Journal of the Korean Data & Information Science Society
2013, **24**(5), 1063–1076

http://dx.doi.org/10.7465/jkdi.2013.24.5.1063
한국데이터정보과학회지

고차원자료에서의 다중검정의 활용[†]

장원철[1]

[1]서울대학교 통계학과

접수 2013년 7월 25일, 수정 2013년 8월 25일, 게재확정 2013년 9월 9일

요 약

현대 과학기술의 발전으로 빅데이터의 시대가 도래하였다, 이러한 빅데이터는 여러가지 과학적 문제에 대한 해답을 제공하지만 반면에 이로 인해 새로운 도전에 직면하고 있다. 마이크로어레이 자료와 같은 고차원자료는 이러한 빅데이터에서 흔히 볼 수 있는 유형중의 하나이다. 본 논문에서는 고차원 자료분석에 많이 쓰이고 있는 대역검정과 동시검정, 그리고 이의 응용에 대한 소개를 한다.

주요용어: 고차원자료, 다중검정, 대역검정, 동시검정, 오발견율.

1. 서론

21세기는 빅데이터의 도래로 많은 분야의 연구들이 새로운 통계분석방법을 요구한다. 현재 빅데이터 분석에 가장 유용한 분석방법들로는 다중검정방법, 구체적으로 대역검정 (global testing)과 동시검정 (simultaneous testing)을 들 수 있겠다. 예를 들면 생물정보학에서 마이크로어레이 자료 분석의 경우 유전자의 갯수가 자료의 갯수보다 많고 각각의 유전자에 대한 발현여부를 통계적 가설검정을 이용하여 알고자 할 때 연구자들은 유전자의 갯수만큼 적게는 수천에서 많게는 수만개의 가설검정을 동시에 행하게 된다. 이러한 다중비교에서 오류 (error rate)의 조정은 필수적이다. 기존의 가설검정방법을 오발견 (false discovery)에 대한 고려없이 사용할 경우, 즉 각각의 유전자에 대한 유의수준 0.05를 이용한 가설검정을 행할 경우, 실제로는 귀무가설들이 모두 참일 경우라도 100개의 가설 중에 평균적으로 5개의 오발견을 하게 된다. 기존의 다중검정에서는 Bonferroni 수정과 같은 FWER (familywise error)의 조절을 통하여 이런 오류를 방지하고자 했으나 빅데이터의 적용에 있어서 너무 보수적인 검정결과, 즉 실제로 기각해야 할 귀무가설보다 아주 적은 수의 기각이 이루어지는 단점이 있었다. 이러한 단점을 보완하기 위해 Benjamini와 Hochberg (1995)는 다른 형태의 오류인 FDR (false iscovery rate)를 제안하고 그의 조정을 통한 다중검정 방법을 제시했다. Google Scholar를 이용한 이 논문인용 횟수 조회 결과는 현재 15,238회이며 Cox의 비례위험모형 (proportional hazard model), Kaplan과 Meier의 생존함수추정 논문들과 더불어 통계학에서 가장 많이 인용되는 논문의 하나로 꼽힌다 (Wit, 2010).

FDR 방법은 사실 1980년대 의학통계에서 쓰였던 다중검정 방법에서 그 유래를 찾을 수 있다. 이후 이 방법론은 다양한 분야의 연구, 신호처리에서 웨이블렛 함수의 분계점 선택 (threshold selection), 천체물리학에서 우주배경복사 (cosmic microwave background)를 이용한 빅뱅이론의 확인, 마이크로어레이 자료의 분석 등에 성공적으로 사용되었다 (Lindsay 등, 2004). Table 1.1은 FDR논문을 인용한 논문들중 가장 많이 인용된 10개의 논문들의 피인용횟수와 관련분야를 정리한것이다 (Wit, 2010).

[†] 이 연구는 서울대학교 신임교수 연구정착금으로 지원되는 연구비에 의하여 수행되었음.
[1] (151-747) 서울특별시 관악구 관악로 1번지, 서울대학교 통계학과, 부교수. E-mail: wcjang@snu.ac.kr

Table 1.1 10 most cited papers that cite Benjamini and Hochberg (1995)

Rank	Article citing Benjamini and Hochberg (1995)	Number of citations	Areas
1	Tusher et al. (2001)	6196	genetics
2	Purcell et al. (2007)	3902	genetics
3	Storey and Tibshirani (2003)	3045	genetics
4	Weisberg et al. (2003)	2659	medicine
5	Genovese et al. (2002)	2001	neuroscience
6	Storey (2002)	1558	statistics
7	Benjamini and Yekutieli (2001)	1413	statistics
8	Willkinson (1999)	982	psychology
9	Patti et al. (2003)	861	medicine
10	Benjamini et al. (2001)	836	behavioral science

이와 같은 다중검정을 행하기 전 많은 경우 분산분석에서의 F-검정처럼 유의한 가설이 하나라도 있는냐를 검정하는 것도 또 다른 관심사항이다. 이러한 대역검정 (global test)은 유전학에서 SNP (single nucleotide polymorphism)분석, 혹은 정보이론 (informatuon theory)에서 통신채널 (communications channel)간의 신호전송 여부를 분석하는 경우에 유용하게 쓰인다.

본 논문에서는 이러한 여러종류의 다중검정 방법론들의 최근동향과 빅데이터의 응용에 대해 소개하고자 한다. 2절에서는 다중검정을 설명하기 위한 통계적 모형을 소개하고 이러한 모형으로 설명될수 있는 과학적 연구가설에 대해 알아본다. 3절에서는 대역검정에 대한 최근의 일련의 연구 결과들을 소개하고 4절에서는 동시검정에 대하여 알아본다. 마지막으로 5절에서는 앞으로의 연구방향에 대한 토의를 한다.

2. 모형

우선 아래와 간단한 같은 모형을 고려하자.

$$y_j = \mu_j + z_j, \quad j = 1, \ldots, n. \tag{2.1}$$

여기서 $z_j \sim N(0,1)$이다.

