응용통계연구 (2010)
23(6), 1103–1113

DOI: 10.5351/KJAS.2010.23.6.1103

이산형 형질에 대한 가족자료 연관성 검정법 FBAT와 형제 전달 불균형 연관성 검정법 S-TDT의 비교

김한상[1] · 오영신[2] · 송혜향[3]

[1]가톨릭대학교 대학원 의학통계학과, [2]가톨릭대학교 대학원 의학통계학과
[3]가톨릭대학교 대학원 의학통계학과

(2010년 8월 접수, 2010년 10월 채택)

요 약

광범위하게 사용되는 가족자료에 근거한 연관성 검정법 FBAT를 형제 전달 불균형 연관성 검정법 S-TDT와 비교하였고, 특히 형제간의 공분산을 고려한 분산추정량을 사용한 수정 S-TDT로써 유전연관성의 정도가 다른 가족자료가 검정통계량값으로 구분될 수가 있다. 모의실험으로 세 검정법을 비교한 결과, 형제의 표현형 자료가 서로 독립이 아닌 경우에 세 검정법 모두의 제 1종 오류가 정해진 유의수준보다 커지며, 또한 FBAT의 검정력이 S-TDT와 수정 S-TDT의 검정력에 미치지 못한다. FBAT 검정법에서 가정하는 조건이 검정법의 효율성에 미치는 영향을 더욱 심도있게 평가하는 연구가 요구된다.

주요용어: 유전연관성, 가족자료, S-TDT, FBAT.

1. 서론

의학의 여러 질환에서 유전연관성 분석을 실시한 논문을 살펴보면 가족자료에 근거한 연관성 검정법(Familiy Based Association Test; FBAT)이 매우 널리 사용되고 있음을 본다. 그 이유는 이 FBAT의 유용성이 매우 광범위하기 때문이다. Rabinowitz와 Laird (2000)에 의해 소개된 FBAT 검정법은 부모로부터 질병자녀에게 전달 또는 비전달된 대립유전자형의 자료를 수집하여 연관성을 알아보는 전달 불균형 연관성 검정법(transmission disequilibrium test; TDT)을 비롯하여 부모의 유전자형 자료가 없는 경우의 분석도 포함하며, 또한 다양한 형태의 자녀의 표현형(phenotype), 즉 질병의 유무(affected/unaffected)인 이산형 형질(trait), 연속형 또는 미완결 생존자료 모두를 하나의 검정통계량에 집약시켜 분석할 수 있다. 그러나 FBAT 검정법은 부모의 유전자형 자료가 주어진 조건하에서 형제들의 표현형이 서로 독립이라는 가정이 요구되며, 따라서 Horvath 등 (2001a)은 토론부분에서 이러한 가정이 근사적으로 옳다는 가정하에서 검정력이 높은 연관성 검정법이라고 강조하였다.

환자의 부모 자료가 수집되기 불가능한 장, 노년기에 발병하는 질병의 유전연관성 연구에서 형제의 이산형 형질, 즉 형제의 질병 유무의 자료에 적용 가능한 전달 불균형 연관성 검정법(Sib Transmission Disequilibrium Test; S-TDT)이 Spielman과 Ewens (1998)에 의해 제안되었다. 이 S-TDT 검정법과 FBAT 검정법은 부모의 유전자형 자료가 없는 경우에 서로 비교될 수 있는 공통점을 가지고 있다. 한편 소수 가족수 자료에서 S-TDT 검정통계량의 분산추정량은 간혹 과소추정(under estimated)되는 경향을 보이며 가족관계를 감안한 분산추정량이 더욱 바람직한 것으로 짐작된다.

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

표 **2.1.** Genotype table with a disease allele A in the ith family

Status	Genotype			Total
	AA	Aa	aa	
Affected	r_{2i}	r_{1i}	r_{0i}	R_i
Unaffected	s_{2i}	s_{1i}	s_{0i}	S_i
Total	n_{2i}	n_{1i}	n_{0i}	N_i

본 논문에서는 이산형 형질에 적용한 FBAT 검정법에서 요구되는 유전자형 자료가 주어진 조건하에서 형제들의 표현형이 서로 독립이라는 가정이 성립되지 않는 경우의 FBAT 검정법의 검정력을 S-TDT 검정법과 또한 가족관계를 감안한 분산추정량을 이용한 수정 S-TDT검정법과 모의실험으로 비교한다.

2. S-TDT

Spielman과 Ewens (1998)가 제안한 S-TDT 통계량은 질병의 유무와 질병에 영향을 미친다고 짐작되는 마커와의 연관성을 검정한다. S-TDT에 앞서서 제안된 TDT 통계량은 부모의 유전자형 자료가 확보되어 질병과 연관되었다고 짐작되는 마커좌위에서 질병자녀에 전달 또는 전달되지 않은 대립유전자 수를 비교하는 검정통계량에 근거하지만, S-TDT는 부모의 자료 대신에 질병형제와 대조형제의 유전자형 자료를 사용한다. S-TDT 검정법은 우선 각 가족의 자료가 다변량 초기하분포한다는 가정하에서 통계량을 도출한 후 가족을 층(stratum)으로 두어 여러 가족의 통계량을 병합하여 분석하는 Mantel과 Haenszel (1959) 형태의 층화분석법이다. S-TDT의 최소 자료 충족 조건은 각 가족에서 적어도 1명의 질병형제와 1명의 대조형제 자료가 제시되어야 하며, 질병형제와 대조형제가 모두 동일 유전자형이어서는 안된다. 이제 총 가족수는 n이며, 한 가족의 자료는 표 2.1과 같다고 하자. 여기서 R_i, S_i와 N_i는 각각 i번째 가족의 질병형제수, 대조형제수 및 총 형제수를 나타낸다. 또한 n_{2i}, n_{1i}와 n_{0i}는 각 유전자형 AA, Aa, aa를 가진 형제수이다.

