가 모두 독립인 경우만 고려할 것이고, 본 연구에서 제시되는 모든 p -value는 우단측 검정에 대한 p -value를 나타낸다.

2.1. Fisher의 역카이제곱법

독립적인 L 개의 연구에서 얻어진 p -value들을 p_i ($i = 1, 2, \dots, L$)라 하자. 귀무가설 하에서 각각의 p_i 들은 구간 $(0, 1)$ 에서 일양분포(uniform distribution)를 따르며, 다음 과정에 따라 $V_i = -2 \ln p_i$ 는 자유도가 2인 카이제곱분포를 따르게 된다.

$$F(v_i) = P[V_i \leq v_i] = P[-2 \ln p_i \leq v_i] = 1 - \left(e^{-\frac{v_i}{2}} \right)$$

$$\therefore f(v_i) = e^{-\frac{v_i}{2}} \sim \exp\left(\frac{1}{2}\right) \equiv \chi_{(2)}^2,$$

여기서 각 연구는 독립이므로,

$$\sum_{i=1}^L V_i = -2 \sum_{i=1}^L \ln p_i \sim \chi_{(2L)}^2 \quad (2.1)$$

가 된다. 통계량 (2.1)에 대한 검정절차는 다음과 같다.

(1) H_1 : 일부연구에서 $\theta_i > 0$

$$-2 \sum_{i=1}^L \ln p_i > \chi_{(2L; \alpha)}^2 \text{ 이면 기각.}$$

(2) H_2 : 일부연구에서 $\theta_i < 0$

$$-2 \sum_{i=1}^L \ln(1 - p_i) > \chi_{(2L; \alpha)}^2 \text{ 이면 기각.}$$

(3) H_2 : 일부연구에서 $\theta_i \neq 0$

$$\text{Max} \left(-2 \sum_{i=1}^L \ln p_i, -2 \sum_{i=1}^L \ln(1 - p_i) \right) > \chi^2_{(2L; \frac{\alpha}{2})} \text{ 이면 기각.}$$

2.2. Stouffer 등의 역정규법

독립적인 L 개의 연구에서 얻어진 p -value들(p_1, p_2, \dots, p_L)을 표준정규점수로 변환한 후, 이들을 합하게 되면 평균이 0이고 분산이 L 인 정규분포를 따르게 된다.

$$z_i = \Phi^{-1}(p_i) \sim N(0, 1), \quad i = 1, 2, \dots, L,$$

$$Z = \frac{\sum_{i=1}^L z_i}{\sqrt{L}} \sim N(0, 1). \quad (2.2)$$

통계량 (2.2)에 대한 검정절차는 다음과 같다.

(1) H_1 : 일부연구에서 $\theta_i > 0$

$$Z > Z_{1-\alpha} \text{ 이면 기각.}$$

(2) H_2 : 일부연구에서 $\theta_i < 0$

$$Z > Z_{\alpha} \text{ 이면 기각.}$$

(3) H_2 : 일부연구에서 $\theta_i \neq 0$

$$|Z| > Z_{\frac{\alpha}{2}} \text{ 이면 기각.}$$

2.3. Tippett의 최소 p 방법

독립적인 L 의 연구에서 얻어진 p -value들을 귀무가설 하에서 일양분포를 따른다. p_1, p_2, \dots, p_L 의 최소값을 $p_{(1)}$ 이라 할 때 유의수준 α 에서 $p_{(1)} \leq 1 - (1 - \alpha)^{1/L}$ 이면 귀무가설을 기각한다.

$$P(\text{Accepting } H_0 | H_0) = \prod_{i=1}^L P(T_i \geq t_{i0} | H_0) \leq (1 - p_{(1)})^L$$

$$\alpha = P(\text{Accepting } H_0 | H_0) \geq 1 - (1 - p_{(1)})^L$$

$$\therefore p_{(1)} \leq 1 - (1 - \alpha)^{\frac{1}{L}} \text{ 이면 } H_0 \text{ 기각.}$$