우리는 다음과 같은 n개의 귀무가설을 고려할 수 있다.

$$H_{0,i} : \mu_i = 0.$$

위와 같은 귀무가설은 사용되는 몇가지 예를 들자면 아래와 같다

- 천문학과 뇌 영상과학과 같은 분야에서 y_i는 각각의 영상자료의 화소에서 관측되는 값으로 생각할 수 있다. 이와 같은 영상자료 분석에서 많은 화소에서 관측되는 값은 0으로 알려져 있다.
- 마이크레어레이 분석에서 각 유전자별로 생성되는 이표본 검정통계량을 y_i로 고려한다. 많은 유전자분석에서 아주 작은 수의 유전자가 질병과 관련이 있다고 믿어진다.
- 전기공학에서 신호처리를 위한 웨이블렛함수를 사용하는 경우에는 웨이블렛의 모수 추정치를 y_i로 고려할 수 있다. 웨이블렛을 이용한 함수추정의 경우 아주 작은 수의 웨이블렛 계수만 0이 아니라고 믿어진다.

다중검정에 있어서 중요한 관심사는 아래 3가지로 요약할 수 있다 (Jin, 2008).

1. 대역검정: 수많은 검정중에서 유의한 가설이 하나라도 존재하는가?

2. 유의가설의 비율: 얼마나 많은 가설이 유의한가?

3. 동시검정: 유의한 가설과 그렇지 않은 가설들은 어떤 가설들인가?

이 논문에서는 1번째와 3번째에 대한 최근의 연구결과에 대해 중점적으로 소개를 하고자 한다. 2번째 문제에 관심이 있는 독자는 Jin (2008)을 참조하기를 권한다.

3. 대역 검정

이 절에서는 대역검정에 대한 최근의 연구를 소개하고자 한다 (Arias 등, 2011). 대역검정은 흔히 **건초더미에서 바늘 찾기 문제 (finding a needle in a haystack)**로 비유될수 있는데 이 경우 귀무가설은

$$H_0 = \bigcap_{i=1}^{n} H_{0,i}.$$

으로 주어지고 여기서 $H_{0,i} : \mu_i = 0$이다.

문제의 단순화를 위하여 대립가설에서는 단 한개의 모평균 $\mu_i = \mu_0 > 0$이고 나머지 모평균들은 0이라 가정하자.

이러한 귀무가설을 검정하기 위해서는 다음과 같은 검정방법들이 사용된다.

- Bonferroni 방법
- 카이제곱 검정
- Higher Criticism
- Adaptive Neyman 검정

3.1. Bonferroni 방법

Bonferroni 방법은 주어진 유의수준 α를 지키기 위해서 각각의 귀무가설 $H_{0,i}$를 유의수준 α/n에서 검정을 시행하고 만약 개개의 귀무가설중에서 하나라도 기각할 경우 전체 귀무가설 H_0를 기각한다. 이 방법이 주어진 유의수준 α를 지키는 것은 아래의 간단한 부등식으로 설명할 수 있다.

$$\begin{aligned} \mathbb{P}_{H_0}(\text{Type I Error}) &= \mathbb{P}\left(\bigcup_{i=1}^{n}\{p_i \le \alpha/n\}\right) \\ &\le \sum_{i=1}^{n}\mathbb{P}(p_i \le \alpha/n) = \alpha. \end{aligned}$$

Bonferroni 방법을 검정통계량 또는 주어진 모형 (2.1)의 y_i를 이용하면 아래와 같이 표현할 수 있다.

$$\min_i p_i \le \alpha/n \Longleftrightarrow \max_i y_i \ge z_{\alpha/n}.$$

이 경우 임계값 $t = z_{\alpha/n}$이 얼마나 큰지 알고 싶다면 Mills ratio을 이용하여 아래와 같은 관계식을 유도할 수 있다.

$$\frac{\phi(t)}{t} \approx \mathbb{P}(Z > t) = \alpha/n.$$

만약 α를 고정하면 충분히 큰 n에 대해서 아래와 같은 근사식이 주어진다.

$$t = z_{\alpha/n} \approx \sqrt{2\log n}\left(1 - \frac{\log\log n}{4\log n}\right) \approx \sqrt{2\log n}.$$

여기서 주목해야 할 사실은 우변의 값이 더 이상 α에 의존하지 않는다는 사실이다. 따라서 검정통계량을 이용하여 Bonferroni 방법을 적용한다면 다음과 같은 경우 귀무가설을 기각한다.

$$\max_i y_i > \sqrt{2\log n}.$$

그러므로 Bonferroni 방법을 사용할 경우 μ_0의 값이 임계값 ($\sqrt{2\log n}$)보다 클 경우 표본크기가 증가하면 검정력은 1으로 수렴한다. 하지만 임계값보다 μ의 값이 작다면 검정력은 α로 수렴하는 것을 쉽게 보일 수 있다. 정리하자면 Bonferroni 방법은 아주 적은 수의 신호 (μ_i)들이 0이 아니고 그 값들이 임계값보다 클 경우 적합한 검정방법으로 추천할 수 있다.

3.2. 카이제곱 검정

만약 모형 (2.1)에서 많은 신호들이 0이 아니지만 그 값은 상대적으로 작을 경우 Bonferroni 방법을 사용하는 것은 효율적이지 않다. 이 절에서는 아래와 같은 보다 일반적인 대립가설을 고려한다.

$$H_1 : \text{적어도 하나의 } \mu_i \neq 0.$$

위와 같은 가정아래서 우리는 $\|y\|^2 = \sum_i y_i^2$ 이 크면 귀무가설을 기각하는 것을 고려할 수 있다. 구체적으로, 귀무가설하에서 $\|y\|^2 \sim \chi^2_n$이므로 우리는 다음과 같은 검정통계량을 고려한다.

$$Y = \frac{\|y\|^2 - n}{\sqrt{2n}}$$

또한 대립가설하에서 $\|y\|^2 = \sum(\mu_i + z_i)^2$가 비중심 카이제곱분포를 따르므로

$$\frac{\|y\|^2 - (n + \|\mu\|^2)}{\sqrt{2n + 4\|\mu\|^2}} \sim N(0,1)$$

을 보일 수 있다. 여기서 $\mu = (\mu_1, \ldots, \mu_n)$이며 $\|\mu\|^2 = \sum_{i=1}^n \mu_i^2$이다.