2.1. S-TDT 통계량

질병과 연관된 대립유전자가 A일 때 S-TDT 통계량은 질병형제의 대립유전자 A의 개수이며 각 가족에서의 통계량의 기댓값과 분산은 질병의 유무와 마커 유전자형간의 유전연관성이 없다는 귀무가설하에서 다변량 초기하분포(multivariate hypergeometric distribution)를 이용하여 구하며, 다시 모든 가족의 기댓값과 분산을 합하여 검정통계량을 구한다. 즉 다변량 초기하분포 가정으로 각 가족의 분할표 주변도수는 고정된 상수이며 랜덤확률변수는 r_{2i}와 r_{1i}뿐이다. S-TDT 통계량인 질병과 연관된 대립유전자 A의 개수는 다음과 같다.

$$Y = \sum_{i=1}^{n} (2r_{2i} + r_{1i}). \tag{2.1}$$

통계량 Y의 기댓값과 분산은 다변량 초기하분포하에서 다음과 같은 r_{2i}와 r_{1i}의 기댓값과 분산, 공분산 추정량을 이용하여 구한다.

$$\begin{aligned}
E(r_{2i}) &= R_i \frac{n_{2i}}{N_i}, & E(r_{1i}) &= R_i \frac{n_{1i}}{N_i}, \\
\widehat{\mathrm{Var}}(r_{2i}) &= R_i S_i \frac{n_{2i}(N_i - n_{2i})}{N_i^2 (N_i - 1)}, & \widehat{\mathrm{Var}}(r_{1i}) &= R_i S_i \frac{n_{1i}(N_i - n_{1i})}{N_i^2 (N_i - 1)}, \\
\widehat{\mathrm{Cov}}(r_{1i}, r_{2i}) &= -R_i S_i \frac{n_{2i} n_{1i}}{N_i^2 (N_i - 1)}.
\end{aligned} \tag{2.2}$$

이로부터 Y의 기댓값과 분산추정량 $\widehat{V}$는 다음과 같다.

$$\begin{aligned} E(Y) &= \sum_{i=1}^{n} E(2r_{2i} + r_{1i}) \\ &= \sum_{i=1}^{n} (2n_{2i} + n_{1i}) \frac{R_i}{N_i}. \\ \widehat{V} &= \sum_{i=1}^{n} \widehat{\mathrm{Var}}(2r_{2i} + r_{1i}) \\ &= \sum_{i=1}^{n} \left\{ 4\widehat{\mathrm{Var}}(r_{2i}) + \widehat{\mathrm{Var}}(r_{1i}) - 4\widehat{\mathrm{Cov}}(r_{2i}, r_{1i}) \right\} \\ &= \sum_{i=1}^{n} R_i S_i \frac{4n_{2i}n_{0i} + n_{1i}(N_i - n_{1i})}{N_i^2(N_i - 1)}. \end{aligned} \tag{2.3}$$

일반적으로 유전연관성 검정통계량을 $U = Y - E(Y)$로 제시하며, U를 정리하면 다음과 같다.

$$U = \sum_{i=1}^{n} \left[(2r_{2i} + r_{1i}) - \frac{R_i}{N_i}(2n_{2i} + n_{1i}) \right]. \tag{2.4}$$

U의 분산은 $\widehat{V}$와 같으므로 S-TDT 유전연관성 검정은 $Z = U/\sqrt{\widehat{V}}$가 대표본하에서 근사 정규분포함을 이용하여 단측으로 시행한다.

2.2. 수정 S-TDT 통계량

S-TDT 검정법에서 다변량 초기하분포 가정하에 구해지는 분산추정량은 양쪽 주변도수가 고정되었음을 가정하고 또한 비복원(without replacement) 추출로 인한 공분산이 고려된 것이며, 동일 가족에 속한 형제가 서로 연관되어 있음을 감안한 형제간의 공분산을 고려한 것은 아니다. 본 논문에서는 서로 다른 가족자료는 독립이지만 동일 가족에 속한 형제간의 공분산을 고려한 분산추정량을 제시하며, 이러한 분산추정량에 근거한 수정 S-TDT 검정통계량을 제안한다. 동일 가족에 속한 형제간의 공분산은 Slager와 Schaid (2001)가 다수 형제 자료에 근거한 질병-대조 연구계획의 유전연관성 연구에서 ITO 행렬방법을 이용하여 구한 것이며, 오영신 등 (2010)에서도 이 방법을 이용하였다. 이 ITO 행렬방법은 Li와 Sacks (1954)가 제안한 것으로 친족관계에 있는 두 명 개체의 유전자형의 결합확률분포(joint probability distribution)로부터 유도한다. 그러나 Slager와 Schaid (2001)와 오영신 등 (2010)은 S-TDT에서의 가족자료의 층화분석법과는 달리 질병-대조 연구계획의 검정통계량을 다루고 있다. 분산식의 자세한 과정은 오영신 등 (2010)에 설명되었으므로 본 논문에서는 간략히 제시한다. 동일 가족에 속한 형제간의 공분산을 고려하기위해 각 형제의 유전자형의 지시벡터를 정의한다. 각 가족에서 질병군과 대조군에 속한 형제의 유전자형 지시벡터가 각각 $\boldsymbol{y}_{ij} = (y_{ij0}, y_{ij1}, y_{ij2})'$와 $\boldsymbol{z}_{ij} = (z_{ij0}, z_{ij1}, z_{ij2})'$이며, 만약 i번째 가족에서 j번째 질병형제의 유전자형이 Aa라면 $\boldsymbol{y}_{ij} = (0, 1, 0)'$이 된다. 여기서 $\boldsymbol{x} = (0, 1, 2)'$는 추세를 반영하는 상수벡터이다.

