응용통계연구 (2009)
22(5), 1059–1072

극소수 샘플에서 유의발현 유전자 탐색에 사용되는 순열에 근거한 검정법

이주형[1] · 송혜향[2]

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

(2009년 5월 접수, 2009년 9월 채택)

요 약

마이크로어레이 극소수 샘플(array) 자료의 분석에서는 유의한 발현수치를 나타내는 유전자를 검정통계량에 의해 결정하는 것이 주요과제이다. 이 때 수 천 또는 수 만개인 유전자의 발현수치로부터 귀무분포(null distribution)의 생성이 필수적이며, 극소수 샘플 자료의 경우에는 순열방법(permutation methods)에 의해 귀무분포를 생성하는 것이 가장 바람직하다. 본 논문에서는 귀무분포 생성에 사용될 수 있는 매우 단순한 검정통계량을 제시하면서 더불어 귀무분포 생성에 적절한 순열방법도 제안한다. 모의실험으로 기존의 검정통계량으로 생성된 귀무분포와 본 논문에서 제안하는 검정통계량의 귀무분포를 비교하며, 실제 자료에 적용하여 유의 유전자를 탐색한다.

주요용어: 순열검정법, 마이크로어레이 자료, 유의 발현유전자, 귀무분포.

1. 서론

마이크로어레이 실험에서 충분한 세포조직이 확보되지 않아서 샘플(array)수가 10개 미만인 경우가 빈번한데 분석의 대상인 유전자는 수 천 또는 수 만개로써, 발현수치 평균이 처리(condition)로 인해 달라지는 또는 종양의 타입에 따라 달라지는 유전자를 결정하는 일이 분석의 주요과제이다. 일부 유전자는 처리에 의해 발현수치가 달라지지 않는 반면에 일부 유전자는 발현수치가 크게 달라지므로 수 천 또는 수 만개 유전자의 발현수치 분포의 양상을 파악하기가 매우 어려운 상황에서 유의 유전자를 밝히기 위해 p값을 정하게 되는데, 이러한 결정에는 귀무가설 하에서의 검정통계량의 분포, 즉 귀무분포(null distribution)를 파악하거나 또는 생성하는 일이 우선되어야 한다.

순열검정법(permutation test)은 Fisher (1935, 1966)에 의해 처음 제안된 후 Pitman (1937)에 의해 구체적으로 여러 검정의 경우로 발전되어 두 군 비교에 대한 순열검정도 이에 포함되어 있다. 오차항의 정규분포 가정이 성립된다면 순열검정법은 모수적 검정법과 비슷한 검정력을 나타내지만 (Hoeffding, 1952; Lehmann과 Stein, 1949), 순열검정법은 실제로 어떤 분포 가정에도 의존하지 않는다. 단지 오차항이 랜덤이고 오차항의 교환성(exchangeability)만이 요구된다 (Welch, 1990). 순열검정법은 그 유용성에도 불구하고 최근까지 계산상의 어려움으로 널리 응용되지 못하였으나 모수적 분포 가정이 요구되지 않는다는 점, 즉 귀무가설 하에서 검정통계량 분포를 구체적으로 가정하지 않아도 된다는 특징으로 인해 마이크로어레이 자료의 수 천 또는 수 만개의 다양한 양상의 발현수치의 분석에 적절한 검정법 중 하나이며, 특히 귀무분포의 생성에 순열방법은 중요한 역할을 한다. 이에 Pan (2003a), Zhao와

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

Pan (2003), Xie 등 (2005), Gao (2006) 등이 기존의 순열검정법의 문제점을 지적하거나 또는 단점을 계속 보완시킨 순열검정법을 제안하였다. 이 논문의 저자들 역시 극소수 샘플자료를 다루고 있으며, 특히 Pan 등 (2003b)에서는 극소수 샘플자료의 우도함수비 검정법(likelihood ratio test, LRT)과 혼합모형 방법(mixture model method, MMM)을 제시하고 있으며, 이 MMM방법은 본 논문에서 설명하는 대립가설 하에서 발현과 비발현 수치의 두 분포가 섞여 있는 경우를 뜻한다. Pan (2003a)은 모의실험에서 총 8개의 마이크로어레이 샘플을 다루고 있으며, 이는 각 처리군의 샘플을 다시 이분하여 이에 근거한 검정통계량인 2.2.2절의 수정 t 검정법의 입장에서 보면 최소의 샘플수이기도 하다.

마이크로어레이 자료의 분석에 주로 사용되는 t 통계량의 문제점은 소수 샘플자료의 경우 각 유전자에서 구한 분산 추정값이 매우 불안정하다는 점이다 (Tusher 등, 2001; Smyth 등, 2003). 그러나 발현수치가 비슷한 유전자는 동일 처리 하에서 분산이 비슷한 크기를 가진다는 사실에 근거하여 (Lee, 2001), 발현수치가 비슷한 다수 유전자 자료의 종합에서 분산을 추정하여 분석하는 방법도 제안되었다 (Jain 등, 2003). 이 저자들의 방법과는 달리 본 논문에서는 평균차 통계량에 근거한 순열검정법을 새로이 제안하며, 특히 유의성검정에서 기본이 되는 귀무분포를 이 순열검정법으로 옳게 생성할 수 있으며 이 귀무분포에 근거하여 유의한 발현수치를 나타내는 유전자를 쉽게 결정하게 된다.

순열검정법이 장차 이용될 수 있는 기반은 앞서 제시된 순열검정법의 적절한 평가 작업에 달려 있다하겠다. 순열검정법은 구체적으로 선정된 검정통계량과 또한 순열의 실시방법에 따라 그 결과가 결정된다. 그러므로 본 논문에서는 이 두 가지 측면에서 총 8개의 마이크로어레이 극소수 샘플로서 순열검정법의 적절성 또는 문제점을, 유의성 검정에서 기본이 되는 귀무분포의 생성에 초점을 두어 알아본다.

Spino와 Pagano (1991)는 쌍체계획법(matched-pairs design)에서 윈저(Winsorized) 평균이나 절사(trimmed) 평균 또는 중위수를 사용한 순열검정법이 이상점이 존재하거나 정규성이 성립치 않는 소량자료의 경우라도 효율성이 매우 높다고 언급하였다. 또한 과거문헌들 (Albers 등, 1976; Chernoff와 Savage, 1958; Hoeffding, 1952; Lehmann과 Stein, 1949)을 뒷받침으로 순열로 생성한 분포는 근사적으로 정규분포한다고 언급하고 있으며 더불어 절사평균의 근사적 정규성도 언급하고 있다. 본 논문의 연구는 Spino와 Pagano (1991)의 연구와 독립된 두 군 계획이라는 점이 다를 뿐이므로 동일하게 효율성이 높을 것이 예상되며, 윈저평균의 근사적 정규성과 순열로 생성된 모든 유전자를 종합한 귀무분포의 정규성을 이용하여 여러 순열검정법을 평가한다.