만약 신호대 잡음비를 다음과 같이 정의하면

$$\theta = \|\mu\|^2/\sqrt{2n},$$

카이제곱 검정통계량은 귀무가설과 대립가설하에서 다음과 같은 분포를 따름을 쉽게 알 수 있다.

$$H_0 : Y \sim N(0,1) \text{ vs } H_1 : Y \sim N\left(\theta, 1 + \frac{4\theta}{\sqrt{2n}}\right)$$

따라서 카이제곱 검정방법은 신호대 잡음비가 1보다 많이 클 경우 검정력도 크고 반대로 θ가 1보다 작을 경우 검정력도 현저히 떨어진다는 것을 알 수 있다, 결론적으로 카이제곱 검정방법을 이용할 경우 신호의 크기 ($\|\mu\|^2$)를 잡음의 크기 ($\sqrt{2n}$)와 비교하면 검정력이 얼마나 좋은지를 알 수 있다.

아래 예들을 살펴보면 Bonferroni 방법과 카이제곱 검정방법은 어떻게 적용범위가 다른지를 이해할 수 있다.

예제 3.1 $\sqrt{n}$개의 μ_i가 10이고 나머지는 모두 0이라 가정하자. 표본크기 (n)가 충분히 큰 경우에는 $10 < \sqrt{2\log n}$이므로 Bonferroni 방법은 검정력이 거의 없음을 알 수 있다. 반면에 신호대 잡음비는 $\theta = \|\mu\|^2/\sqrt{2n} = (100\sqrt{n})/\sqrt{2n} \gg 1$이므로 카이제곱 검정을 사용할 경우 검정력이 매우 높음을 알 수 있다.

예제 3.2 $n^{1/5}$개의 μ_i가 $\sqrt{4\log n}$이고 나머지는 0이라 가정하자. 이 경우 신호대 잡음비는 $\theta = \|\mu\|^2/\sqrt{2n} = (n^{1/5}4\log n)/\sqrt{2n}$이므로 n이 커지면 0으로 수렴한다. 따라서 이 경우에는 카이제곱 검정을 사용하면 검정력이 거의 없음을 알 수 있다. 반면에 $\sqrt{4\log n} > \sqrt{2\log n}$이므로 Bonferroni 방법은 효과적임을 알 수 있다.

3.3. Higher Criticism

Donoho와 Jin (2004)은 논문주제의 동기 부여를 설명하기 위해 Donoho가 Princeton 대학생 시절에 있었던 John Tukey에 관한 일화를 소개했다. 1976년 John Tukey는 수업시간에 아래와 같은 심리학실험을 언급했다.

"심리학자 A는 연구과제의 일부로 250개의 가설검정을 한 결과 그중 11개가 유의수준 5% 에서 귀무가설을 기각하는 것을 발견하고 매우 기뻐했다. 하지만 통계학자 B는 모든 귀무가설이 참이라도 평균적으로 250× 0.05 =12.5개의 유의한 결과가 나올것이라는 걸 지적하고 11개만 유의하다고 나온건 오히려 실망스러운 결과라는걸 지적했다"

Tukey는 이러한 문제점을 풀기위해 아래와 같은 검정통계량을 사용하여 2단계 유의수준 검정 (second-level significance testing)을 할 것을 제안한다.

$$HC_{\alpha,n} = \frac{(\#\text{ significant at level }\alpha) - n\alpha}{\sqrt{n\alpha(1-\alpha)}}.$$

Tukey는 위의 검정 통계량의 값이 2보다 크다면 1단계에서 기각한 귀무가설중 일부는 실제로 유의하다는 결론을 내릴 것을 권했다. Donoho와 Jin (2004)은 이를 확장하여 고정된 유의수준을 사용하는 대신 유의수준은 특정 구간 $(0, \alpha_0)$에 있는 하나의 점이라 가정하고 Higher Criticism이라는 새로운 검정통계량을 제시했다.

$$HC_n^* = \max_{0<\alpha\le\alpha_0} HC_{\alpha,n}.$$

Higher Criticism을 설명하기 위해 Donoho와 Jin (2004)는 다음과 같은 모형을 고려했다.

$$H_0 : X_i \overset{iid}{\sim} N(0,1) \text{ vs } H_1 : X_i \overset{iid}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n N(\mu_n, 1).$$

여기서

$$\begin{aligned} \epsilon_n &= n^{-\beta}, & \frac{1}{2} < \beta < 1, \\ \mu_n &= \sqrt{2r\log n}, & 0 < r < 1, \end{aligned}$$

라 하자. 여기서 β는 신호의 희소성 (sparsity), r은 신호의 크기를 나타낸다. 즉 β가 클수록 신호의 갯수가 적어지고 r이 클수록 신호의 크기는 커진다. 3절 처음에서 대역 검정에 사용했던 가정, 즉 대립가설하에서 하나의 신호만 양수인 경우는 $\beta = 1$와 $r = 1$인 경우임을 알 수 있다.

실제 대역검정문제에 Higher Criticism을 적용하기위해 Donoho와 Jin (2004)은 다음과 같은 알고리즘을 제안했다.

1. 먼저 각각의 귀무가설하에서 유의확률, $p_1, \ldots, p_n$을 다음과 같이 계산한다.

$$p_i = 1 - \Phi(X_i) = \mathbb{P}(Z > X_i)$$

2. 유의확률의 순서통계량, $p_{(1)} \le p_{(2)} \le \cdots \le p_{(n)}$을 이용하면 Higher Criticism 통계량은 다음과 같이 표현할 수 있다.

$$HC_n^* = \max_{1\le i\le\alpha_0 n} \frac{\sqrt{n}(i/n - p_{(i)})}{\sqrt{p_{(i)}(1-p_{(i)})}} = \max_{\frac{1}{n}\le t\le\alpha_0} \frac{\sqrt{n}(F_n(t) - t)}{\sqrt{t(1-t)}}$$

여기서 $F_n(t)$는 유의확률의 경험적 누적분포함수 (empirical cumulative density function)이며 $t = p_{(i)}$이다.