동일 가족에 속한 형제간의 공분산을 고려한 귀무가설하에서의 분산추정량 $\widehat{V_A}$는 방금 정의한 지시벡터를 사용하여 다음과 같다.

$$V_A = \sum_{i=1}^{n} \left(1 - \frac{R_i}{N_i}\right)^2 \boldsymbol{x}' \left[\sum_{j=1}^{N_i} \mathrm{Var}(\boldsymbol{y}_{ij}) + 2 \sum_{j<j'} \sum \mathrm{Cov}(\boldsymbol{y}_{ij}, \boldsymbol{y}_{ij'}) \right] \boldsymbol{x}$$

표 **2.2.** 질병형제 3명과 대조형제 3명 가족자료(R_i=3, S_i=4, N_i=7)의 S-TDT 검정통계량 결과

가족	(r_2, r_1, r_0), (s_2, s_1, s_0)	U	S-TDT		수정 S-TDT	
			$\widehat{V}$	Z	$\widehat{V}_A$	Z_A
1	(0, 3, 0), (0, 0, 4)	1.7143	0.4898	2.4495	1.1208	1.6193
2	(3, 0, 0), (0, 0, 4)	3.4286	1.9592	2.4495	2.6989	2.0870
3	(2, 1, 0), (0, 4, 0)	1.1429	0.4082	1.7889	1.3057	1.0002
4	(2, 0, 1), (0, 0, 4)	2.2857	1.6327	1.7889	2.2491	1.5241

$$+\sum_{i=1}^{n}\left(\frac{R_i}{N_i}\right)^2 \boldsymbol{x'}\left[\sum_{j=1}^{N_i}\mathrm{Var}(\boldsymbol{z}_{ij}) + 2\sum_{j<j'}\sum \mathrm{Cov}(\boldsymbol{z}_{ij}, \boldsymbol{z}_{ij'})\right]\boldsymbol{x}. \tag{2.5}$$

귀무가설 하에서 $\mathrm{Var}(\boldsymbol{y}_{ij}) = \mathrm{Var}(\boldsymbol{z}_{ij})$은 다항분포(multinomial distribution)의 공분산행렬로 $\sigma_{ikk} = p_{ik}(1-p_{ik})$, $\sigma_{ikk'} = p_{ik}p_{ik'}$이 되며, 여기서 $p_{ik}\,(k=0,1,2)$는 유전자형 aa, Aa, AA의 확률이다. 또한 $\mathrm{Cov}(\boldsymbol{y}_{ij}, \boldsymbol{y}_{ij'})$는 i번째 가족의 j번째와 j'번째 질병형제의 공분산이며, $\mathrm{Cov}(\boldsymbol{z}_{ij}, \boldsymbol{z}_{ij'})$는 i번째 가족의 j번째와 j'번째 대조형제의 공분산을 나타내는데 여기서 이러한 형제간의 공분산은 Li와 Sacks (1954)의 ITO 행렬방법을 이용하여 계산한다. p_i가 위험인자인 대립유전자 A의 비율(allele frequency)일 때 유전자형 $p_{ik}(k=0,1,2)$의 추정량으로써 우선 $\hat{p}_i = (2n_{2i}+n_{1i})/2N_i$와 $\hat{q}_i = 1-\hat{p}_i$으로 $\hat{p}_{i0} = \hat{q}_i^2$, $\hat{p}_{i1} = 2\hat{p}_i\hat{q}_i$, $\hat{p}_{i2} = \hat{p}_i^2$을 계산하고 이를 대입하여 분산추정량 $\widehat{V}_A$을 구한다. 참고로 형제간 공분산을 고려하지 않는다면 AS-TDT 통계량의 분산은 S-TDT 통계량의 분산과 미소한 차이를 제외하고 동일하다. 즉 AS-TDT 통계량의 경우 분모가 N_i^3인 반면에 S-TDT 통계량의 경우 $N_i^2(N_i-1)$의 차이가 있을 뿐이다. 이제 유전연관성 검정은 $Z_A = U/\sqrt{\widehat{V}_A}$가 대표본하에서 근사 정규분포함을 이용하여 시행한다.

S-TDT 검정통계량의 분산추정량을 살펴보면 다변량 초기하분포하에서 구해진 식 (2.3)에 제시된 분산추정량 $\widehat{V}$은 표 2.1의 분할표의 주변도수에만 의존하는 것을 볼 수 있다. 즉 R_i, S_i와 N_i가 고정되었다고 가정할 때, 분산은 유전자형의 주변도수 n_{2i}, n_{1i}와 n_{0i}에 의해 결정되며, 가족자료에서 주로 n_{2i} 또는 n_{0i}이 0인 경우에 작은 값을 가지고, n_{1i}이 0인 경우에 큰 값을 가진다.

몇 가족자료로써 $\widehat{V}$값과 검정통계량 Z값을 살펴보면 표 2.2의 가족 1과 3은 n_{2i}가 0으로써 분산이 작은 값을 가지는 경우이고, 이와 반대로 가족 2와 4는 n_{1i}이 0으로써 분산이 큰 값을 가지는 경우이다. 이와 같은 분산추정량값이 유전연관성 검정통계량값에 영향을 미치고 있다. 표 2.2의 가족 1과 2를 비교해 보면 가족 2에서 질병형제와 대조형제의 유전자형이 양극단으로 가장 달라서 가족 2의 자료가 가족 1보다도 더욱 연관성이 높지만(가족 1, U=1.7; 가족 2, U=3.4), 가족 1의 분산이 매우 작아서 결과적으로 두 가족의 검정통계량 Z값은 서로 동일한 값을 가진다. 가족 3과 4의 경우에도 가족 4의 자료에서 연관성이 높으나 검정통계량 Z값에서 이러한 차이가 드러나지 않는다. 더욱 예민한 분산추정량이 요구된다 하겠다.