2. 방법

두 군 비교에 사용되는 t 검정에서 p값 유의성 결정은 관측된 자료로부터 구한 검정통계량 Z의 값을 귀무분포와 견주어 결정한다. 마이크로어레이 관련 논문에서는 t 검정통계량을 기호 t 대신에 Z로 표기하는 것이 통례이므로 여기서도 Z로 표기한다 (Zhao와 Pan, 2003; Efron 등, 2001). 마이크로어레이 자료 분석에서도 처리 또는 종양의 타입이 다름에 의해 유의한 발현을 나타내는 유전자 탐색에 중요한 역할을 하는 것이 귀무분포의 생성이다. 이제 귀무분포를 생성하는 귀무통계량(null statistic)을 z로 나타낼 때 z 분포, 즉 귀무분포의 한 가지 생성방법은 두 군에 속한 샘플 자료를 여러 번 랜덤 순열하여 모든 유전자에 대해서 계산된 Z값들의 분포로서 구성하는 것이며, 다수 유전자수를 이용하여 모든 유전자에 대한 Z값으로 구성되는 귀무분포는 양극단 부분을 사용해서 검정해야 하는 경우에 매우 안정적이다. 또한 두 군 비교의 t 검정에서는 귀무통계량 z가 검정통계량 Z와 동일하지만, 경우에 따라서는 다를 수 있음을 2.2.2절에서 보게 된다.

만약 일부 유전자가 발현을 나타내는 경우에는 두 군의 샘플자료를 랜덤 순열로써 구한 Z의 귀무분포는 실제로 모든 유전자가 발현을 나타내 보이지 않는 경우보다도 큰 변동을 보이게 되어 Z의 귀무분포를

생성하였다고 간주할 수가 없다. 이러한 현실에서 적절한 귀무통계량의 제시 또는 귀무분포의 생성이 유의한 유전자를 밝히는데 중요하다. 우선 순열방법을 설명하며, 다음으로 마이크로어레이 극소수 샘플의 경우에 순열검정법에서 사용될 수 있는 검정통계량을 설명한다.

2.1. 순열방법

전체 샘플자료에 적용하는 순열방법으로는 전체 샘플의 모든 가능한 순열의 경우를 단지 1회씩만 고려하는 정확한(exact) 순열방법과 랜덤하게 임의의 회수의 순열을 선정하는 랜덤 순열방법이 있다. 일반적으로 샘플수가 커서 전체 순열의 경우수가 클 때에는 랜덤 순열방법을 선택하게 된다. 이와 달리 전체 샘플을 이분할한 상태에서 순열방법을 적용할 수도 있으며, Pan (2003a)이 제안한 새로운 검정법에서 이와 같은 순열방법을 설명하게 된다.

정확한 순열방법 또는 랜덤 순열방법에서 순열의 반복수를 B라 하자. 마이크로어레이 전체 샘플자료가 X행렬로 표현되어서 G개의 유전자 행(row)과 총 샘플인 $n = n_1 + n_2$ 열(column)로 구성되었다. 여기서 n_1은 대조군의 샘플수, n_2는 처리군의 샘플수이며 본 논문에서는 두 군에 동일 샘플수가 할당된 균형자료를 다루게 된다. 검정통계량 Z가 선정되어 관측된 샘플자료로부터 모든 유전자에 대해서 검정통계량 $Z_1, \ldots, Z_G$의 값이 계산되었을 때 귀무가설 하에서 이 통계량의 결합분포 또는 주변분포를 알 수 없는 상황에서 모수적 가정을 하지 않는다면 순열검정법이 바람직하다. 검정통계량 $Z_1, \ldots, Z_G$의 귀무가설 하에서의 결합분포는 샘플자료 행렬 X의 전체 열을 랜덤하게 순열하여 처음 n_1개 샘플을 대조군으로 다음 n_2개의 샘플을 처리군으로 분류할 때 이와 같은 순열방법은 여러 유전자간의 연관성을 그대로 유지하면서도 처리 또는 대조군 소속 여부와 유전자 발현수치가 무관하도록 진행하는 방법이다.

본 논문에서는 모든 가능한 순열의 경우수를 모두 고려하는 정확한 순열방법을 가지고 경우에 따라서는 문제가 될 수 있는 랜덤 순열방법을 평가한다. 랜덤 순열방법에서는 관측된 샘플자료를 포함하여 전체 순열 경우수보다 작은 회수의 샘플자료의 랜덤순열을 고려하며, 예를 들어 총 샘플수가 8인 경우에 Pan (2003a)은 랜덤순열의 반복수 B를 50회로 정하여 랜덤순열을 추출한다. 이 샘플자료의 경우에 실제로 전체 순열의 경우수는 총 $\binom{8}{4} = 70$가지로 정확한 순열방법에서는 이 모든 순열의 경우를 고려하여 반복수 B는 70이다. 그러나 랜덤순열의 경우에는 반복수가 적을 때에는 중복되는 순열의 경우가 나타날 가능성도 생기고 또한 어떤 순열은 추출되지 못하여, 구해진 귀무분포가 실제 비대칭이 될 수 있으며 따라서 정확한 귀무분포를 반영하지 않을 수 있다. 아래 2.2.2절에서 소개하는 Pan (2003a)이 제시하는 수정 t 검정법의 경우에는 각 군에서, 즉 이분할한 상태에 적용한 순열방법으로 총 순열의 경우수가 $\binom{4}{2}^2 = 36$가지이므로 랜덤하게 50회를 반복한다면 어떤 순열의 경우는 2회 이상 반복 추출되거나 또는 한 번도 선택되지 않는 순열의 경우도 있어 구해진 분포가 비대칭이 되기 쉽다.

본 논문에서는 마이크로어레이 극소수 샘플자료의 분석에 초점을 두었다. 이와 같이 샘플수가 적은 경우에는 랜덤 순열방법을 이용하는 것보다 정확한 순열방법으로 귀무분포를 추정하는 것이 어렵지 않으며 귀무분포를 더욱 정확히 반영하는 방법이 된다. 적절한 검정통계량과 귀무통계량이 선정되었다면 정확한 순열검정법의 경우에는 귀무분포가 대칭을 이루고 귀무가설을 충실히 반영하는 귀무분포를 생성할 수 있다. 순열방법의 자세한 과정은 모의실험에서 설명한다.

2.2. 검정통계량

일반적으로 많이 사용하는 t 검정통계량을 우선 설명하며, 다음으로 Pan (2003a)에 의해 제시된 수정된 t 검정법을 설명한다.