3. 만약 $HC_n^* \geq \sqrt{(1+\epsilon)2\log\log n}$이면 귀무가설 H_0를 기각한다. 실제문제에서는 ϵ=1을 사용한다.

Donoho와 Jin (2004)는 Higher Criticism이 적절한 범위의 r와 β에서 최적의 검정력을 가질수 있다는 걸 증명하였다, 먼저 소개한 2가지 방법과 비교할 경우 $\beta \in (0, 1/2)$일 경우, 즉 신호의 크기는 작지만 상대적으로 많은 수의 신호가 존재하는 경우에는 카이제곱 방법이 최적이며 $\beta \in (1/2, 1)$인 경우에는 Higher Criticism의 사용을 권장한다. 반면에 Bonferroni 방법은 $\beta \in (3/4, 1)$ 인 경우 (신호의 갯수가 상대적으로 아주 적은 경우)에 효율적이다.

3.4. Adaptive Neyman 검정

Fan (1996)은 카이제곱 검정의 검정력을 높이기 위해 μ에서 일부분만 사용하여 검정통계량을 만들것을 제안했다. 구체적으로, 만약 대부분의 신호가 처음 m번째에 있다면 $\sum_{i=1}^{m} y_i^2$이 검정 통계량으로 적합할 것이다. 하지만 m은 실제 상황에서는 알 수 없기 때문에 Fan은 아래와 같은 목적함수를 최소로 만드는 값을 m의 추정치로 사용할 것을 추천했다.

$$\widehat{m} = \underset{m:1\leq m\leq n}{\operatorname{argmax}} \left\{ \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (y_i^2 - 1) \right\}.$$

Fan은

$$T_{AN}^* = \left\{ \frac{1}{\sqrt{2\widehat{m}}} \sum_{i=1}^{\widehat{m}} (y_i^2 - 1) \right\}.$$

이 큰 값을 가지면 귀무가설을 기각할 것을 제안하고 극단이론 (extreme value theroy)을 사용하여 위의 검정 통계량의 점근분포를 보였다.

하지만 대부분의 신호가 앞부분에 있지 않을 경우 위의 검정통계량은 효율적이지 않을 수 있다. 이러한 경우 순서에 관계없이 일정크기 이상의 신호들만 사용하여 검정통계량을 만드는 것이 보다 합리적일 것이다. Fan (1996)은 이러한 경우에는 아래와 같은 분계점 (thresholding)을 이용한 검정통계량을 사용할 것을 권장했다,

$$\widehat{T}_H^* = \sum_{i=1}^{n} y_i^2 I(|y_i| > \widehat{\delta})$$

여기서

$$\widehat{\delta} = \sqrt{2\log(n\widehat{a}_n)}, \quad \widehat{a}_n = \min\left[4\left(\max_{1\leq i\leq n} |y_i| \right)^{-4}, \log^{-2} n \right]$$

이며 Fan은 위의 검정 통계랑의 점근분포가 정규분포로 수렴함을 보였다. Fan의 방법들은 기존의 카이제곱 방법보다 검정력이 좋으므로 카이제곱의 적용경우와 마찬가지로 신호자체의 갯수는 상대적으로 많지만 크기가 작을 경우 사용을 권장한다.

3.5. 적용사례와 R package

대역검정의 가장 대표적인 적용사례로는 유전학에서의 SNP분석, 유전자 망 (gene network)의 비교와 같은 고차원 자료의 분산분석을 들 수 있다. 또한 함수적 자료분석의 경우 함수들을 기저함수 (basis function)을 사용하여 확장할 경우 기저함수들의 계수들의 추정치를 모형 (2.1)과 같이 나타낼수 있다 (Fan, 1996). 대역검정을 이용한 함수적 분석은 다양한 분야에서 찾아 볼 수 있는데 예를 들면 천문학에서 변광성의 변광 주기의 변화탐지 (Park 등, 2011), 경영학에서 광고효과 (Fan과 Lin, 1998)등을 들 수 있다.

이 장에서 소개한 방법들은 대부분 R package로 이미 구현되어 있어서 실제 자료분석에 별도의 프로그래밍없이 적용이 가능하다. Higher criticism의 경우 R package fdrtool을 사용하면 쉽게 구현할 수 있으며 Adaptive Neyman 검정의 경우는 R package DEGraph에서 AN.test 명령어를 사용하면 된다.

4. 동시검정

이 절에서는 신호가 존재할 경우 (즉 0이 아닌 모평균들이 존재할 경우) 실제 신호가 어디에 있는지 알아보는 검정방법에 대해 알아보겠다, 전통적으로 n개의 가설검정을 동시에 시행할 경우 결과는 Table 4.1과 같이 정리할 수 있다. 여기서 n과 전체기각 횟수 R을 제외하고는 모든 변수가 관측되지 않으며 우리는 일반적으로 V와 R의 크기의 조정에 관심이 있다.

Table 4.1 Number of errors committed when testing n hypotheses

	accept H_0	reject H_0	Total
H_0 true	U	V	n_0
H_0 false	T	S	n_1
Total	$n-R$	R	n

다중검정에서 전통적인 오류 조정방법으로는 FWER을 조정하는 방법을 들 수 있겠다. 이 방법는 귀무가설이 참일때 한번이라도 귀무가설을 기각하는 확률 ($\mathbb{P}\{V \geq 1\}$)을 일정수준 이하로 되도록 전체기각 횟수 R을 조정하는 방법이다. 하지만 이러한 방법들은 전반적으로 기각하는 귀무가설의 숫자가 매우 적어서 빅데이터 분석과 같이 많게는 수백만개의 가설검정을 할 경우 적합하지 않는 경우가 종종 있다. 이러한 단점을 보완하기 위하여 Benjamin와 Hochberg (1995)는 새로운 개념의 오류인 FDR을 소개하였다.