표 2.2의 가족자료 예에서 계산한 수정 S-TDT 검정통계량값을 앞에서 설명한 S-TDT 검정통계량값과 비교해 보면 동일 가족에 속한 형제간의 공분산을 고려한 분산 $\widehat{V}_A$가 다변량 초기하분포 가정하의 분산 $\widehat{V}$보다 크다. 또한 가족 1과 2의 분산추정량값이 서로 다르게 계산되어 결과적으로 질병형제와 대조형제의 유전자형이 양극단으로 가장 달라서 연관성이 높은 가족 2의 검정통계량 Z_A값이 2.1로써 가족 1의 Z_A값 1.6보다 커서 가족 1과 가족 2의 검정결과가 구분된다. 마찬가지로 연관성이 높은 가족 4의 검정통계량 Z_A값이 1.5로 가족 3의 Z_A값 1.0보다 커서 두 가족의 검정결과가 구분된다. 그러나 S-TDT 검정통계량의 분산과 수정 S-TDT 검정통계량의 분산을 비교해 보면 형제간의 공분산을 고려한 분산 $\widehat{V}_A$가 공분산을 고려하지 않은 분산 $\widehat{V}$보다 크기 때문에 수정 S-TDT 검정통계량값은 S-TDT

검정통계량값보다 덜 유의한 것으로 나타났다. 한편 분산 $\widehat{V}_A$가 분산 $\widehat{V}$보다 항상 큰 것은 아니므로 모의실험으로 두 검정통계량의 검정력이 비교되어야 한다.

3. FBAT

TDT 검정법은 부모의 유전자형 자료와 질병자녀의 유전자형 자료가 요구되는 반면에, S-TDT 검정법은 질병형제와 대조형제의 유전자형 자료가 요구되며, 이러한 연구계획과 검정통계량의 다양함은 질병에 따라서 더욱 바람직한 연구계획이 다르기 때문이다. 희귀질병의 경우에는 질병형제의 자료가 더욱 유용하고, 모집단 질병율(population prevalence)이 높은 경우에는 질병형제와 대조형제의 자료를 모두 사용하는 연구계획이 바람직하다 (Lange와 Laird, 2002). FBAT 검정법은 개체에게서 수집된 여러 표현형 자료를 동일 검정통계량에 표현할 수 있다는 장점뿐만 아니라, 부모와 질병형제의 자료만 있는 경우나 부모의 자료가 없이 질병형제와 대조형제의 자료만이 있는 경우를 동일 검정통계량에 표현할 수 있으며 또한 여러 형태의 표현형 자료를 검정통계량에 병합하여 사용할 수도 있다. 이러한 여러 장점 때문에 FBAT 검정법은 널리 이용되고 있으나 매우 포괄적인 반면에 여러가지 조건을 전제로 하여 검정통계량의 분포가 유도되고 있음이 사실이다. Laird 등 (2000)이 제시한 FBAT 통계량 U_F와 분산추정량 $\widehat{V}_F$는 다음과 같다.

$$U_F = \sum_{i=1}^{n}\sum_{j=1}^{N_i} T_{ij}[X_{ij} - E(X_{ij})]. \tag{3.1}$$

$$\widehat{V}_F = \sum_{i=1}^{n}\sum_{j=1}^{N_i} T_{ij}^2 \mathrm{Var}(X_{ij}). \tag{3.2}$$

여기서 Y_{ij}가 i번째 가족의 j번째 형제의 질병여부를 나타내어 질병형제일 때 1이고, 대조형제일 때 0으로 지정할 때 T_{ij}는 $T_{ij} = Y_{ij} - \gamma$로 정의한다. 여기서 사용된 γ $(0 < \gamma < 1)$는 질병형제와 대조형제에 대한 가중치 상수(constant offset)이며 γ값이 0인 경우에는 질병형제에 모든 가중치를 부여한 것이고 질병형제와 대조형제를 모두 사용하고자 할 때 0이 아닌 상수를 채택한다. 일반적으로 상수 γ는 귀무가설하에서 U_F의 분산이 최소가 되도록 정한다 (Horvath 등, 2001a). 한편 식 (3.1)의 X_{ij}는 자녀의 마커에서의 질병과 관련된 대립유전자의 개수 0, 1, 2, 다시 말하면 유전자형을 나타내는 확률변수이다. FBAT 검정법에서는 부모의 유전자형과 자녀의 표현형 형질 Y_{ij}를 알고 있다고 가정하며 따라서 부모의 유전자형과 자녀의 표현형 형질 Y_{ij}는 고정된 상수로 취급하므로 T_{ij}도 역시 상수이다. 자녀의 유전자형만이 랜덤확률변수이다.

