형제자료에 대한 양적형질의 유전자 관련성 분석방법의 비교

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 민 지

형제자료에 대한 양적형질의 유전자 관련성 분석방법의 비교

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2003년 6월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 민 지

김민지의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2003년 6월 일

감사의 글

어느덧 2년이라는 시간이 지났습니다. 처음 대학원을 들어올 때의 설레임과 기대감을 다시 한번 되새겨보게 됩니다. 2년여의 대학원 생활동안 고생스럽고 짜증나는 일도 많았지만 그보다 더 많은 것을 배울 수 있는 소중한 시간이었습니다. 이제는 다시 초심으로 돌아갈 때가 온 듯합니다. 사회인으로 첫발을 내딛게 되는 지금 이 순간 저를 이 자리에 있게 해주신 모든 분들께 감사의 말을 전하고자 합니다.

먼저 이 논문이 완성되기까지 지속적인 가르침으로 저를 다듬어주신 김동기 교수님께 감사드립니다. 유전통계라는 새로운 분야를 알게 해 주시고, 관심갖게 해주신, 그리고 항상 온화하고 따뜻한 마음으로 대해주시던 임길섭 교수님, 바쁘신 와중에도 논문이 완성되기까지 도와주신 박상언 교수님, 지금까지 통계를 공부할 수 있는 계기가 되어주시고 힘이 되어주신 조진남 교수님, 김동건 교수님께도 진심으로 감사드립니다.

대학원 생활에서 가장 큰 힘이 되어주시고 이 논문이 완성되기까지 끝없는 질책과 격려를 아끼지 않은 기준오빠에게 진심으로 감사하다는 말을 전하고 싶습니다. 항상 밝은 웃음으로 대해주시는 그의 영원한 반려자 박춘선 선생님께도 감사드립니다. 때로는 어이없는 농담으로 웃음을 주시지만, 처음 대학원생활에 익숙하지 않던 그 시절부터 지금까지 항상 든든한 힘이 되어주시고 많은 것을 가르쳐주신 성민오빠에게 진심으로 감사드립니다. 친구같지만 타인을 배려할 줄 아는 나의 동기 원열오빠, 힘든 일이 있을 때 웃음을 주는 무영오빠, 언제나 미소를 잃지 않고 나에게 많은 도움을 주었던 찬미언니, 부족한 나를 잘 따르고 믿어주었던 나의 술친구 정숙에게는 미안함과 고마운 마음이 듭니다. 많은 시간을 함께 보내지는 않았지만 따뜻한 마음을 가지고 계신 윤주, 효미, 영진, 명희, 현선언니, 우선, 성재, 영선, 현철오빠, 봉섭씨, 학부동기로 그리고 대학원 선배로 다시 만난 경민언니, 지금은 휴학하고 없지만 항상 잘해주지 못해서 미안한 은정이, 동기 장섭씨, 이제 막 대학원 생활을 시작한 수옥씨, 신영, 은혜, 은정, 혜리씨에게도 감사드립니다.

늘 저를 고민하게 해주신 지선하 교수님께 감사드립니다. 같이 보낸 시간이 길지는 않지만 저에게 희망과 미소를 주신 국민건강증진연구소 식구 윤지은, 주미현, 한순실, 최승주 선생님께도 진심으로 감사드립니다.

항상 나의 위안이 되어주고 늘 말없이 날 격려해준 나의 사랑하는 친구 찬희

에게 진심으로 감사드립니다. 공통의 추억을 간직해 더 소중한 초등학교 동창 보람, 춘선, 필주, 혜진, 소소한 일상과 인생사에 대해 같이 고민하고 서로가 서로에게 용기가 되어준 남은, 연희, 사춘기시절 서로의 고민을 나누었던 재명, 종철, 재원, 언제나 새로움을 추구하고 도전하려는 나의 남동생 우석에게 감사의 말을 전합니다.

무엇보다 지금의 저를 세상에 있게 해주시고 저의 안식처가 되어주시며 인생의 디딤돌이 되어주신 부모님께 깊이 감사드립니다. 지금 이 순간, 자립심과 항상 노력하는 자세를 갖게 해주시고 세상에 맞서 당당하게 살아갈 수 있게 해주신 부모님의 사랑이 더욱 크게 느껴집니다. 타국에 홀로 떨어져 외로움과 싸우며 공부하고 있는 친구같은 동생 혜지와 항상 나의 짜증을 묵묵히 받아주고 늘 우리집의 활력소가 되어주는 착한 막내동생 예리에게도 진심으로 감사드립니다. 새삼 부모님과 동생의 사랑이 더욱 절실하게 느껴집니다. 그 동안 바쁘고 힘들다는 핑계로 잘해드리지 못한 것에 죄송한 마음이 앞섭니다.