2.2.1. t **검정법** 우선 한 유전자에 대한 검정을 고려한다. 두 군에서의 유전자 발현수치는 $(X_1, \ldots, X_{n_1})$과 $(Y_1, \ldots, Y_{n_2})$이고, 각 군의 평균은 μ_1과 μ_2라고 하자. 여기서 n_1과 n_2는 각 군의 샘플수이다. 검정의 귀무가설은 $H_0 : \mu_1 = \mu_2$이다. 두 군의 발현수치 자료의 등분산을 가정하는 경우에는 t 검정통계량은

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2}} \tag{2.1}$$

이다. 여기서 $\bar{X} = \sum_{i=1}^{n_1} X_i/n_1$와 $\bar{Y} = \sum_{i=1}^{n_2} Y_i/n_2$은 표본평균이고, s^2은 등분산의 합병추정량으로서 $s^2 = \{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2\}/(n_1 + n_2 - 2)$이다. 두 군의 발현수치 자료의 등분산을 가정할 수 없는 경우에는 Welch t 검정통계량을 사용한다.

이 Z 통계량에 근거한 귀무가설 하에서의 순열방법에 의해 생성된 검정통계량값 z는 총 n개의 수치 $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ 중 랜덤하게 처음 n_1개를 뽑아 첫 번째 군의 발현수치로 정하고 나머지 n_2개를 두 번째 군의 발현수치로 정한 후 모든 유전자에 대해 이와 같은 순열방법으로 위의 식 (2.1)에 의해 구한 z 검정통계량의 값으로 귀무분포가 구해진다. 이 t 검정법을 Pan (2003a)은 기존(current) 검정법이라고 부르고 있다. Pan (2003a)은 전체 샘플자료에 랜덤 순열방법을 채택하였고 랜덤순열의 반복수 B를 50으로 정하였다.

Efron 등 (2001)과 Tusher 등 (2001)의 논문에서는 식 (2.1)의 검정통계량 Z 또는 귀무통계량 z의 분모가 0에 가까운 값이 되어 Z 또는 z의 값이 매우 큰 값이 되는 것을 피하기 위해서 분모에 작은 상수값인 s_0를 더하는 것을 제안하였다. 이와 같은 수정은 합리적인 해결방법인 것 같지만 상수값 s_0의 구체적인 결정이 어렵고, 발현수치 자료가 정규분포한다 하여도 분포에 상수값 s_0를 더한 검정통계량은 더 이상 t 분포를 따르지 않는다. 그러므로 본 논문에서는 일반적으로 사용하는 t 검정통계량을 그대로 사용한다.

2.2.2. 수정 t **검정법** 방금 설명한 t 검정법에서 검정통계량의 분자는 귀무가설하에서 평균이 0이다. 귀무가설 하에서든 대립가설 하에서든 검정통계량의 분자가 항상 0이 되는 다음의 귀무통계량 z_{Pan}을 Pan (2003a)은 제안하였다. 다시 말하면 대조군과 처리군의 각 군을 같은 크기의 표본수로 이분할하여 귀무통계량 z_{Pan}을 생성한다. 이는 처리군의 발현수치의 분산이 대조군의 비발현수치의 분산과 다를 수 있음을 감안하여 독립적으로 대조와 처리의 각 군에서 랜덤 순열방법을 적용하여 귀무분포를 생성하는 것이다. 순열방법으로는 랜덤순열 또는 정확한 순열방법을 채택할 수 있다. 검정통계량 Z_{Pan}은 귀무통계량 z_{Pan}과 유사한 형태를 지니며 분자는 2.2.1절의 Z의 분자와 동일하다. 구체적으로 검정통계량과 귀무통계량은 다음과 같다.

$$Z_{Pan} = \frac{\left(\sum_{i=1}^{n_{11}} \frac{X_i}{n_{11}} + \sum_{i=n_{11}+1}^{n_1} \frac{X_i}{n_{12}}\right)/2 - \left(\sum_{i=1}^{n_{21}} \frac{Y_i}{n_{21}} + \sum_{i=n_{21}+1}^{n_2} \frac{Y_i}{n_{22}}\right)/2}{\sqrt{\left(\frac{s_{11}^2}{n_{11}} + \frac{s_{12}^2}{n_{12}}\right)/4 + \left(\frac{s_{21}^2}{n_{21}} + \frac{s_{22}^2}{n_{22}}\right)/4}}, \tag{2.2}$$

$$z_{Pan} = \frac{\left(\sum_{i=1}^{n_{11}} \frac{X_i}{n_{11}} - \sum_{i=n_{11}+1}^{n_1} \frac{X_i}{n_{12}}\right)/2 + \left(\sum_{i=1}^{n_{21}} \frac{Y_i}{n_{21}} - \sum_{i=n_{21}+1}^{n_2} \frac{Y_i}{n_{22}}\right)/2}{\sqrt{\left(\frac{s_{11}^2}{n_{11}} + \frac{s_{12}^2}{n_{12}}\right)/4 + \left(\frac{s_{21}^2}{n_{21}} + \frac{s_{22}^2}{n_{22}}\right)/4}}, \tag{2.3}$$

여기서 s_{11}^2, s_{12}^2, s_{21}^2 그리고 s_{22}^2는 $(X_1, \ldots, X_{n_{11}})$, $(X_{n_{11}+1}, \ldots, X_{n_1})$, $(Y_1, \ldots, Y_n)$ 그리고 $(Y_{n_{21}+1}, \ldots, Y_{n_2})$의 표본분산이다. 예를 들어서 $s_{11}^2 = \sum_{i=1}^{n_{11}} (X_i - \sum_{j=1}^{n_{11}} X_j/n_{11})^2/(n_{11} - 1)$이며, 다른 표본

분산도 같은 방법으로 구한다. 대조군과 처리군의 발현수치 X_i와 Y_i의 정규분포 가정하에서 Z_{Pan}과 z_{Pan}의 분모는 분자와 독립이다. 그러므로 z_{Pan}으로 생성된 귀무분포는 Z_{Pan}의 귀무분포로서 매우 적절한 것으로 보인다. 이 검정법을 Pan (2003a)은 새로운 방법이라고 부르며, 랜덤순열을 채택하여 50회 반복을 시행하였다.

대조군과 처리군의 샘플수가 각각 4일 때 t 검정법과 수정 t검정법의 Z와 Z_{Pan}의 참귀무분포는 각각 자유도 6과 4의 t 분포에 따른다 (Pan, 2003a).

2.2.3. 평균차 또는 윈저평균차 통계량 일반적인 t 검정법과 Pan (2003a)의 t 검정법에서 사용된 검정통계량은 t 통계량 또는 수정 t 통계량으로서 분자는 평균차로 구성되었고 분모는 두 군 또는 이분할한 부분군의 자료로부터 구한 표준편차이다. 그러나 극소수 샘플에서는 각 군에서 4개의 샘플수에 근거하여, 또는 이분할한 자료에서는 2개의 샘플수에 근거하여 표준편차를 계산하기 때문에 이 표준편차 추정량이 매우 불안정할 수 밖에 없다. 따라서 구해진 검정통계량의 값도 매우 불안정한 값을 나타낸다. 그러므로 평균차 $\overline{D} = \overline{X} - \overline{Y}$를 검정통계량으로 사용하는 방법을 본 논문에서는 제안한다. 여기서 $\overline{X}$와 $\overline{Y}$는 식 (2.1)에서 정의한 바와 같이 각 유전자마다 계산하게 되는 각 군의 샘플의 평균이며 따라서 $\overline{D}$는 이 두 군의 평균차이다. 모든 유전자를 총망라하여 귀무분포를 생성하는 경우 더욱 적절한 귀무분포를 생성하기 위해 단순 평균이 아닌 여러 다른 평균을 채택할 수 있으며 윈저평균 (또는 절사평균)에 근거한 샘플의 평균차 $\overline{D}_W$ 검정통계량이 적절하다.