FBAT 통계량 U_F와 앞에서 설명한 S-TDT 통계량의 관계를 살펴보면, 우선 FBAT 통계량 U_F는 S-TDT 통계량 U와 다른 형태를 보인다. 구체적으로 두 통계량이 서로 일치하는 경우를 알아보기 위해 식 (3.1)의 FBAT 통계량 U_F를 표 2.1의 기호로써 표현해 보면 다음과 같다. 동일 부모의 자녀들은 $E(X_{ij})$값이 같으므로, 잠정적으로 가족을 나타내는 첨자만을 사용하여 $E(X_i)$로 표시한다.

$$U_F = \sum_{i=1}^{n} \left[2\left\{(1-\gamma)r_{2i} - \gamma s_{2i}\right\} + \left\{(1-\gamma)r_{1i} - \gamma s_{1i}\right\} - (1-\gamma)R_i E(X_i) + \gamma S_i E(X_i)\right]. \tag{3.3}$$

FBAT에서 γ값을 지정하여 S-TDT 통계량 U와 FBAT의 통계량 U_F를 일치시키려면 각 가족에서 질병형제수와 대조형제수를 사용한 상수 $\gamma = R_i/N_i$으로 지정되어야 한다. 다시 말하면 식 (3.3)에 $\gamma = R_i/N_i$을 대입하면 $E(X_i)$가 포함된 부분은 서로 삭제되고 나머지 부분이 바로 식 (2.4)의 S-TDT 통계량 U와 동일하다. 즉 각 가정의 γ를 다르게 선택하지 않고서는 두 통계량 U와 U_F가 같아질 수 없다.

표 **3.1.** Conditional probabilities with no parent's genotype available.

가족 형태	유전자형	유전자형에 대한 확률
II	AA, Aa	$P(AA) = \frac{n_2}{N_i},\ P(Aa) = \frac{n_1}{N_i},$ $P(AA, AA) = \frac{n_2(n_2-1)}{N_i(N_i-1)},$ $P(Aa, Aa) = \frac{n_1(n_1-1)}{N_i(N_i-1)},\ P(AA, Aa) = \frac{n_2 n_1}{N_i(N_i-1)}.$
III	AA, aa or AA, Aa, aa	$P(AA) = P(aa) = \frac{4^{N_i-1} - 3^{N_i-1}}{4^{N_i} - 2(3)^{N_i} + 2^{N_i}},$ $P(Aa) = \frac{2(4)^{N_i-1} - 4(3)^{N_i-1} + 2^{N_i}}{4^{N_i} - 2(3)^{N_i} + 2^{N_i}},$ $P(AA, AA) = P(aa, aa) = 0.5P(AA, Aa) = 0.5P(aa, Aa),$ $= P(AA) = P(aa) = \frac{4^{N_i-2} - 3^{N_i-2}}{4^{N_i} - 2(3)^{N_i} + 2^{N_i}},$ $P(AA, aa) = \frac{4^{N_i-2}}{4^{N_i} - 2(3)^{N_i} + 2^{N_i}},$ $P(Aa, Aa) = \frac{4^{N_i-1} - 8(3)^{N_i-2} + 2^{N_i}}{4^{N_i} - 2(3)^{N_i} + 2^{N_i}}.$

일반적으로 FBAT 통계량의 프로그램에서는 모든 가족에 걸쳐 공통된 γ를 채택한다. FBAT 통계량의 프로그램에서 명시하지 않으면 디폴트(default)는 $\gamma = 0$이며 이러한 0값은 질병형제만으로 검정통계량을 생성하는 것이며 본 논문에서 이 0값을 채택하였다. 이제 검정통계량의 기댓값과 공분산을 구하는 과정을 설명한다. 기댓값 $E(X_{ij})$와 분산 $\mathrm{Var}(X_{ij})$과 공분산 $\mathrm{Cov}(X_{ij}, X_{ij'})$는 부모의 자료가 없는 경우에 동일 가족에 속한 형제의 유전자형의 모든 가능한 경우를 고려하여 구한다. 예를 들어서, 어떤 한 가족의 세 자녀의 유전자형이 AA, Aa와 Aa이고 첫번째 자녀가 질병형제일 때 모든 가능한 경우의 유전자형 순열은 (AA, Aa, Aa), (Aa, AA, Aa), (Aa, Aa, AA)이 되며, 각 경우의 확률은 동일하게 1/3이다. 따라서 질병형제만으로 생성하는 통계량의 기댓값은 $E(X_i) = 2(1/3) + 1(2/3) = 3/4$이고, 분산은 $\mathrm{Var}(X_i) = (2 - 4/3)^2/3 + (1 - 4/3)^2 2/3 = 2/9$이며 공분산은 0이 된다. 임의의 가족자료의 경우에 대해서 모든 가능한 경우의 순열에 대한 통계량 계산에 도움이 되도록 Horvath 등 (2001b)은 유전자형의 확률을 공식화하여 제시하였다. 이제 i번째 가족에서의 A 대립유전자의 갯수가 W_i일 때, 다시 말하면 식 (3.1)에서 $W_i = \sum_{j=1}^{N_i} T_{ij} \cdot X_{ij}$일 때 FBAT의 통계량은 $U_F = \sum_i [W_i - E(W_i)]$이 되고, W_i의 기댓값 $E(W_i)$와 분산 $\mathrm{Var}(W_i)$은 다음과 같다.

$$E(W_i) = \sum_j T_{ij} E(X_{ij}), \tag{3.4}$$

$$\begin{aligned}\widehat{\mathrm{Var}}(W_i) &= \sum_j T_{ij}^2 \mathrm{Var}(X_{ij}) + \sum_j \sum_{j' \neq j} T_{ij} T_{ij'} \mathrm{Cov}(X_{ij}, X_{ij'}) \\ &= \left(\sum_j T_{ij} \right)^2 \sum_g \sum_{g'} \left[X(g) \left(P\left(gg'\right) - P(g)p\left(g'\right) \right) X\left(g'\right)^T \right] \\ &\quad + \sum_j T_{ij}^2 \left[\sum_g X(g) X(g)^T P(g) - \sum_g \sum_{g'} X(g) P\left(gg'\right) X\left(g'\right)^T \right]. \end{aligned} \tag{3.5}$$