2.4. Simes의 순위통계량 p 방법

L 개의 p 값 순위통계량($p_{(1)} < p_{(2)} < \dots < p_{(L)}$)에 근거하여 Simes가 제안한 방법은 다음과 같다.

$$\Pr \left[\bigcup_{i=1}^L \left(p_{(i)} \leq \frac{i\alpha}{L} \right) \right] \leq \alpha.$$

이 검정법은 최소한 하나의 i 에 대해 $p_{(i)} \leq i\alpha/L$ 이면 유의수준 α 에서 귀무가설을 기각하게 된다.

2.5. Zaykin 등의 절삭곱방법

독립적인 L 개의 연구에서 얻어진 p -value들 중에 고정된 값 τ 보다 작은 p -value들의 곱을 검정통계량으로 한다. 즉,

$$W = \prod_{i=1}^L p_i^{I(p_i \leq \tau)} \quad (2.3)$$

이다. 여기서 $I(\cdot)$ 는 지시함수이고 통계량 (2.3)의 분포는 다음과 같다. 다음 분포의 유도 과정은 기존 연구 (Zaykin 등, 2002)에 제시되어 있다.

$$\begin{aligned} \Pr(W \leq w) &= \sum_{k=1}^L \Pr(W \leq w | k) \Pr(k) \\ &= \sum_{k=1}^L \binom{L}{k} (1 - \tau)^{L-k} \left[w \sum_{s=0}^{k-1} \frac{(k \ln \tau - \ln w)^s}{s!} I(w \leq \tau^k) + \tau^k I(w \geq \tau^k) \right]. \end{aligned}$$

2.6. 지수분포에 기초한 제안된 p -value의 결합 방법

Tippett의 최소 p 방법과 마찬가지로 독립적인 L 개의 연구에서 얻어진 p -value들을 귀무가설 하에서 일양분포를 따르고, p_1, p_2, \dots, p_L 의 최소값을 $p_{(1)}$ 이라 할 때 다음과 같은 근사분포를 구할 수 있다.

$$\lim_{L \rightarrow \infty} Lp_{(1)} \sim \exp(1). \quad (2.4)$$

식 (2.4)에 제시된 근사분포를 이용한 검정절차는 다음과 같다.

(1) H_1 : 일부연구에서 $\theta_i > 0$

$$Lp_{(1)} < \exp(1; 1 - \alpha) \text{이면 기각.}$$

(2) H_2 : 일부연구에서 $\theta_i < 0$

$$Lp_{(1)} > \exp(\alpha) \text{이면 기각.}$$

(3) H_2 : 일부연구에서 $\theta_i \neq 0$

$$Lp_{(1)} < \exp\left(1; 1 - \frac{\alpha}{2}\right) \text{ 또는 } Lp_{(1)} > \exp\left(\frac{\alpha}{2}\right) \text{이면 기각.}$$

3. 모의시험

본 3절에서는 Fisher의 역카이제곱법, Stouffer 등의 역정규법, Tippett의 최소 p 방법, Simes의 순위통계량 p 방법, Zaykin 등의 절삭곱방법과 본 연구에서 제시한 지수분포에 기초한 p -value 결합 방법의 검정력을 비교하기 위한 모의시험을 시행하였다. 3.1절에서는 귀무가설이 옳은 경우의 난수추출 방법과 대립가설이 옳은 경우의 난수추출 방법에 대해 설명하고, 3.2절에서는 모의시험의 계획과 3.1절에서 설명한 방법을 토대로 시행한 모의시험의 결과에 대해 알아볼 것이다.