윈저평균은 예를 들어 1,000개의 유전자가 있고 전체 샘플수가 8일 때 총 8,000개의 자료를 순위로 나열하여 총 자료 중 5%인 400개에 해당하는 자료를, 또는 정해진 다른 비율에 해당하는 자료를 교체해 주는 것으로서 양극단의 자료에서 최소 발현수치부터 200번째까지의 작은 발현수치들을 201번째 발현수치로 모두 교체하고 또한 7801번째부터 최고 발현수치까지의 큰 발현수치들을 7800번째 발현수치로 모두 교체한 후 평균을 구한다.

이러한 평균차 검정통계량에 정확한 순열방법이나 랜덤 순열방법을 적용하여 귀무분포를 생성한다. 그러나 어느 정도까지의 자료를 윈저 변환시켜야 하는지의 결정이 어려우며, 그러한 이유로 만약 단순 평균차가 적절하다고 판단된다면 더욱 추천되는 검정통계량이다. 단순 평균차 또는 윈저평균차는 정규분포에 매우 가까운 분포임을 가정할 수 있고 이러한 근사적 정규분포성을 다음 절에서 제시하는 QQ plot으로 확인할 수 있다.

2.3. 유의 발현유전자의 결정방법

본 논문에서 초점을 두었던 적절한 귀무분포의 생성은 유의 발현유전자의 결정에 사용되는 경계값(threshold value) A를 구하기 위해서이다. 모의실험으로 귀무분포의 생성에 적절한 검정통계량과 순열방법이 결정된다면 이를 이용하여 귀무분포로부터 미리 정해진 유의수준에 해당하는 경계값 A를 우선 구한다. 다음으로 양측 검정의 경우에, 실제 관측된 자료로부터 구한 검정통계량의 절대값이 이 경계선보다 큰 유전자를 유의 발현유전자로 결정한다. 구체적으로 추천되는 검정통계량과 순열방법은 모의실험의 결과를 설명한 후 밝히게 된다.

3. 모의실험

모의실험에서 이미 알고 있는 참귀무분포(true null distribution)를 이용하여 각 순열방법과 각 검정통계량에 의한 귀무분포 생성의 적절함을 평가한다. 즉 참귀무분포는 비발현 유전자 자료만을 가지고 정확한 순열방법으로 생성된 분포를 지칭한다. t 검정법과 수정 t 검정법의 경우 QQ plot와 히스토그램에

실선으로 제시된 참귀무분포의 기준선은 Pan (2003a)에서와 같이 t 분포로 정하였고 총 샘플수가 8인 경우 자유도는 각각 6과 4이다. 평균차를 사용한 순열검정법의 경우에는 순열로 생성된 귀무분포의 근사적인 정규성에 의존하여 참귀무분포 기준선을 여러 문헌에서 뒷받침된 정규분포로 정하였다 (Spino와 Pagano, 1991; Albers 등, 1976; Chernoff와 Savage, 1958; Hoeffding, 1952; Lehmann와 Stein, 1949). 여러 검정법과 순열방법으로 생성된 귀무분포가 기준선과 일치하는 경우 생성된 귀무분포는 참귀무분포에 근사한다고 말할 수 있다.

본 논문의 순열방법에 의한 귀무분포 생성과 이를 이용한 검정의 과정은 다음과 같다.

1. 귀무통계량을 선택한다.
2. 순열의 b번째 반복에서 그 순열의 순서대로 총 샘플(열)을 재배열하여 두 군으로 구분한다. 배열된 총 샘플(열)에서 처음 n_1개는 첫째 군이고 나머지는 둘째 군이다.
3. 이와 같이 배열된 두 군에 근거하여 각 유전자(행)의 발현수치로 귀무통계량을 계산하여 z_j^b $(j = 1, \ldots, 1000)$를 얻는다.
4. 반복수 B만큼 2번과 3번 과정을 반복한다.
5. 총 유전자수×반복수만큼 계산된 귀무통계량 z_j^b $(b = 1, \ldots, B; j = 1, \ldots, 1000)$으로부터 귀무분포를 생성한다.
6. 위의 방법으로 생성된 귀무분포 $f_0(z)$에서 미리 정해진 유의수준에 해당하는 경계값 A를 아래와 같이 구한다.

$$\alpha = \int_{|z|>A} f_0(z)dz, \tag{3.1}$$

   여기서 z_j^b를 단순히 z로 표기하였다.
7. 각 유전자의 검정통계량 Z를 계산하여 경계값 A와 비교하여 유의한 유전자를 탐색한다. 즉 $|Z| > A$인 경우에 유의하다. 유의성 검정에서 경계값 A가 모든 유전자에 대해 동일한 값으로 사용된다.

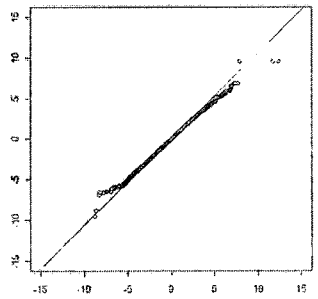
각 순열방법과 각 검정통계량에 의한 귀무분포 생성의 적절함의 평가에는 Pan (2003a)에서 제시된 FP(estimated numbers of false positives)와 TP(estimated total positives)를 추정하여 알아본다. 이는 총 유전자, 즉 1,000개 중 귀무분포로 정한 A를 사용하여 FP는 귀무통계량의 절대값이 A보다 큰 경우수이며, 그러나 반복수 B로 평균한 경우수가 되며, TP는 검정통계량의 절대값이 A보다 큰 경우수이다. 모의실험에서는 실제로 어떤 유전자가 비발현 유전자인지 알고 있으므로 비발현 유전자 자료만에 근거하여 생성된 귀무분포로부터 FP를 결정한다. 이 FP와 TP는 다음과 같이 표현된다.

$$\text{FP} = \frac{1}{B}\sum_{b=1}^{B} \#\{z : |z^b| > A\}, \tag{3.2}$$

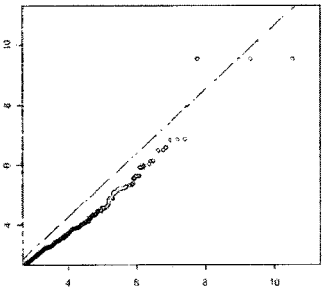
$$\text{TP} = \#\{Z : |Z| > A\}. \tag{3.3}$$