식 (3.4)와 (3.5)는 모든 경우를 포괄하는 FBAT 통계량을 수식으로 표현하였으며, 본 논문에서 다루

표 **4.1.** Spielman과 Ewens (1998)의 세 가족자료 분석결과

가족	(r_2, r_1, r_0), (s_2, s_1, s_0)	U	S-TDT		수정 S-TDT		FBAT		
			$\widehat{V}$	Z	$\widehat{V}_A$	Z_A	U_F	$\widehat{V}_F$	Z_F
1	(2, 1, 0), (0, 4, 0)	1.4286	0.4082	1.7889	1.3057	1.0002	1.1430	0.4080	1.7890
2	(0, 1, 0), (0, 2, 2)	0.4000	0.2400	0.8165	0.2928	0.7392	0.4000	0.2400	0.8165
3	(1, 0, 0), (0, 1, 2)	1.2500	0.6875	1.5076	0.6035	1.6090	1.0000	0.7780	1.1340
Total		2.7929	1.3357	2.4166	2.2020	1.8821	2.5430	1.4260	2.1295

표 **4.2.** Sherrington 등 (1988)의 자료분석 결과

가족	(r_2, r_1, r_0), (s_2, s_1, s_0)	U	S-TDT		수정 S-TDT		FBAT		
			$\widehat{V}$	Z	$\widehat{V}_A$	Z_A	U_F	$\widehat{V}_F$	Z_F
1	(0, 0, 1), (0, 1, 0)	−0.5000	0.2500	−1.0000	0.1250	−1.4142	−0.5000	0.2500	−1.0000
2	(5, 1, 0), (1, 3, 0)	1.4000	0.6400	1.7500	2.0352	0.9814	1.4000	0.6400	1.7500
3	(0, 1, 0), (2, 0, 0)	−0.6667	0.2222	−1.4142	0.1790	−1.5757	−0.6667	0.2222	−1.4144
4	(2, 1, 0), (0, 4, 1)	1.6250	0.7701	1.8518	1.9428	1.1659	2.0000	1.4510	1.6600
5	(2, 0, 0), (0, 2, 0)	1.0000	0.3333	1.7321	0.4375	1.5119	1.0000	0.3333	1.7320
6	(0, 1, 0), (2, 2, 0)	−0.4000	0.2400	−0.8165	0.2928	−0.7392	−0.4000	0.2400	−0.8165
Total		2.4583	2.4556	1.5688	5.0123	1.0981	2.8330	3.1370	1.6000

고 있는 대립유전자가 둘인 경우의 통계량의 구체적인 유도과정은 부록에 제시하였다. 식 (3.5)에서 $X(g)$와 $P(g)$는 한 개체의 유전자형과 이 유전자형에 대한 확률이며, $P(gg')$는 두 형제 유전자형의 결합확률을 나타낸다. 부모의 자료가 없는 경우의 $P(g)$와 $P(gg')$를 표 3.1에 제시하였다 (Horvath 등, 2001b). 표 3.1의 type 2는 유전자형 aa를 가진 형제가 없는 가족의 경우이며, type 3는 유전자형 Aa를 가진 형제가 없는 가족의 경우 또는 형제의 자료에서 모든 유전자형이 있는 가족의 경우에 해당한다.

이제 $\sum_i \mathrm{Var}(W_i) = \mathrm{Var}(U_F) \equiv V_F$을 이용하여 FBAT 유전연관성 검정은 검정통계량 $Z_F = U_F/\sqrt{\widehat{V}_F}$가 대표본하에서 근사정규분포함을 이용하여 시행한다. 여러 검정통계량의 차이가 검정력에 미치는 영향을 모의실험으로 알아보게 된다.

4. 예제자료

첫 번째 예제 자료는 Spielman과 Ewens (1998)의 논문에 사용한 것으로 총 세 가족의 자료이며 이 자료를 이용하여 각각 S-TDT, 수정 S-TDT, FBAT 검정통계량으로 분석한 결과가 표 4.1에 제시되었다.

표 4.1의 분석결과를 살펴보면 S-TDT의 각 가족의 분산이 모두 작은 값을 가지기 때문에 결과적으로 모든 가족을 병합한 Z값(2.4166)이 다른 두 분석결과보다도 큰 것을 확인할 수 있다. 반대로 수정 S-TDT의 분산이 S-TDT에 비해 조금 크기 때문에 Z_A값(1.8821)이 작다. FBAT의 Z_F값(2.1295)은 두 분석결과의 중간 정도이다.

두 번째 예제 자료는 Sherrington 등 (1988) 논문의 가계도 자료로써 정신분열증(schizophrenia)과 연관되었다고 짐작되는 마커에서의 유전자형과 질병의 여부에 대한 정보이다. 제 2장에서 설명한 S-TDT의 최소 자료 충족조건이 만족되는 여섯 가족의 형제자료가 제시되었다.

표 4.2의 분석결과를 살펴보면 S-TDT와 FBAT는 n_{2i} 또는 n_{0i}이 0인 가족의 경우에 비교적 작은 분산을 가져서 검정통계량 Z값이 크다. 특히 가족 2의 경우 수정 S-TDT의 분산은 S-TDT의 분산보다 매우 크다. 가족 1, 3, 6의 경우에는 연관성과는 반대의 결과, 그러나 유의하지는 않은 결과가 도출되었음을 볼 수 있다.