3.1. p -value 난수추출 방법

귀무가설이 $H_0 : \mu = \mu_0$ 이고 대립가설이 $H_1 : \mu = \mu_1 > \mu_0$ 인 경우를 생각해보자. 그러면 표본 평균 \bar{x} 를 귀무가설이 옳다는 가정 하에 다음과 같이 표준화 할 수 있다.

$$z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}. \quad (3.1)$$

수식 (3.1)은 표준정규분포 $N(0, 1)$ 을 따르게 된다. 그런데 만약 대립가설이 사실이라면

$$z_1 = \frac{\sqrt{n}(\bar{x} - \mu_1)}{\sigma} \quad (3.2)$$

가 $N(0, 1)$ 을 따를 것이다. 수식 (3.1)과 (3.2)의 차이를 γ 라 하면

$$z - z_1 = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} - \frac{\sqrt{n}(\bar{x} - \mu_1)}{\sigma} = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} = \gamma \quad (3.3)$$

가 된다. 그러므로 대립가설이 사실인데 수식 (3.1)로 표준화를 하였다면, 이는 평균이 γ 이고 분산이 1인 정규분포 $N(\gamma, 1)$ 을 따르게 된다. 이러한 관계를 이용하여 사실인 귀무가설에 대한 p -value와 거짓인 귀무가설에 대한 p -value를 발생시키는 것이다.

먼저, 귀무가설이 사실이란 가정 하에서 각각의 귀무가설에 대한 우단측 검정의 p -value는 다음과 같다.

$$p = P(Z > z) = 1 - \Phi(z) \quad (3.4)$$

여기서, $z = \Phi^{-1}(p)$ 이고, $\Phi(z)$ 는 표준정규분포의 누적확률함수이므로, p 는 범위 $(0, 1)$ 인 일양분포를 따른다. 그러므로 옳은 귀무가설에 대한 p -value는 수식 (3.4)에 따라, 범위 $(0, 1)$ 인 일양분포를 따르는 난수를 발생시켜 얻는다.

또한, 각각의 대립가설이 옳다라는 가정 하에서 우단측 검정의 p -value는

$$\begin{aligned} P(Z > z + \gamma) &= 1 - \Phi(z + \gamma) \\ &= 1 - \Phi(\Phi^{-1}(1 - p) + \gamma) \\ &= \Phi(-\Phi^{-1}(1 - p) - \gamma) \\ &= \Phi(\Phi^{-1}(p) - \gamma) \end{aligned} \quad (3.5)$$

이다. 거짓인 귀무가설에 대한 p -value를 발생시키면, 범위 $(0, 1)$ 인 일양분포를 따르는 난수 p 를 발생시키고 이를 식 (3.5)의 마지막 항과 같이 계산하여 얻는다.

3.2. 모의시험의 계획 및 결과

3.1절에서 설명한 방법으로 난수를 뽑아 10000번 반복하는 모의시험을 시행하였다. Fisher의 역카이제곱법, Stouffer 등의 역정규법, Tippett의 최소 p 방법, Simes의 순위통계량 p 방법, Zaykin 등의 절삭곱 방법, 그리고 지수분포에 기초한 p -value 결합방법의 검정력을 비교하였다.

위에서 언급한 p -value 결합방법들의 검정력을 비교하기 전에 제 1종오류가 잘 제어가 되는지 Table 3.1에 정리하였다. Table 3.1에서 Fisher는 Fisher의 역카이제곱법, Stouffer는 Stouffer 등의 역정규법, Tippett의 Tippett의 최소 p 방법, Simes는 Simes의 순위통계량 p 방법, TP는 Zaykin 등의 절삭곱방법, 그리고 Exp는 본 연구에서 제시한 지수분포에 기초한 p -value의 결합방법을 나타낸다. Table 3.2에서는 연구의 수에 따른 p -value 통합 방법들의 검정력을 비교하였다.