3.1. 모의실험

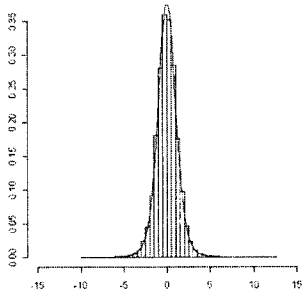
첫 번째 모의실험 샘플 자료는 전체 유전자의 수를 $G = 1000$개로 정하였을 때 유전자 발현수치 평균이 두 군의 처리에서 다른 유전자를 50%로 정하였다. 처음 500개의 유전자는 유전자 발현수치는 대조군과 처리군이 서로 동일 분포하며, 다음 500개의 유전자는 유전자 발현수치는 대조군과 처리군이 서로 다른 분포를 따른다고 가정한다. 즉, $g = 1, \ldots, 500$일 때 $X_{gi}, Y_{gi} \sim N(0, 1)$이고 $g = 501, \ldots, 750$일 때



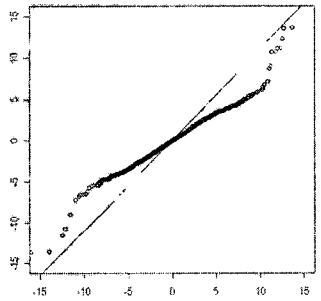
(a) Random permutation with t statistic



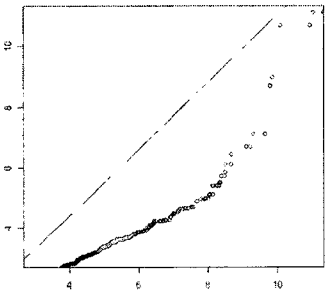
(b) Enlargement of upper part



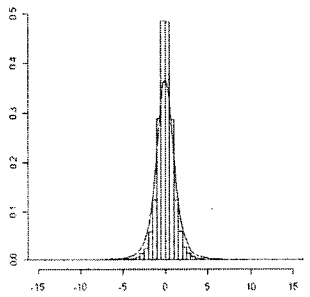
(c) Random permutation with t statistic



(d) Each group random permutation with t_m statistic



(e) Enlargement of upper part



(f) Each group random permutation with t_m statistic

그림 **3.1.** QQ plot & Histogram: empirical vs true null distribution(solid line)

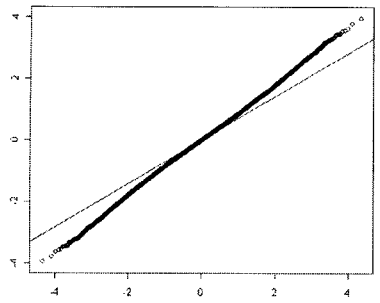
$X_{gi} \sim N(0,1)$, $Y_{gi} \sim N(2,1)$이며 $g = 751, \ldots, 1000$일 때 $X_{gi} \sim N(0,1)$, $Y_{gi} \sim N(-2,1)$이다. 각 군의 샘플수는 4로 총 샘플수는 8이며, 각 군의 샘플수는 $n_1 = 4$, $n_2 = 4$로써 $i = 1, \ldots, 4$이다.

랜덤 순열방법에서 반복수 B는 50이고 정확한 순열방법에서 반복수 B는 모든 순열의 경우수인 70이다. 순열방법에서는 정확한 순열방법, 랜덤 순열방법 그리고 이분할 랜덤 순열방법을 비교하게 된다. 검정통계량은 t 검정통계량, 수정 t 검정통계량, 평균차 또는 윈저평균차 (또는 절사평균차) 검정통계량을 위의 여러 순열방법과 병합하여 비교한다.

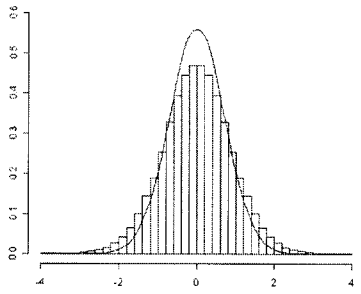
두 번째 모의실험 샘플 자료는 양의(positive) 발현 유전자 경우이다. 전체 유전자의 수를 $G = 1000$개로 정하였을 때 유전자 발현수치 평균이 두 군의 처리에서 다른 유전자를 10%로 정하여, 처음 900개 유전자의 발현 수치는 대조군과 처리군이 서로 동일한 분포이며, 나머지 100개 유전자의 발현수치는 대조군과 처리군이 서로 다른 분포를 따른다고 가정한다. 즉, $g = 1, \ldots, 900$일 때 즉 $X_{gi}, Y_{gi} \sim N(0,1)$이고 $g = 901, \ldots, 1000$일 때 $X_{gi} \sim N(0,1)$, $Y_{gi} \sim N(4,1)$이다. 각 군의 샘플수는 4로 총 샘플수는 8이며, 각 군의 샘플수는 $n_1 = 4$, $n_2 = 4$로써 $i = 1, \ldots, 4$이다.

3.2. 모의실험 결과

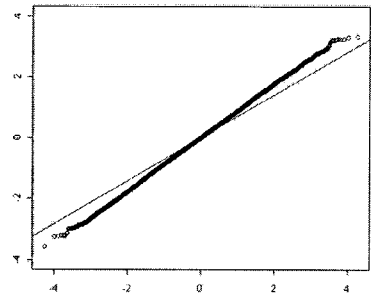
모의실험의 결과는 QQ plot과 히스토그램을 이용하여 위에서 제시한 참귀무분포 기준선과 비교한다. 그림 3.1의 t 검정통계량의 경우에는 기준선은 (a)는 자유도 6의 t 분포이고 (c)는 자유도 4의 t 분포이다. 그림 3.2~3.4의 평균차를 사용한 순열검정법의 경우에는 참귀무분포의 기준선은 추정된 평균과 분



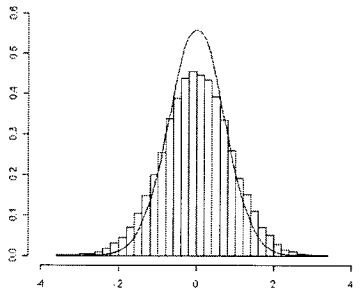
(a) Exact permutation with $\overline{D}$



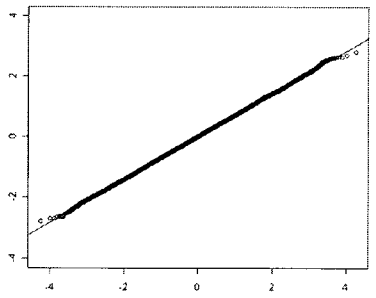
(b) Exact permutation with $\overline{D}$



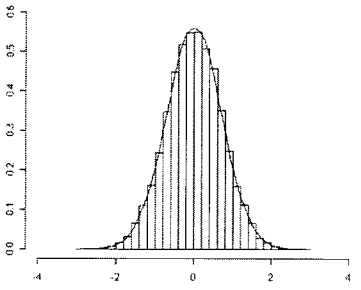
(c) Random permutation with $\overline{D}$