표 **5.1.** S-TDT, 수정S-TDT, FBAT 검정통계량의 유의수준과 검정력 비교

분포차 a	상관성	2형제, 3형제			3형제		
		S-TDT	수정 S-TDT	FBAT	S-TDT	수정 S-TDT	FBAT
0	0	0.046	0.119	0.046	0.047	0.072	0.060
	0.3	0.048	0.119	0.050	0.060	0.087	0.072
	0.5	0.059	0.139	0.059	0.068	0.090	0.076
0.3	0	0.638	0.744	0.599	0.562	0.628	0.463
	0.3	0.724	0.813	0.684	0.611	0.670	0.513
	0.5	0.811	0.880	0.782	0.677	0.734	0.595
0.5	0	0.959	0.984	0.945	0.912	0.940	0.870
	0.3	0.972	0.987	0.959	0.932	0.952	0.885
	0.5	0.983	0.988	0.978	0.956	0.975	0.919
0.8	0	0.995	0.997	0.994	0.995	0.997	0.995
	0.3	1.000	1.000	1.000	0.998	0.999	0.997
	0.5	1.000	1.000	1.000	1.000	1.000	1.000

분포차 a	상관성	3형제, 4형제			4형제		
		S-TDT	수정 S-TDT	FBAT	S-TDT	수정 S-TDT	FBAT
0	0	0.049	0.054	0.071	0.045	0.034	0.093
	0.3	0.053	0.055	0.076	0.062	0.052	0.101
	0.5	0.067	0.074	0.093	0.083	0.070	0.112
0.3	0	0.486	0.499	0.336	0.487	0.456	0.288
	0.3	0.557	0.570	0.380	0.530	0.492	0.294
	0.5	0.591	0.603	0.425	0.557	0.534	0.343
0.5	0	0.836	0.849	0.723	0.849	0.827	0.648
	0.3	0.882	0.904	0.789	0.880	0.865	0.712
	0.5	0.897	0.904	0.789	0.893	0.876	0.717
0.8	0	0.990	0.992	0.984	0.995	0.995	0.980
	0.3	0.997	0.997	0.988	0.997	0.997	0.983
	0.5	0.999	0.999	0.994	0.998	0.998	0.983

5. 모의실험

모의실험은 2형제와 3형제의 조합 및 3형제와 4형제의 조합, 3형제 또는 4형제만의 네 가지 경우에서 각각 200명 정도의 가족자료를 생성하였다. 대립유전자 A의 비율은 0.2로 고정시키고, 형제간의 표현형 자료를 우선 다변량 정규분포로써 생성하며 대립가설하에서 유전자형 AA의 표현형 자료의 분포평균은 a, Aa의 평균은 0, aa의 평균은 $-$a로 정하였고, 동일가족에 속한 형제의 표현형 자료간의 상관성은 0, 0.3, 0.5로 변화시켜 생성한 후 이변량화시켜 큰 수치를 질병군으로 채택하였다. 모의실험에서 $\alpha = 0.05$ 수준에 해당하는 제 1종 오류와 검정력 결과가 다음 표 5.1에 제시되었다.

모의실험의 결과를 비교해 보면 형제의 표현형 자료가 서로 독립인 경우에는 S-TDT의 제 1종 오류가 모든 경우에서 정해진 $\alpha = 0.05$ 수준에 가깝다. 검정력은 전체적으로 형제수가 작아져 가족의 수가 많아질수록, 또 형제의 표현형 자료의 상관성이 커질수록 커진다. 이러한 결과는 분산에 따라서 결정되어지며 3형제 이하의 자료에서는 수정 S-TDT의 분산이 다른 두 통계량의 분산보다 작아져서 검정력이 높아지며 반대로 4형제 이상에서는 분산이 다른 두 통계량의 분산보다 커져서 검정력이 S-TDT보다 낮아지지만 큰 차이는 아니다. 3형제와 4형제를 병합한 자료에서 유의수준이 $\alpha = 0.05$ 수준에 가까우므로

검정력을 비교할 수 있으므로 검정력을 살펴보면 수정 S-TDT의 검정력이 다른 두 통계량보다 조금 높은 것을 확인할 수 있다.

6. 결론

본 논문에서는 가족자료 연관성 검정법으로써 S-TDT와 FBAT 검정통계량의 차이와 효율성을 비교해 보았다. 앞서 소개한 검정통계량들은 모두 가족을 층으로 한 층화분석법에 기초한다. 층화분석을 하게 되면 각 가족 내에서의 영향성이 고려된다는 장점이 있지만 또한 각 층의 자료수가 너무 작아지기 때문에 분산이 불안정하게 나타난다는 단점이 있다. 따라서 분산을 안정화시키기 위해 수정 S-TDT방법을 제안하였으며 이를 요즘 많이 사용되고 있는 FBAT방법을 사용해서 비교해 보았다. 모의실험으로 비교한 결과, 형제수에 따라 그리고 형제의 상관성에 따라 각 검정통계량값이 차이가 있는 것으로 나타났다. 특히 다수 형제 가족수의 비중이 커질수록 S-TDT와 수정 S-TDT 결과에 많은 차이가 있으며 4형제 이상부터는 S-TDT의 분산이 작아지면서 Z값이 전체적으로 커지는 경향이 나타나고, 반대로 수정 S-TDT는 분산이 S-TDT보다 상대적으로 커져서 Z값이 작아지는 경향을 보인다. 따라서 부모의 유전자형 자료가 없는 경우에 연구자는 각 연구 상황에 따라서 최적의 분석법을 선택하여 사용해야 한다. 현대 사회의 소수 가족자료의 분석에서 작은 표본수의 문제는 간과할 수 없는데 본 논문에서 이데 대해 충분히 밝히지 못하였으며 장차 광범위한 모의실험으로 밝혀져야 한다.