다중검정에서 FWER과 FDR중 어떤 오류를 조정하는 것이 합리적인지 결정하기위해서 다음 2가지 경우를 고려해보자.

(a) 10개의 기각된 가설중 4개의 오발견

(b) 100개의 기각된 가설중 20개의 오발견

(a)의 경우 오발견 비율은 40% 이고 (b)의 경우는 20%이다. 하지만 오발견의 절대 갯수는 (b)의 경우가 훨씬 많다. 만약 주어진 문제에서 (a)가 더 심각한 실수라고 생각된다면 FDR을 조정하고 그렇지 않다면 FWER을 조정하는 다중검정방법을 사용할 것을 권장한다.

4.1. FWER 조정방법

전체 검정의 갯수가 상대적으로 적을 경우 FWER을 조정하는 대표적인 방법으로는 Bonferroni 방법을 들 수 있다. 3절에서 소개한 것처럼 Bonferroni 방법의 경우 개개의 가설에서 유의확률이 α/n보다 작을 경우 그 귀무가설을 기각한다. Bonferroni 방법이 FWER을 조정하는 것은 아래의 부등식를 통해 쉽게 보일 수 있다.

$$\text{FWER} = \mathbb{P}(V \geq 1) \leq \mathbb{E}(V) \leq \frac{n_0}{n}\alpha \leq \alpha$$

Bonferroni 방법은 일반적으로 너무 적은 수의 귀무가설을 기각하여서 (즉 보수적이어서) FWER을 조정하는 일련의 다른 순차적 검정방법들이 개발되었다. 이러한 방법들은 귀무가설들을 검정통계량의 크기에 따라 순차적 검정을 통하여 귀무가설들을 기각하는데 검정통계량이 커지는 순서에 따라 순차적 검정을 할 경우는 Step-up, 그 반대의 경우는 Step-down 방법이라고 부른다.

먼저 Step-down 방법중의 하나인 Holm 방법에 대해 알아보자. FWER을 α수준 이하로 조정하기 위해서 Holm 방법은 아래과 같은 알고리즘을 사용한다. 먼저 유의확률들을 순서대로 정렬하고 각각의 유의확률들의 순서통계량에 대응하는 귀무가설들을 $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(n)}$이라 하자.

1 단계: 만약 $p_{(1)} \leq \alpha/n$이면, $H_{0,(1)}$를 기각하고 2단계로 간다. 그렇지 않으면 모든 귀무가설 $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(n)}$ 을 받아들이고 여기서 멈춘다.

i 단계: 만약 $p_{(i)} \leq \alpha/(n-i+1)$이면, $H_{0,(i)}$를 기각하고 $i+1$단계로 간다. 그렇지 않으면 귀무가설 $H_{0,(i)}, H_{0,(i+1)}, \ldots, H_{0,(n)}$ 을 받아들이고 여기서 멈춘다.

n 단계: 만약 $p_{(1)} \leq \alpha$이면, $H_{0,(n)}$를 기각한다. 그렇지 않으면 귀무가설 $H_{0,(n)}$을 받아들인다.

Holm 방법을 사용시 Bonferroni 방법에 비해 일반적으로 더 많은 귀무가설을 기각하고 FWER을 같은 수준으로 조정하기 때문에 검정력은 더 높다는 것을 알 수 있다.

이와 대조되는 Step-up 방법으로는 Hochberg 방법을 들 수 있다. FWER을 α수준 이하로 조정하기 위해서 Hochberg방법은 아래와 같은 알고리즘으로 구현할 수 있으며 Step-down 방법과 비교시 가장 큰 차이점은 한번에 여러개의 가설을 기각할 수 있다는 점이다.

1 단계: 만약 $p_{(n)} \leq \alpha$이면, $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(n)}$을 모두 기각하고 여기서 멈춘다. 그렇지 않으면 $H_{0,(n)}$을 받아들이고 2단계로 간다. .

n − i + 1 단계: 만약 $p_{(i)} \leq \alpha/(n-i+1)$이면, $H_{0,(i)}, H_{0,(i+1)}, \ldots, H_{0,(n)}$을 모두 기각하고 여기서 멈춘다. 그렇지 않으면 $H_{0,(i)}$을 받아들이고 $n-i+2$단계로 간다.

n 단계: 만약 $p_{(1)} \leq \alpha/n$이면, $H_{0,(1)}$을 기각하고 그렇지 않으면 $H_{0,(1)}$ 을 받아들인다.

Holm 방법과 Hochberg 방법을 비교하자면 Holm방법은 유의확률들을 오름차순으로 나열한 후 귀무가설을 아래 조건을 만족하기 전까지 계속 기각한다.

$$p_{(i)} > \frac{\alpha}{n-i+1}.$$

반면에 Hochberg 방법은 유의확률을 내림차순으로 정렬한 후에 아래 조건을 만족하기 전까지 $H_{(i)}$를 기각하지 않는다.

$$p_{(i)} \leq \frac{\alpha}{n-i+1}.$$

위의 조건을 만족한 경우 나머지 귀무가설 $H_{0,(1)}, \ldots, H_{0,(i)}$를 모두 기각한다. Holm 방법과 Hochberg 방법 모두 같은 임계점들을 사용하지만 일반적으로 Hochberg 방법이 여러개의 귀무가설을 한꺼번에 기각할 수 있으므로 Holm 방법보다 검정력이 좋다고 할 수 있다.