Table 3.1. Power levels when all hypotheses are true

	연구의 수						
	2	5	10	15	20	30	50
Fisher	0.0499	0.0495	0.0514	0.0490	0.0521	0.0468	0.0470
Stouffer	0.0500	0.0519	0.0505	0.0484	0.0529	0.0497	0.0470
Tippett	0.0499	0.0502	0.0527	0.0499	0.0509	0.0505	0.0522
Simes	0.0503	0.0498	0.0533	0.0505	0.0512	0.0504	0.0528
TP	0.0503	0.0502	0.0542	0.0515	0.0509	0.0493	0.0514
Exp	0.0505	0.0503	0.0528	0.0501	0.0510	0.0505	0.0523

Table 3.2. Comparison of power levels

		연구의 수						
		2	5	10	15	20	30	50
$\gamma = 0.8$	Fisher	0.2887	0.5156	0.7600	0.8876	0.9527	0.9921	0.9996
	Stouffer	0.3043	0.5634	0.8144	0.9251	0.9748	0.9968	0.9999
	Tippett	0.2328	0.2882	0.3277	0.3473	0.3708	0.3977	0.4299
	Simes	0.2367	0.3001	0.3472	0.3700	0.3984	0.4308	0.4701
	TP	0.2367	0.3556	0.5237	0.6480	0.7447	0.8628	0.9649
	Exp	0.2349	0.2896	0.3266	0.3479	0.3711	0.3977	0.4301
$\gamma = 1.0$	Fisher	0.3887	0.6828	0.9045	0.9730	0.9950	0.9998	1.0000
	Stouffer	0.4087	0.7305	0.9342	0.9872	0.9983	1.0000	1.0000
	Tippett	0.3111	0.3919	0.4509	0.4805	0.5170	0.5607	0.6069
	Simes	0.3172	0.4070	0.4823	0.5164	0.5623	0.6163	0.6731
	TP	0.3172	0.4866	0.7034	0.8269	0.8976	0.9657	0.9969
	Exp	0.3137	0.3937	0.4514	0.4811	0.5172	0.5608	0.6071
$\gamma = 1.3$	Fisher	0.5521	0.8674	0.9849	0.9991	0.9999	1.0000	1.0000
	Stouffer	0.5768	0.9001	0.9935	0.9998	1.0000	1.0000	1.0000
	Tippett	0.4470	0.5747	0.6592	0.7014	0.7434	0.7961	0.8444
	Simes	0.4568	0.5964	0.7057	0.7557	0.8033	0.8582	0.9170
	TP	0.4568	0.6919	0.8919	0.9644	0.9884	0.9983	1.0000
	Exp	0.4495	0.5762	0.6599	0.7020	0.7436	0.7961	0.8444

Table 3.3은 이상값이 존재하는 경우 각 방법들의 검정력을 비교한 것이다. 연구의 수를 10개로 고정시킨 후 거짓인 귀무가설의 개수를 하나씩 증가시키며 각 방법들의 검정력을 비교하였다. Table 3.4는 Table 3.3과 동일한 경우이나 연구의 수를 20개로 증가시켜 모의시험을 시행한 결과이다.

Table 3.3과 Table 3.4의 모의시험 결과는 임의로 p -value들을 발생한 것으로 값이 큰 p -value가 존재하는 경우 문제점이 생기는 방법들을 알아보기 어렵다. 따라서 Table 3.5는 귀무가설이 거짓인 경우와 확실히 귀무가설이 사실인 경우가 존재하도록 한 모의시험의 결과로 거짓인 귀무가설의 개수만큼 값이 큰 p -value를 의도적으로 생성하여 포함시킨 것을 이용해 검정력을 구한 것이다. Table 3.5는 연구의 수가 10개인 경우의 검정력을 비교한 결과이고, Table 3.6은 동일한 경우에 대해 연구의 수가 20개인 경우의 검정력을 비교한 결과이다.