이제 새로운 세상으로 나아가는 기로에 서 있습니다. 많은 사람들에게 받은 은혜를 감사의 말로 다 갚을 수는 없지만 앞으로 더 발전하는 모습을 보이는 것이 이분들께 은혜를 보답하는 일이라 믿고 항상 노력하겠습니다.

2003년 7월

김 민 지 올림

목 차

표 차 례

국 문 요 약

형제자료에 대한
양적형질의 유전자 관련성 분석방법 비교

Spielman 등(1993)에 의해 개발된 TDT(transmission/disequilibrium test)는 가족자료를 기초로 한 질적형질의 유전자 관련성(association)을 분석하기위한 방법이다. 최근 이러한 TDT는 Allison(1997) 등에 의해 질적형질뿐만 아니라 양적형질에까지 확장되고 있다.

본 논문에서는 양적형질의 유전자 관련성 분석에서 다중대립유전자(multiple alleles)가 존재하고, 부모정보를 모를 때에도 활용할 수 있는 Allison의 혼합효과모형과 Fulker의 최대우도분산성분방법을 비교한다. 이들은 또한 인구집단 층화나 혼합성으로 인해 발생할 수 있는 의사(spurious)관련성을 보정할 수 있는 방법들이다.

Allison 방법은 각 가족간의 형제들을 임의효과로, 표식자의 유전형을 고정효과로 지정하고 양적형질을 종속변수로 하는 혼합모형을 이용하여 관련성 분석을 수행하는데, 공변량의 효과도 동시에 고려할 수 있다. Fulker 방법은 형제쌍간 평균과 형제쌍내 평균으로 표식유전자의 효과를 분할하여 관련성을 분석한다.

위 두 가지 방법을 일 병원 심장혈관유전체연구센터의 실제자료에 적용한 결과 두 방법에서 모두 표식자와 형질간에 관련이 없다는 결과를 나타냈다. 모의실험을 통하여 두 방법의 타당성을 비교한 결과, Allison 방법은 Fulker에 비해 형제가 2명, 3명, 5명인 경우에 모두 검정력이 높았고, 형제수가 증가할수록 검정력도 현저히 증가함을 보였다. 제1종 오류율은 Allison 방법이 Fulker에 비해 대체적으로 낮았으나 형제수에 상관없이 제1종 오류율은 큰 차이가 없는 것으로 나타났다.

핵심되는 말 : 양적형질, 인구집단층화, 형제자료, 관련성검정, Allison 방법, Fulker 방법

제 1 장 서 론

1.1 연구배경

유전학에서 인간에게 나타나는 형질은 크게 질적형질(qualitative trait)과 양적형질(quantitative trait)로 구분되어 진다. 질적형질은 질병이 있고 없음에 의한 이분형 또는 범주형으로 구분된다. 이처럼 질병을 연속적인 수치가 아닌 범주로 나타내는 형질을 질적형질이라고 한다. 반면 양적형질은 혈압, 지능, 몸무게 등과 같이 연속적인 수치로 나타내는 형질을 일컫는다. 질적형질에 관한 유전적연구는 임상적 진단의 중요성 때문에 강조되어 왔고 양적형질에 관한 연구보다 먼저 시작되었다. 그러나 알코올 중독, 우울증, 당뇨병, 비만, 고혈압 등과 같은 임상적 관심분야에서 대부분의 형질은 양적형질이기 때문에 질적형질에 비해 더 많은 정보를 갖고 있는 양적형질을 이용한 유전적 연구로 집중되고 있다.

인간에게 있어서 양적형질에 영향을 주는 유전자(gene)의 위치를 확인하는 것은 도전적인 일로 남아있다. 연관성(linkage)에 관한 연구는 많은 경우에 실질적인 검정력 문제를 갖고 있다(Blackwelder and Elston 1982; Risch and Merikangas 1996; Allison and Schork 1997; Collins et al. 1997). 표식유전자가(marker locus)가 형질유전자(trait locus)일 때, 또는 표식자유전자가 형질유전자에 연관되어 있고(linked), 관련되어있는(associated) 경우, 관련성(association)에 관한 연구는 실질적으로 더 검정력이 높지만 인구집단 층화(population stratification) 또는 인구집단 혼합성(population admixture)로 인한 중첩(confounding)에 영향을 받는다(Ewens and Spielman 1995).