4.2. FDR 조정방법

Benjamin와 Hochberg (1995)는 새로운 개념의 오류인 FDR과 이를 조정할 수 있는 방법을 제안한다. FDR를 정의하기 위해 우리는 우선 FDP (false discovery proportion)을 소개한다 (Genovese와 Wasserman, 2004). 만약 각각의 유의확률 p_i이 주어진 절사값 (cutoff) t보다 작은 경우 거기에 대응하는 개개의 귀무가설 $H_{0,i}$을 기각한다고 하면 FDP는 다음과 같이 정의할 수 있다.

$$\mathrm{FDP}(t) = \frac{V}{R} = \frac{\sum I\{p_i \leq t\}(1-H_i)}{\sum I\{p_i \leq t\}}$$

여기서 만약 i번째 귀무가설 $H_{0,i}$가 참이면 $H_i = 0$이라고 하고 그렇지 않으면 $H_i = 1$이라 하자. 또한 분모와 분자 모두 0일경우 FDP는 0으로 간주한다. 주어진 절사값 T에 대해서, FDR은 FDP의 기대값

으로 다음과 같이 정의된다.

$$\text{FDR} = \mathbb{E}\left(\text{FDP}(T)\right).$$

만약 FDR을 q이하로 조정하고 싶다면 다음과 같은 알고리즘으로 절사값 T를 계산할 수 있다 (Benjamin와 Hochberg, 1995).

1. 먼저 유의확률을 다음과 같이 순서대로 정렬한다: $p_{(1)} < \cdots < p_{(n)}$.
2. 아래와 같이 R을 정의한다.
$$R = \max\left\{i : p_{(i)} < q\frac{i}{n}\right\}$$
3. FDR 방법의 절사값인 $T = p_{(R)}$이라 정의한다.
4. 만약 $p_i \le T$이면 거기에 해당하는 귀무가설 $H_{0,i}$을 기각한다.

FDR 방법과 4.1절에서 소개한 FWER 방법의 일종인 Hochberg 방법들은 Step-up 방법으로 간주할 수 있다. 이 두방법을 비교하기위해 만약 FWER과 FDR을 같은 수준 ($q = \alpha$)에서 조정한다고 가정하자. 이 경우 각각의 방법에서 사용된 절사값들의 비는 아래와 같다.

$$\text{T}_{\text{FDR}}/\text{T}_{\text{Hochberg}} = \frac{i/n}{1/(n-i+1)} = i\left(1 - \frac{i-1}{n}\right)$$

만약 $i \approx n/2$일 경우 위의 절사값의 비는 $n/4$이므로 FDR 조정방법의 절사값이 훨씬 큼을 알 수 있다. 따라서 표본크기가 클 경우 만약 두 오류의 수준을 동일하게 준다면 FDR 조정방법이 훨씬 많은 귀무가설을 기각함을 알 수 있다.

FDR 조정 방법을 이론적 배경을 이해하기 위하여 유의확률의 분포를 다음과 같은 혼합물 분포를 가정할 수 있다 (Genovese와 Wasserman, 2004).

$$F = \pi_0 U + \pi_1 F_1$$

여기서 $\pi_0 + \pi_1 = 1$이며 π_0는 귀무가설이 참인 비율을 나타내고 U는 균등분포, 즉 귀무가설하에서 생성된 유의확률들의 누적 분포함수, F_1은 대립가설들하에서의 생성된 유의확률들의 누적 분포함수를 나타낸다.

Benjamini와 Hochberg는 유의확률들이 서로 독립이라는 가정하에서 주어진 t에 대해 아래와 같은 부등식을 성립함을 보였다. .

$$\text{FDR} = \mathbb{E}\left(\text{FDP}\right) = \mathbb{E}\left(\frac{V}{R}\right) \le \mathbb{E}\left(\frac{V}{R}|R>0\right) = \frac{\pi_0 t}{\widehat{F}(t)} \le \frac{t}{\widehat{F}(t)}$$

여기서 $\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(p_i \le t)$이다. 만약 $t = p_{(i)}$라 정의하면 $\widehat{F}(t) = \widehat{F}(p_{(i)}) = i/n$이 된다. 따라서 위식의 우변은

$$RHS = \frac{p_{(i)}}{i/n} = q \Rightarrow p_{(i)} = q\frac{i}{n}$$

이므로 우리는 Benjamini와 Hochberg의 방법이 FDR을 q이하로 조정함을 알 수 있다.

Storey (2001)는 위의 방법이 π_0를 추정하지 않고 단순히 상계값인 1을 사용한점을 주목하고 π_0를 자료로부터 추정하는 것을 제안했다. 뿐만 아니라 Storey는 FDR의 변형인 또다른 오류인 pFDR (positive FDR)을 조정할 것을 제안했다. pFDR은 다음과 같이 정의된다.

$$pFDR = \mathbb{E}\left(\frac{V}{R}|R>0\right)$$

주목할 것은 $FDR = pFDR \cdot \mathbb{P}(R>0) \leq pFDR$이라는 점이다. 하지만 동시검정을 통한 빅데이터의 분석시 대부분의 경우 $\mathbb{P}(\widehat{R>0}) = 1-(1-t)^n \approx 1$로 간주 될 수 있기때문에 실제로는 두 오류의 값은 매우 근사하다고 할 수 있으며 두방법 결과의 차이는 π_0를 어떻게 추정하는 여부에 달려있다고 할 수 있다.

pFDR은 또다른 장점은 베이지안방법을 이용하여 사후확률로 해석할 수 있다는 점이다,

$$pFDR = \frac{\pi_0 t}{F(t)} = \frac{\pi_0 \mathbb{P}(p \leq t | H=0)}{\mathbb{P}(p \leq t)} = \mathbb{P}(H=0 | p \leq t)$$

즉 pFDR은 유의확률이 주어진 절사값 t보다 작을 경우 귀무가설이 참일 경우의 사후확률로 해석할 수 있다. 그런데 많은 경우 우리는 개개의 귀무가설이 참일 사후확률에 관심이 있지만 pFDR은 개개의 귀무가설에 대한 정보는 제공하지 않는다. 이러한 문제점을 해결하기 위해 Efron (2010)은 새로운 오류개념인 지역 오발견율 (local false discovery rate)을 소개한다. 지역 오발견율을 소개하기위해서 먼저 z_i를 유의확률들의 프로빗 변환값이라 하자. 그러면 위의 혼합물 분포는 다음과 같은 Brown-Stein 모형을 이용하여 소개할 수 있다. 먼저 (δ_i, z_i) 가 독립이고 아래의 위계모형에서 생성된다고 가정하자.

$$\delta \sim g(\delta) = \pi_0 g_0(\delta) + \pi_1 g_1(\delta) \tag{4.1}$$

$$z|\delta \sim N(\delta, 1) \tag{4.2}$$

여기에서 $g_0(\delta) = \phi_{\mu,\sigma^2}(\delta)$이고 ϕ_{μ,σ^2}는 평균과 분산이 μ와 σ^2인 정규분포 확률밀도함수이며 확률 밀도함수 g_1의 값은 알지 못한다고 가정한다. 즉 귀무가설하에서 δ의 사전분포는 $N(\mu, \sigma^2)$를 따르고 대립가설하에서 g_1을 따른다고 가정한다.