Table 3.1에서 모든 방법들의 유의수준이 0.05에 가까운 값을 나타냈다. 즉, 모든 방법들이 제 1종 오류를 잘 제어했으므로 검정력을 비교하는 데 무리가 없을 것이다. Table 3.2에서 모든 방법들은 연구의 수와 γ 값이 커질수록 검정력이 커지고 있음을 확인할 수 있다. 특히, Fisher와 Stouffer의 방법이 다른 방법들에 비해 높은 검정력을 나타냈고, 그 다음으로 TP 방법의 검정력이 높게 나타났다. 다중비교 방법

Table 3.3. Comparison of power levels according to the number of alternative hypothesis ($L = 10$)

$\gamma = 1.0$	거짓인 귀무가설의 개수									
	1	2	3	4	5	6	7	8	9	10
Fisher	0.1068	0.1794	0.2800	0.3952	0.5047	0.6178	0.7142	0.7937	0.8607	0.9046
Stouffer	0.0929	0.1530	0.2430	0.3517	0.4767	0.5999	0.7137	0.8102	0.8878	0.9350
Tippett	0.1002	0.1483	0.1900	0.2355	0.2762	0.3159	0.3536	0.3851	0.4186	0.4501
Simes	0.1005	0.1497	0.1935	0.2433	0.2865	0.3301	0.3691	0.4062	0.4450	0.4784
TP	0.0999	0.1700	0.0542	0.2293	0.3272	0.3885	0.4709	0.5415	0.6506	0.6995
Exp	0.1004	0.1487	0.1904	0.2358	0.2766	0.3163	0.3540	0.3859	0.4195	0.4509

Table 3.4. Comparison of power levels according to the number of alternative hypothesis ($L = 20$)

$\gamma = 1.0$	거짓인 귀무가설의 개수									
	2	4	6	8	10	12	14	16	18	20
Fisher	0.1324	0.2720	0.4387	0.6048	0.7591	0.8656	0.9249	0.9680	0.9851	0.9951
Stouffer	0.1137	0.2267	0.3822	0.5530	0.7230	0.8556	0.9288	0.9748	0.9915	0.9978
Tippett	0.1117	0.1692	0.2249	0.2355	0.2775	0.3239	0.3742	0.4073	0.4489	0.5211
Simes	0.1124	0.1735	0.2310	0.2891	0.3404	0.3937	0.4350	0.4841	0.5250	0.5622
TP	0.1251	0.2382	0.3562	0.4635	0.5797	0.6743	0.7353	0.8197	0.8607	0.8970
Exp	0.1117	0.1694	0.2250	0.2776	0.3241	0.3744	0.4074	0.4491	0.4887	0.5213

에서 비롯된 Tippett, Simes의 방법과 본 연구에서 제시된 방법은 앞에 언급된 방법들과 달리 검정력이 높지 않았다. 특히, $\gamma = 0.8$ 인 경우에는 0.5도 넘지 못하는 낮은 검정력을 보였다. 또한 본 연구에서 제시한 방법과 Tippett 방법의 검정력이 거의 일치하고 연구의 수가 많아질수록 그 차이가 점점 줄어드는 것을 확인 할 수 있었다.

Table 3.3에서 거짓인 귀무가설의 개수가 증가할수록 검정력도 증가하는 양상을 보인다. 귀무가설의 개수가 적은 경우에는 모든 방법들의 검정력이 매우 낮게 나타났지만 Fisher와 Stouffer의 방법은 거짓인 귀무가설의 개수가 전체 연구 수의 절반을 넘어가면서부터 높은 검정력을 나타냈다. Table 3.4의 결과 또한 연구의 수가 10개인 경우인 Table 3.3의 결과와 동일한 양상이 나타났다.