(d) Random permutation with $\overline{D}$



(e) Each group random permutation with $\overline{D}_m$



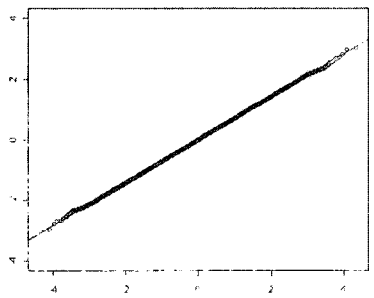
(f) Each group random permutation with $\overline{D}_m$

그림 3.2. QQ plot & Histogram: empirical vs true null distribution(solid line)

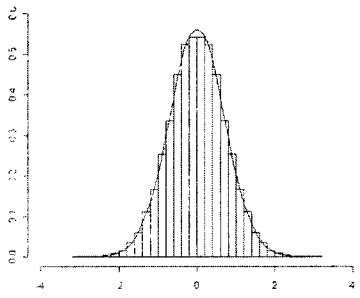
산을 사용한 정규분포이다. 그림에 사용된 자료수는 유전자수(1,000)×순열의 반복수인 B이다. 구체적으로 정확한 순열방법의 경우에는 B가 70이므로 각 그림마다 총 자료수는 70,000개, 랜덤 순열방법의 경우에는 B를 50으로 정하여 각 그림마다 총 자료수는 50,000개이다.

첫 번째 모의실험 샘플 자료의 결과를 QQ plot과 히스토그램으로 살펴보자.

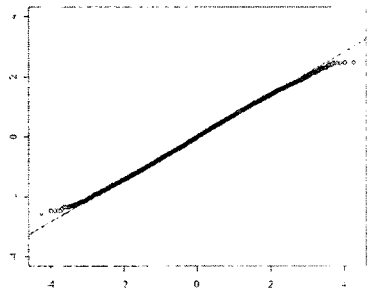
그림 3.1에서 (a)는 랜덤 순열방법을 적용한 t 검정통계량으로 구한 귀무분포의 QQ plot이고, (d)는 이분할 랜덤 순열방법을 적용한 수정 t 검정통계량으로 구한 귀무분포의 QQ plot이다. 여기서 (a), (b)와 (c)의 기준선은 참귀무분포로서 자유도가 6인 t 분포이고, 수정 t 검정통계량을 적용한 (d), (e)와 (f)의 기준선은 참귀무분포로서 자유도가 4인 t 분포이다 (Pan, 2003a). 수정 t 검정통계량의 경우 검정통계



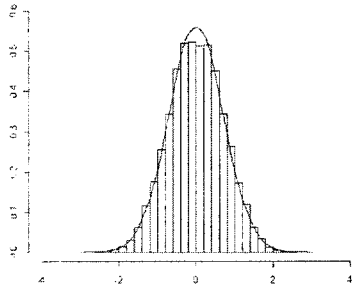
(a) Exact permutation with $\overline{D}_W$



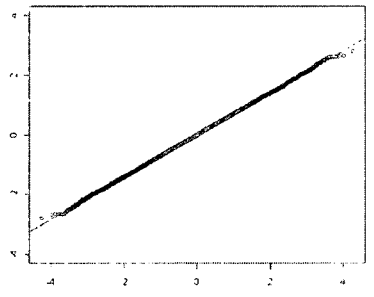
(b) Exact permutation with $\overline{D}_W$



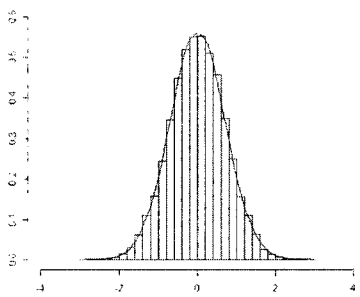
(c) Random permutation with $\overline{D}_W$



(d) Random permutation with $\overline{D}_W$



(e) Each group random permutation with $\overline{D}_{W,m}$



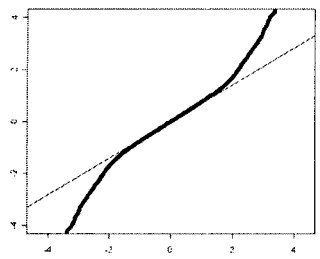
(f) Each group random permutation with $\overline{D}_{W,m}$

그림 **3.3.** QQ plot & Histogram: empirical vs true null distribution(solid line) with winsorized data

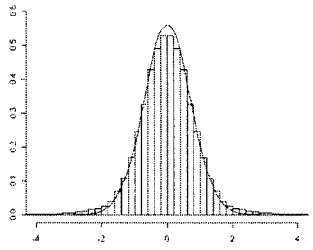
량 값의 범위가 $-15 \sim 15$로 매우 변동이 크며 기준선과 비교해 볼 때 옳은 귀무분포 생성이라고 볼 수 없다. t 통계량은 수정 t 통계량에 비해 기준선에 좀 더 가깝지만 두 통계량 모두 참귀무분포의 기준선과 일치한다고 볼 수 없다. 그러므로 극소수 샘플 자료에서 t 검정통계량은 분산에 의해 큰 영향을 받기 때문에 좋은 통계량이 될 수 없다.

다음은 본 논문에서 제시하는 평균차 검정통계량의 결과를 살펴본다. 순열방법에 따른 평균차 검정통계량의 QQ plot과 히스토그램은 다음과 같다.

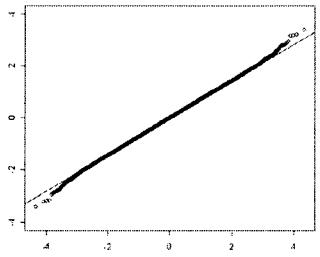
그림 3.2의 (a), (c)와 (e)에서 순열방법의 차이에 대한 결과를 비교할 수 있다. (a)는 정확한 순열방법, (c)는 랜덤 순열방법, 그리고 (e)는 이분할 랜덤 순열방법을 적용한 경우이다. 세 순열방법에서 QQ



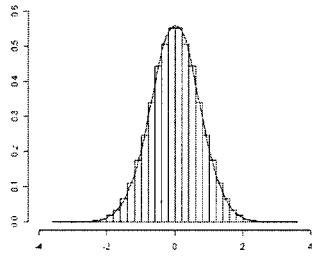
(a) Observed data



(b) Observed data



(c) Winsorized data



(d) Winsorized data

그림 **3.4.** QQ plot & Histogram of Exact permutation with $\overline{D}$: empirical vs true null distribution(solid line)

plot을 살펴보면 t 검정통계량의 경우보다도 안정적이다. 특히 극소수 샘플 자료에서는 단순 평균차 통계량을 사용하는 경우에 이분할 순열방법은 정확한 참귀무분포를 생성할 수 있다. 그러나 평균차 통계량은 정확한 순열방법으로는 참귀무분포에 근접하다고 볼 수 없다.