만약 귀무가설하에서 검정통계량의 확률밀도함수를 $f_0(z)$, 대립가설하에서는 $f_1(z)$이라고 하면 Brown-Stein 모형하에서 확률밀도함수는 다음과 같이 정의된다.

$$f_i(z) = \int f(z|\delta) g_i(\delta) d\delta, \quad i = 1, 2.$$

따라서 Efron (2010)의 정의에 따라서 지역 오발견율 (fdr)은 다음과 같이 정의된다.

$$\text{fdr}(z) = \Pr\{H_{i,0} = \text{ 참} | Z_i = z\} = \pi_0 f_0(z)/f(z).$$

여기에서 $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$이다. 정의에서 알 수 있듯이 지역 오발견율은 주어진 유의확률의 프로빗 변환값하에서 각각의 귀무가설이 참일 사후확률로 해석할 수 있다. 또한 이 정의에 따르면 동시검정의 문제는 확률밀도함수들 f_0, f의 추정문제로 생각할 수 있다. 귀무가설하에서 유의확률들은 균등분포를 따르므로 이의 프로빗변환의 분포 (f_0)는 표준정규분포를 따른다고 가정하는 것이 일반적이다. 하지만 고차원 다중검정 문제의 경우, Efron (2010)은 f_0가 표준정규분포를 따르지 않을 수 있다는 것을 여러가지 사례연구를 통해 지적하고 f_0의 분포가 정규분포라는 것만 가정하고 평균과 분산은 각각 자료에서 추정할 것을 제안했다. Efron은 귀무가설하에서 생성된 유의확률들의 프로빗 변환은 관측된 자료의 중앙부분 (즉 유의확률의 분포에서는 오른쪽 부분)에서 주로 관측될것이라는 걸 착안하여 그 부분의 자료를 이용하여 평균과 분산을 추정하자고 제안했다. 또한 주변분포 f의 추정을 위해서는 Lindsey's 방법을 사용하고 이를 바탕으로 지역 오발견율을 추정할 것을 제안했다.

4.3. 적용사례와 R package

동시검정은 표 1.1에서 볼 수 있듯이 많은 분야에 응용되고 있다. 마이크로어레이 분석이 가장 대표적인 예이며 뇌영상자료, 천체관측자료, 신호처리 등 여러분야에서 적용되고 있다.이 장에 소개된 대부분의 방법들 (Holm, Hochberg, Bonferroni, Benjamini와 Hochberg)는 별도의 R package를 사용할 필요 없이 p.adjust 명령어로 실행할 수 있다. 각 방법들의 비교를 위해서 아래의 R 코드를 이용하여 다음과 같은 간단한 컴퓨터 모의실험을 하였다.

1. 900개의 변수를 표준정규분포에서 생성하고 100개의 변수를 $N(3,1)$에서 생성한다.
2. 유의 확률을 계산한 후에 FWER 조정방법들 (Holm, Hochberg, Bonferroni)과 FDR 조정방법 (Benjamini와 Hochberg)을 통하여 조정된 유의확률 (adjusted p-value)들을 계산한다.
3. 조정된 유의확률들이 FWER=0.05 과 FDR=0.05이하인 귀무가설들을 찾아낸다.

```
require(graphics)

set.seed(412)
 x <- rnorm(1000, mean =c(rep(0, 900), rep(3, 100))) #900개의 잡음과 100개의 신호를 생성한다
p <- 2*pnorm(sort(-abs(x))) # 유의확률의 계산

p.adjust.M <-c("holm","hochberg","bonferroni", "BH", "none")
p.adj  <- sapply(p.adjust.M, function(meth) p.adjust(p, meth))
round(p.adj,3)

pdf("mc.pdf")
matplot(p, p.adj, ylab="p.adjust(p, meth)", type ="l", lty=1:5,
    main = "P-value adjustments")
legend(0.7, 0.6, p.adjust.M, col = 1:5,lty=1:5)
dev.off()
```

Figure 4.1은 각각의 다중검정방법들에 의해 조정된 유의확률과 원래의 유의확률의 관계를 보여준다. 유의실험 결과 FWER를 조정하는 3방법 모두 19개의 같은 귀무가설을 기각했고 이 중 오발견은 없었다. 반면에 FDR를 조정하는 방법의 경우 58개의 귀무가설을 기각했으며 이 중에 오발견의 갯수는 2개였다. 만약 원래 유의확률이 0.05이하인 귀무가설들을 기각한다면 기각하는 귀무가설의 갯수는 107개이지만 이중 오발견의 갯수는 26개이다.