Table 3.5의 검정력은 전반적으로 Table 3.3의 검정력보다 낮아졌다. 거짓인 귀무가설의 수가 연구 수의 절반보다 많을 때는 나머지 옳은 귀무가설에 대한 *p*-value를 모두 일반적인 경우보다 더 큰 값으로 생성하여 포함시켰다. 또한, 큰 *p*-value는 0.5와 1.0 사이에 난수를 발생시켜 얻었다. 그 결과, 앞의 결과와 비교했을 때 모든 방법들의 검정력이 낮아졌다. 앞에서 다른 방법들에 비해 항상 높은 검정력을 보였던 Fisher와 Stouffer의 방법은 거짓인 귀무가설의 개수가 적을 때에 오히려 다른 방법들보다 검정력이 낮게 나타났다. Fisher의 방법은 거짓인 귀무가설의 개수가 연구 수의 절반을 넘어가면서부터 검정력 다시 높아졌고, Stouffer의 방법은 거짓인 귀무가설의 개수가 9개와 10개일 때 제일 높은 검정력이 나타났다. 거짓인 귀무가설의 개수가 5개 이하인 경우에는 TP의 검정력이 제일 높게 나타나고 나머지 방법들의 검정력은 비슷하게 나타났다. 이 결과를 이상값이 존재하는 경우에는 Fisher와 Stouffer의 방법보다 TP 방법의 검정력이 더 높았다. Table 3.6의 검정력 역시 Table 3.4의 검정력보다 낮게 나타났고, 연구의 수를 10개로 하여 동일한 방법으로 모의시험한 Table 3.5의 결과와 동일한 양상이 나타났다.

4. 결론 및 고찰

본 연구에서는 메타분석 방법 중 가장 단순한 방법인 *p*-value를 통합하는 여러가지 방법들을 알아보고 지수분포를 이용하여 *p*-value를 통합하는 새로운 방법을 제안하였다. 기존의 *p*-value 통합 방법들과 본

Table 3.5. Comparison of power levels according to the number of alternative hypothesis when large p -values containe ($L = 10$)

$\gamma = 1.0$	거짓인 귀무가설의 개수									
	1	2	3	4	5	6	7	8	9	10
Fisher	0.0711	0.0931	0.1259	0.1641	0.1971	0.3630	0.5397	0.7007	0.8251	0.9046
Stouffer	0.0511	0.0516	0.0498	0.0550	0.0524	0.1928	0.4221	0.6567	0.8379	0.9350
Tippett	0.0954	0.1408	0.1800	0.2197	0.2573	0.3010	0.3434	0.3789	0.4157	0.4501
Simes	0.0953	0.1419	0.1826	0.2251	0.2642	0.3113	0.3563	0.3989	0.4414	0.4784
TP	0.0871	0.1408	0.1863	0.2654	0.3126	0.4144	0.5006	0.5687	0.6397	0.6995
Exp	0.0931	0.1412	0.1804	0.2200	0.2576	0.3014	0.3438	0.3797	0.4163	0.4509

Table 3.6. Comparison of power levels according to the number of alternative hypothesis when large p -values containe ($L = 20$)

$\gamma = 1.0$	거짓인 귀무가설의 개수									
	2	4	6	8	10	12	14	16	18	20
Fisher	0.0790	0.1247	0.1761	0.2322	0.2996	0.5633	0.7907	0.9243	0.9755	0.9951
Stouffer	0.0553	0.0576	0.0627	0.0678	0.0768	0.3306	0.6782	0.9105	0.9816	0.9978
Tippett	0.1073	0.1605	0.2131	0.2623	0.3071	0.3610	0.3984	0.4434	0.4859	0.5211
Simes	0.1078	0.1638	0.2177	0.2706	0.3205	0.3774	0.4235	0.4758	0.5216	0.5622
TP	0.1066	0.1907	0.2799	0.3702	0.4718	0.6024	0.7080	0.7976	0.8520	0.8970
Exp	0.1073	0.1607	0.2132	0.2624	0.3073	0.3612	0.3985	0.4436	0.4860	0.5213

연구에서 제안한 방법의 검정력을 모의시험을 통해 비교하였다.