부록. 대립유전자가 둘인 경우의 FBAT 통계량의 기댓값과 분산

대립유전자가 둘인 경우의 식 (3.4)의 (3.5)의 구체적인 유도과정은 다음과 같다. 이제 Y_{ij}가 질병형제의 경우 1이고 대조형제일 때 0으로 지정하며, 또한 $T_{ij} = Y_{ij} - \gamma$이며, 여기서 $\gamma = 0$으로 둔다. 식 (3.4)와 (3.5)는 R_i, $E(X_i|S)$, $V(X_i|S)$와 $\mathrm{Cov}(X_i, X_i\prime)$으로 표현된다. 즉

$$\begin{aligned} E(W_i) &= \sum_j T_{ij}E(X_{ij}) = R_iE(X_i|S). \\ \mathrm{Var}(W_i) &= \sum_j T_{ij}^2\mathrm{Var}(X_{ij}) + \sum_j \sum_{j'\neq j} T_{ij}T_{ij'}\mathrm{Cov}(X_{ij}, X_{ij'}) \\ &= R_i\mathrm{Var}(X_i|S) + R_i(R_i-1)\mathrm{Cov}(X_i, X_i\prime). \end{aligned}$$

위에서 각각의 항인 $E(X_i|S)$, $V(X_i|S)$와 $\mathrm{Cov}(X_i, X_i\prime)$는 다음과 같다.

$$\begin{aligned} E(X_i|S) &= 2P(AA) + P(Aa) \\ \mathrm{Var}(X_i|S) &= (2 - E(X_i|S))^2P(AA) + (1 - E(X_i|S))^2P(Aa) + (0 - E(X_i|S))^2P(aa). \\ \mathrm{Cov}(X_i, X_i') &= [4P(AA, AA) + 4P(AA, Aa) + P(Aa, Aa)] \\ &\quad - [4P(AA)P(AA) + 4P(AA)P(Aa) + P(Aa)P(Aa)]. \end{aligned}$$

이제 표 3.1에 제시된 가족형태 II의 각각의 확률값을 대입하여 기댓값과 분산을 구한다.

$$\begin{aligned} E(W_i) &= R_iE(X_i|S) = (2n_{2i} + n_{1i})\frac{R_i}{N_i}. \\ \widehat{\mathrm{Var}}(W_i) &= \frac{R_iS_i}{N_i^2(N_i-1)}\left(4n_{2i}n_{0i} + n_{1i}(n_{2i} + n_{0i})\right). \end{aligned}$$

가족형태 III의 경우에 기댓값과 분산은 다음과 같이 구해진다.

$$E(W_i) = R_i E(X_i|S) = R_i.$$
$$\widehat{\mathrm{Var}}(W_i) = R_i \frac{2(4)^{N_i-1} - 2(3)^{N_i-1}}{4^{N_i} - 2(3)^{N_i} + 2^{N_i}} + R_i(R_i - 1)\left(\frac{4^{N_i} - 20(3)^{N_i-2} + 2^{N_i}}{4^{N_i} - 2(3)^{N_i} + 2^{N_i}} - 1\right).$$

참고문헌

오영신, 김한상, 송혜향 (2010). 형제자료에 근거한 유전연관성 추세 검정법의 비교, <응용통계연구>, **23**, 845–855.

Horvath, S., Xu, X. and Laird, N. M. (2001a). The family based association test method: Strategies for studying general genotype-phenotype associations, *European Journal of Human Genetics*, **9**, 301–306.

Horvath, S., Xu, X. and Laird, N. M. (2001b). *The Family Based Association Test Method: Computing Means and Variances for General Statistics*, Technical report.

Laird, N. M., Horvath, S. and Xu, X. (2000). Implementing a unified approach to family based tests of association, *Genetic Epidemiology*, **19**, S36–42.

Lange, C. and Laird, N. M. (2002). Power calculations for a general class of family-based association tests: Dichotomous traits, *American Journal of Human Genetics*, **71**, 575–584.

Li, C. C. and Sacks, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices, *Biometrics*, **10**, 347–360.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, **22**, 719–748.

Rabinowitz, D. and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information, *Human Heredity*, **50**, 211–223.

Sherrington, R., Brynjolfsson, J., Petursson, H., Potter, M., Dudleston, K., Barraclough, B., Wasmuth, J., Dobbs, M. and Gurling, H. (1988). Localization of a susceptibility locus for schizophrenia on chromosome 5, *Nature*, **336**, 164–167.

Slager, S. L. and Schaid, D. J. (2001). Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects, *American Journal of Human Genetics*, **68**, 1457–1462.

Spielman, R. S. and Ewens, W. J. (1998). A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test, *American Journal of Human Genetics*, **62**, 450–458.

Comparison of the Family Based Association Test and Sib Transmission Disequilibrium Test for Dichotomous Trait

Han-Sang Kim[1] · Young-Sin Oh[2] · Hae-Hiang Song[3]

[1]Department of Biostatistics, Graduate School, The Catholic University of Korea
[2]Department of Biostatistics, Graduate School, The Catholic University of Korea
[3]Department of Biostatistics, Graduate School, The Catholic University of Korea

(Received August 2010; accepted October 2010)

Abstract

An extensively used approach for family based association test(FBAT) is compared with the sib transmission/disequilibrium test(S-TDT), and in particular the adjusted S-TDT, in which the covariance among related siblings is taken into consideration, can provide a more sensitive test statistic for association. A simulation study comparing the three test statistics demonstrates that the type I error rates of all three tests are larger than the prespecified significance level and the power of the FBAT is lower than those of the other two tests. More detailed studies are required in order to assess the influence of the assumed conditions in FBAT on the efficiency of the test.

Keywords: Genetic association test, family data, FBAT, S-TDT.

[3]Corresponding author: Professor, Department of Biostatistics, Graduate School, Integrated Research Center for Genome Polymorphism, The Catholic University of Korea, 505 Banpo-Dong, Seocho-Gu, Seoul 137-701, Korea. E-mail: hhsong@catholic.ac.kr