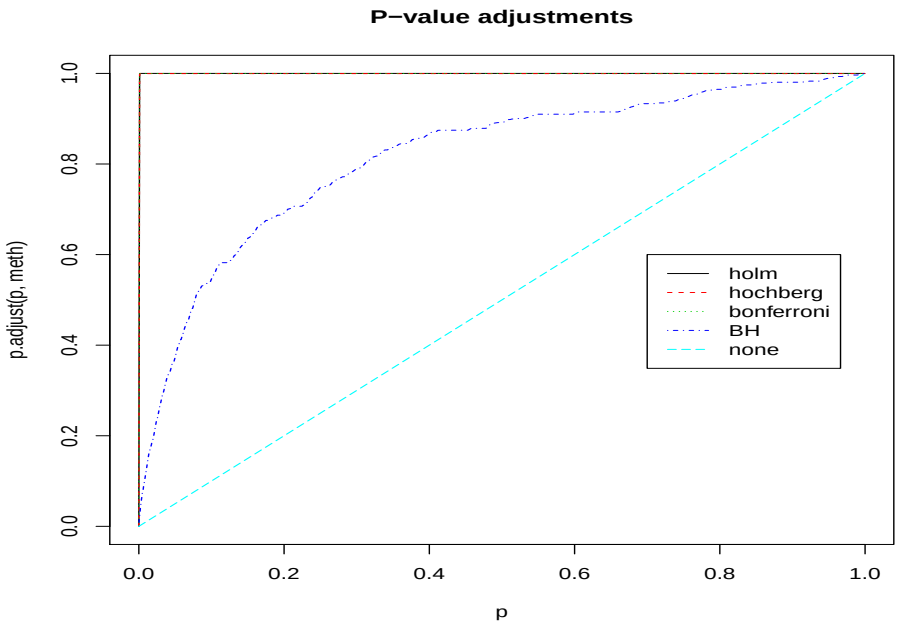

Figure 4.1 P-value adjustments with multiple comparison methods

지역 오발견율의 경우 R package locfdr 또는 fdrtool을 사용할 것을 추천한다. pFDR은 R Package samr을 이용하여 구현할 수 있다. 이 논문에서는 소개되지 않았지만 재표집 (resampling)을 이용한 FWER 조정방법의 경우 기존의 FWER 방법에 비해 검정력이 높은 것으로 알려져 있으며 이를 포함한 다양한 다중검정 방법이 R package multtest에 구현되어있다.

5. 결론

본 논문에서는 다중 검정방법의 최근 개발동향에 대한 소개를 다루었다.. 생물정보학이나 신호처리 분야를 포함한 많는 과학분야에서 다중검정은 빅데이터 분석에 핵심적인 역할을 하고 있으며 최근 이 분야에 대한 활발한 연구가 이루어지고 있다. 하지만 많은 경우 빅데이터 분석시 새로운 다중검정 방법에 대한 이해를 하지 못하고 기계적으로 적용하는 사례가 종종 있다. 본 논문에서는 현재 많이 쓰이는 다중검정방법들의 가정과 특성에 대해 알아보고 각각의 경우에 대해 최적인 다중비교방법들에 대한 소개를 하였다. 또한 동시검정에 비해서 상대적으로 주목받지 못한 대역검정에 대한 자세한 소개를 함으로서 이 분야에 보다 활발한 연구가 이루어지기를 바란다. 본 논문은 y_j가 서로 독립인 경우에 한하여 다중검정방법들을 소개하였다. 실제 많은 경우 y_j은 독립이 아닌 경우가 많으며 현재 기존의 방법들을 독립이 아닌 경우에도 적용할 수 있도록 확장하는 많은 연구가 이루어지고 있다.

References

Arias-Castro, E., Candés, E. J. and Plan, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Annals of Statistics*, **39**, 2533–2556.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. and Golani, I (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research*, **125**, 279–284.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Donoho, D. L. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures, *Annals of Statistics*, **32**, 962–994.

Efron, B. (2010). *Large-scale inferece: Empirical Bayes methods for estimation, testing, and prediction*, Cambridge University Press, Cambridge.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, **91**, 674–688.

Fan, J. and Lin, S.-K.(1998). Test of significance when data are curves. *Journal of the American Statistical Association*, **93**, 1007–1021.

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using false discovery rate. *Neuroimage*, **15**, 870–878.

Genovese, C. R. and Wasserman, L. A. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, **32**, 1035–1061.

Jin, J. (2008). Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society B*, **70**, 461–493.

Lindsay, B. G., Kettenring, J. and Siegmund, D. O. (2004). A report on the future of statistics. *Statistical Science*, **19**, 387–413.

Park, C., Ahn, J., Hendry, M. and Jang, W. (2011). Analysis of long period variable stars with nonparametric tests for trend detection. *Journal of the American Statistical Association*, **106**, 832–845.

Patti, M. E., Butte, A. J., Crunkhorn, S., Cusi, K., Berria, R., Kashyap, S., Miyazaki, Y., Kohane, I., Costello, M., Saccone, R., Landaker, E. J., Goldfine, A. B., Mun, E., DeFronzo, R., Finlayson, J., Kahn, C. R. and Mandarino, L. J. (2003). Coordinate reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. *Proceedings of the National Academy Sciences of USA*, **100**, 8466–8471.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J. and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, **100**, 9440–9445.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy Sciences of USA*, **98**, 5116–5121.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of micorarrays applied to the ionizing radiation response. *Proceedings of the National Academy Sciences of USA*, **98**, 5116–5121.

Weisberg, S. P., McCann, D., Desai, M., Rosenbaum, M., Leibel, R. L. and Ferrante, A. W. (2003). Obesity is associated with macrophage accumulation in adipose tissue. *Journal of Clinical Investigation*, **112**, 1796–1808.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, **54**, 594–604.

Wit, E. (2010). Comments on Discovering the false discovery rate by Benjamini. *Journal of the Royal Statistical Society B*, **72**, 410-411.

Journal of the Korean Data & Information Science Society
2013, **24**(5), 1063–1076

http://dx.doi.org/10.7465/jkdi.2013.24.5.1063
한국데이터정보과학회지

Multiple testing and its applications in high-dimension

Woncheol Jang[1]

[1]Department of Statistics, Seoul National University

Received 25 July 2013, revised 25 August 2013, accepted 9 September 2013

Abstract

The power of modern technology is opening a new era of big data. The size of the datasets affords us the opportunity to answer many open scientific questions but also presents some interesting challenges. High-dimensional data such as microarray are common in big data. In this paper, we give an overview of recent development of multiple testing including global and simultaneous testing and its applications to high-dimensional data.

Keywords: False discovery rate, global test, high-dimensional data, multiple test, simultaneous test.

[1] Associate professor, Department of Statistics, Seoul National University, Seoul 151-747, Republic of Korea. E-mail: wcjang@snu.ac.kr