일반적인 경우, Fisher가 제안한 방법인 역카이제곱법과 Stouffer 등이 제안한 방법인 역정규법의 검정력이 제일 높았다. 그 다음으로 Zaykin 등이 제안한 절삭법 방법의 검정력이 높았고, 다중비교에서 비롯된 Tippett이 제안한 최소 p 방법, Simes가 제안한 순위통계량 p 방법과 본 연구에서 제안한 지수분포에 기초한 p -value의 통합 방법이 비슷한 검정력을 보였다. 또한 Tippett의 방법과 본 연구에서 제안한 방법의 검정력은 매우 비슷하였으며 연구의 수가 많아질수록 검정력이 거의 같아지는 현상을 나타냈다. 모든 방법들의 공통적인 현상으로 연구의 수가 많을수록 그리고 γ 값이 커질수록 검정력이 높아지는 것을 알 수 있었다.

이상값이 존재하는 경우에는 Fisher와 Stouffer가 제안한 방법의 검정력이 제일 낮았으며 Zaykin 등이 제안한 절삭법 방법의 검정력이 제일 높게 나타났다. 일반적인 경우와 마찬가지로 나머지 방법들은 비슷한 수준의 검정력을 보였다.

본 연구에서 여러가지 p -value 통합 방법의 검정력을 비교해 본 결과, 본 연구에서 제안한 방법은 이상값이 존재하는 경우, Fisher와 Stouffer가 제안한 방법보다 높은 검정력을 보였다. 일반적인 경우에는 Fisher가 제안한 방법과 Stouffer가 제안한 방법의 검정력이 높았으나 이상값이 존재하는 경우에는 Zaykin 등이 제안한 절삭법 방법의 검정력이 더 높았다. 하지만 이상값이 존재하는 경우에, 거짓인 귀무가설의 개수가 적을 때의 검정력은 매우 낮은 검정력을 보여 이를 해결하기 위한 새로운 연구가 필요할 것으로 보인다.

References

- Edgington, E. S. (1972). An additive model for combining probability values from independent experiments, *Journal of Psychology*, **80**, 351–363.

- Fisher, R. A. (1932). *Statistical Methods for Research Workers*, Oliver and Boyd, London.
- Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika*, **73**, 751–754.
- Song, H. H. (1998). *Meta-Analysis for Study of Medicine, Nursing, and Social Science*, ChungMunGak, Korea.
- Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. and Williams, R. M. (1949). *The American Soldier, Vol 1. Adjustment During Army Life*, Princeton University Press, Princeton.
- Tippett, L. H. (1931). *The Methods of Statistics*, Williams and Norgate, London.
- Wilkinson, B. (1951). A statistical consideration in psychological research, *Psychol Bull*, **48**, 156–158.
- Zaykin, D., Zhivotovsky, L., Westfall, P. H. and Weir, B. S. (2002). Truncated product method for combining p -values, *Genetic Epidemiology*, **22**, 170–185.

메타분석에서 새로운 P -Value 결합 방법

선정연^a · 김동재^{a,1}

^a가톨릭대학교 의학통계학과

(2013년 8월 13일 접수, 2013년 10월 11일 수정, 2013년 10월 11일 채택)

요약

메타분석은 이전의 연구들의 결과를 결합하기 위해 많은 분야에서 널리 사용된다. 메타분석 방법들 중에 p -value의 통합은 가장 단순한 방법이며 Tippett (1931), Fisher (1932) Stouffer 등 (1949) 등이 p -value를 통합하는 다양한 방법들을 제안하였다. 이 논문에서 우리는 지수분포에 기초한 새로운 p -value의 통합 방법을 제안하였다. 또한, 몬테카를로 모의시험 연구를 통해 제안된 방법과 이전의 방법들의 검정력을 비교하였다.

주요용어: 메타분석, p -value 통합, 다중비교.

¹교신저자: (137-701) 서울 서초구 반포동, 가톨릭대학교 의학통계학과, 교수. E-mail: djkim@catholic.ac.kr