본 논문에서 제시한 윈저평균차 검정통계량을 적용한 QQ plot과 히스토그램은 다음과 같다.

그림 3.3의 (a), (c)와 (e)의 QQ plot에서 볼 수 있듯이 세 순열방법 모두에서 윈저평균차 검정통계량으로 생성된 귀무분포가 참귀무분포 기준선과 거의 일치하는 것을 볼 수 있다. 정확한 순열방법과 랜덤 순열방법에서는 윈저 자료의 비율이 발현수치를 나타내 보이는 유전자수의 비율에 비례적으로 커져야 참귀무분포 기준선에 근접하게 되고, 이분할 랜덤 순열방법의 경우에는 윈저 변환자료의 비율이 낮아도 참귀무분포 기준선에 근접함을 여러 차례 모의실험으로 파악할 수 있었다. 그러므로 윈저 자료를 사용하는 경우 세 순열방법에 큰 차이가 없이 모두 적절한 귀무분포를 생성할 수 있음을 알 수 있다.

이제 두 번째 모의실험의 경우로서 10%가 극도로 큰 양의 발현 유전자인 결과를 살펴보자. 정확한 순열방법으로 (a)에는 단순 평균차 통계량을, (c)에는 윈저평균차 통계량을 사용한 QQ plot과 히스토그램은 다음과 같다.

그림 3.4에서 단순 평균차 통계량의 경우인 (a)와 (b)를 살펴보면 그림 3.2에서와 마찬가지로 정확한 순열방법으로 단순 평균차 통계량은 적절한 귀무분포를 생성하는 방법이라 할 수 없으나 이는 10%가 매우 극도로 양의 발현 유전자인 것도 그 한 가지 이유이다. 반면에 윈저평균차 검정통계량은 그림 3.3에서와 마찬가지로 매우 적절한 귀무분포가 생성됨을 알 수 있다. 즉 윈저평균차 검정통계량은 매우 극단의 이상점 수치에도 로버스트한 방법이다.

이제 귀무분포로부터 추정된 경계값 A를 이용하여 검정통계량값의 꼬리부분의 분포로부터 FP와 TP를 계산한 결과는 다음과 같다.

표 3.1. 유의수준 α에서 FP와 TP - 양측검정

$\alpha/2(\%)$	Number	정확한 순열방법			랜덤 순열방법			이분할 랜덤 순열방법		
		$\overline{D}$	t	$\overline{D}_W$	$\overline{D}$	t	$\overline{D}_W$	$\overline{D}_m$	t_m	$\overline{D}_{W,m}$
0.5	FP	37.51	12.11	18.43	35.18	9.68	16.48	7.92	9.74	7.58
	TP	282	156	338	278	165	345	278	118	278
1	FP	59.80	22.31	32.37	62.40	19.60	29.60	19.10	20.96	18.36
	TP	338	217	376	343	220	376	348	188	348
2	FP	92.97	40.77	57.26	95.76	37.70	57.68	38.86	39.22	37.7
	TP	389	301	425	389	309	425	391	251	391
2.5	FP	107.17	51.57	68.26	112.00	47.94	69.74	47.76	48.52	46.48
	TP	405	338	452	407	338	453	408	208	408
5	FP	174.97	100.31	129.77	182.94	97.88	131.76	99.00	98.60	97.66
	TP	480	439	498	480	439	498	480	408	480

표 4.1. 유의수준 1%에서 의심되는 발현유전자

Gene name							
Gene 26	T95018	Gene 100	T5218	Gene 190	D63874	Gene 228	J03040
Gene 241	M36981	Gene 365	X14958	Gene 444	T59878	Gene 495	H20426
Gene 745	D16469	Gene 780	H40095	Gene 1042	R36977	Gene 1558	R49416
Gene 1671	M26383	Gene 1770	U17899	Gene 1771	J05032	Gene 1772	H08393
Gene 1900	X56597	Gene 1905	R23907				

표 3.1에서 보면 t 검정통계량과 수정 t 검정통계량의 경우 다른 평균차 검정통계량에 비해 상대적으로 TP의 수치가 작아서 그림 3.1에서와 같은 결론으로 t 검정통계량과 수정 t 검정통계량에 근거한 귀무통계량 z와 z_{Pan}은 극소수 샘플 자료에서 귀무분포를 생성하기에 적절하지 않다. 평균차 검정통계량의 FP와 TP를 살펴보면 순열방법에 따라 TP는 모든 유의수준에서 비슷한 수치를 보이지만 FP는 이분할 랜덤 순열방법에서 가장 작은 수치를 보인다. 그림 3.2에서와 같이 이분할 랜덤 순열방법을 적용한 평균차 검정통계량으로 적절한 귀무분포가 생성됨을 알 수 있다. 또한 윈저평균차 검정통계량의 FP는 작고 TP는 크다. 특히 정확한 순열방법과 랜덤 순열방법에서는 윈저 변화의 비율이 20%로 높게 정하여 TP가 매우 크다. 이분할 랜덤 순열방법은 윈저 변화의 비율이 6.25%로 매우 낮게 정하여 TP의 변화는 크지 않았지만 FP가 작아져서 더욱 정확하게 발현유전자를 발견할 수 있었다.

4. 적용 사례

본 논문에서 분석한 올리고뉴클레오티드(oligonucleotide) 마이크로어레이 자료는 정상인 22명과 악성 종양 환자 42명의 결절로부터 채취한 조직이다 (Alon 등, 1999). 자료의 표준화 과정은 조희진과 송혜향 (2004)에서와 같은 방법으로 시행하였다. 본 논문의 목적에 맞추어 각 군에서 랜덤하게 4개의 샘플을 추출하였으며 총 2,000개의 유전자를 분석하였다.

그림 3.3과 3.4에 제시된, 모의실험에서 좋은 결과를 보여준 정확한 순열방법을 이용한 윈저평균차 검정통계량으로 귀무분포를 생성하고 경계값을 구하여 유의수준 1%에서 유의 발현유전자로 탐색된 총 18개의 유전자를 표 4.1에 열거하였다.

총 18개의 유전자가 유의수준 1%에서 발현유전자로 탐색되었고 다른 검정방법의 결과와 비교해 보면, 정확한 순열방법을 이용한 단순평균차 검정통계량의 경우와 이분할 순열방법을 이용한 단순평균차 검정

통계량의 경우에 1% 유의수준에서 유의한 유전자수는 각각 20, 21개가 탐색되었고 거의 모든 유전자가 일치하였다. 그러나 이분할 랜덤 순열방법을 이용한 수정 t 검정통계량의 경우에 1% 유의수준에서 유의한 유전자는 총 22개가 되었고 표 4.1의 결과와 비교할 때 Gene72, 365, 444, 780, 1042, 1671의 6개만이 일치하였다. 즉, 평균차를 이용한 결과는 거의 비슷한 결과를 보이지만 t 검정통계량을 이용한 경우는 매우 다른 결과를 보이는 사실로부터 순열방법의 차이보다는 검정통계량의 차이가 결과에 큰 영향을 미치는 것을 알 수 있다.

5. 결론

본 연구의 목적은 마이크로어레이 극소수 샘플 자료에서 유의한 유전자를 탐색하기 위한 귀무분포 생성에 적절한 검정통계량의 모색이다. 상당수 또는 극히 일부만의 발현유전자가 섞여 있는 마이크로어레이 자료에서 귀무분포를 옳게 파악 또는 생성하지 못한다면 유의한 발현유전자의 탐색이 불가능하다. 이러한 과정에서 극소수 샘플의 어려움에도 불구하고 수 천개 또는 수 만개 유전자의 장점을 이용하는 순열방법을 이용하는 것이 가장 바람직하다. 본 연구의 모의실험에서 밝혀진 바는 일반적인 t 검정통계량과 Pan (2003a)이 제시한 수정 t 검정통계량을 어떤 순열방법과 병합하여도 참귀무분포를 생성하지 못한다는 점이다. 다시 말하면 마이크로어레이 극소수 샘플 자료에서는 검정통계량의 분모인 표준편차가 매우 불안정하여 귀무분포 생성에 적절하지 않으며, 그러나 본 연구에서 다루고 있지 않은 샘플수가 충분히 큰 경우에는 표준편차가 비교적 안정되므로 적절한 귀무분포가 생성될 것으로 짐작된다.

모의실험의 결과, 윈저평균차 검정통계량이 어떠한 순열방법에서도 참귀무분포에 가장 근접한 결과를 제시하고 더욱이 소수의 유전자가 발현된 경우에도 참귀무분포에 근접한 결과를 제시하였다. 윈저평균차를 사용하는 경우에 총 유전자 중에서 발현유전자수의 비율을 알 수 있다면 좋은 귀무분포를 생성할 수 있다. 만약 이 비율이 낮은데도 극도로 높은 비율의 자료를 윈저 변환시킨다면 이로부터 생성된 분포는 참귀무분포의 분산보다 작은 분산을 가지게 되어 적절한 귀무분포가 될 수 없다. 그러므로 윈저평균차를 사용하는 경우에는 발현유전자수의 비율을 되도록 정확히 추측하여야 하며 현실적으로 마이크로어레이 자료를 많이 다루는 연구자는 경험적으로 적당한 비율을 알 수 있다. 단순평균차 검정통계량을 사용하는 경우에는 이분할 랜덤 순열방법의 적용으로 옳은 귀무분포가 생성된다.

순열방법에서는 정확한 순열방법과 랜덤 순열방법의 차이가 크지 않았다. 극소수 샘플자료인 경우에는 정확한 순열방법의 경우수가 비교적 작으므로 정확한 순열방법의 선택이 더욱 바람직하다. 이로서 대칭 귀무분포가 만들어지고 참귀무분포에 가장 근접한 순열방법이기 때문이다. 그러나 샘플수가 큰 경우에는 정확한 순열방법의 경우수가 극단적으로 커지게 되므로 정확한 순열방법은 귀무분포를 생성하는데 계산시간이 매우 오래 걸리게 되므로 현실적으로 사용이 불가능하다. 그러므로 샘플수가 큰 경우에는 약간의 비대칭이 되는 귀무분포가 생성된다하더라도 귀무분포를 생성할 수 있다는 것이 중요한 문제이므로 랜덤 순열방법도 좋은 방법이라 하겠다. 만약 랜덤 순열방법으로 비대칭을 되도록 줄이기를 원한다면 랜덤순열의 반복수를 증가시켜 주는 것이 한 가지 해결책이다.

참고문헌

조희진, 송혜향 (2004). 변수가 관측치 보다 많은 자료에서 표식 유전자를 찾기 위한 방법, 가톨릭대학교 의과대학 의학통계학과 석사학위 논문집.

Albers, W., Bickel, P. J. and van Zwet, W. R. (1976). Asymptotic expansions for the power of distribution free tests in the one-sample problem, *Annals of Statistic*, 4, 108–156.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed

by oligonucleotide arrays, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745–6750.

Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics, *Annals of Mathematical Statistics*, **29**, 972–994.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of American Statistical Association*, **96**, 1151–2001.

Fisher, R. A. (1935, 1966). *The Design of Experiments (1st ed., 8th ed.)*. Oliver& Boyd, Edinburgh.

Gao, X. (2006). Construction of null statistics in permutation-based multiple testing for multi-factorial microarray experiments, *Bioinformatics*, **22**, 1486–1494.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations, *Annals of Mathematical Statistics*, **23**, 169–192.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connel, M. and Lee, J. K. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays, *Bioinformatics*, **15**, 1945–1951.

Lee, J.K. (2001). OAnalysis issues for gene expression array data, *Clinical Chemistry*, **47**, 1350–1352.

Lehmann, E. L. and Stein, C. (1949). On the theory of some non-parametric hypotheses, *The Annals of Mathematical Statistics*, **20**, 28–45.

Pan, W. (2003a). On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression, *Biometrics*, **19**, 1333–1340.

Pan, W., Lin, J. and Le, C. (2003b). A mixture model approach to detecting differentially expressed genes with microarray data, *Functional Integrative Genomics*, **3**, 117–124.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations, *Journal of the Royal Statistical Society*, **4**, 119–130.

Smyth, G. K., Yang, Y. H. and Speed, T. (2003). Statistical issues in cDNA microarray data analysis, *Methods in Molecular Biology*, **224**, 111-136.

Spino, C. and Pagano, M. (1991). Efficient calculation of the permutation distribution of trimmed means, *Journal of the American Statistical Association*, **86**, 729–737.

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

Welch, W. J. (1990). Construction of permutation Tests, *Journal of the American Statistical Association*, **85**, 693–698.

Xie, Y., Pan, W. and Khodursky, A. B. (2005). A note on using permutation-based false discovery rate estiimates to compare different analysis methods for microarray data, *Bioinformatics*, **21**, 4280–4288.

Zhao, Y. and Pan, W. (2003). Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments, *Bioinformatics*, **19**, 1046–1054.

Permutation-Based Test with Small Samples for Detecting Differentially Expressed Genes

Ju-Hyoung Lee[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 May 2009; accepted September 2009)

Abstract

In the analysis of microarray data with a small number of arrays, the most important task is the detection of differentially expressed genes by a significance test. For this purpose, one needs to construct a null distribution based on a large number of genes and one of the best way for constructing the null distribution for a small number of arrays is by means of permutation methods. In this paper we propose simple test statistics and permutation methods that are appropriate in constructing the null distribution. In a simulation study, we compare the null distributions generated by the proposed test statistics and permutation methods with the previous ones. With an example microarray data, differentially expressed genes are determined by applying these methods.

Keywords: Permutation test, micorarray data, differentially expressed genes, null distribution.

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