현재 환자-대조군연구와 같은 관련성에 관한 연구는 유전적으로 복합형질(complex trait)의 분석에서 중요한 역할을 하고 있다. 환자-대조군연구가 갖는 문제점은 대조군의 선택에 있다. 비적절한 대조군이 선택되었을 때 위(僞)양성(false positive)으로 인한 관련성이 나타나는 문제가 있다. 예를 들면, 대조군이 비적절하

게 선택되었을 때 다른 하위그룹에 대한 인구집단 층화는 의사(spurious)관련성을 야기시킬 수 있다. 이러한 점을 보완하여 대조군과 환자군의 표본이 잘 배합되는 것을 보장하는 Spielman et al.에 의해 개발된 TDT(transmission/disequilibrium test)는 가족자료를 이용한(family-based) 질적형질에 관한 검정력이 좋은 관련성 검정(association test)이라고 할 수 있다. 이 방법은 부모에게서 질병이 있는 자녀에게로 전이된 표식대립유전자(marker allele)의 빈도와 전이되지 않은 표식대립유전자의 빈도를 χ^2 통계량을 이용하여 비교한다. 따라서 TDT는 질병이 없는 형제(sib) 또는 질병이 있는 여러 명의 가족구성원에 관한 자료를 필요로 하지 않고 관련성이 존재하는 경우 높은 검정력을 가지며, 부적절한 대조군의 존재로 인한 위(僞)양성도를 주지 않는다.

TDT는 여러 가지 방법으로 확장되어왔다(e.g., Curtis and Sham 1995; Sham and Curtis 1995; Morris et al. 1997). 그 중 주목할만한 두 가지 중에 첫째는 Allison(1997)이 양적형질을 사용하여 TDT를 확장한 방법과 Rabinowitz(1997)가 다중대립유전자(multiallelelic loci)를 사용할 수 있는 경우로 확장한 방법이고, 둘째는 최근 Spielman과 Ewens(1998), Curtis(1997), Boehnke와 Langefeld(1998)가 부모에 관한 정보를 필요로 하지 않는 관련성 검정방법이다. 이 방법은 형제자료(sibship) 중에 적어도 한 명의 자녀는 질병에 걸려있어야 하고 한 명은 질병에 걸리지 않아야 한다는 조건과 모든 형제들(sibling)은 동일한 유전형(genotype)을 갖지 않는다는 조건을 필요로 한다.

본 논문의 목적은 부모에 관한 정보를 이용할 수 없을 때 형제들 간에 양적형질을 적용할 수 있는 관련성 검정방법 두 가지를 소개하고 비교하는 것이다. 많은 복합형질은 부모에 관한 정보를 얻을 수 없는 경우가 많고, 또한 재정적인 한계나 실질적인 제약조건 등으로 발생하는 문제가 종종 있기 때문에 이러한 방법을 이용하는 것은 중요하다고 할 수 있다. 특히 고연령에 발생하는(late-onset) 질병형질에 대하여 연구될 때 중요하다고 할 수 있다. 예를 들면, 표현형(phenotype)이 고연력에 발생하는 질병을 가진 개인들간의 제지방체중(lean-body-mass; sarcopenia [Rosenberg 1997]) 손실률일 때 부모에 관한 적절한 표본을 조사하기

가 매우 어렵다.

1.2 연구목적 및 방법

최근 핵가족(nuclear family)에서 대가족(extended family)에 관한 자료를 다루는 관련성 분석방법이 많이 연구되고 있다. 그러나 실제로 부모에 관한 자료를 수집하기 어려운 질병에 관하여 관련성 연구를 수행할 때 형제자료를 사용하여 이러한 모형으로 분석할 수 없는 상황이 발생하거나 형제자료만 있어도 분석가능하더라도 모형자체가 형제자료만을 이용하는 방법보다 많은 정보를 필요로하기 때문에 그 정보를 알지 못한 채 분석을 수행하게 된다면 많은 편의(bias)를 갖는 결과를 나타낼 수 있다. 또한 인구집단 층화나 혼합성으로 인하여 실제로 관련성이 없는 경우에 관련성이 있다는 결과를 나타낼 수 있는 문제가 있다. 예를 들면, 수집된 자료가 병원자료인 경우에 대부분의 개체는 환자이기 때문에 자료의 분포가 왜곡될 우려가 매우 높다. 이런 경우 인구집단 층화나 혼합성에 관한 문제를 보정할 수 없는 방법을 이용하여 관련성 검정을 수행하게 된다면 잘못된 결과를 도출하게 될 것이다.

여기서는 부모에 관한 정보를 알 수 없고 인구집단 층화나 혼합성을 보정하여 의사관련성 결과를 배제하기 위한 두 가지 방법에 관하여 논의할 것이다. 앞으로 소개할 두 방법은 모두 위와 같은 문제에 잘 적합되는 방법이고 이 두 가지 방법에 관하여 비교된 논문은 현재까지 발표되지 않고 있다.

따라서 이 논문에서는 연관성이 존재하는 경우에 관하여 관련성 여부를 검정하는 것을 기본 가정으로 하여 부모에 관한 정보를 사용할 수 없을 때 형제자료만으로 표식자유전형(marker genotype)과 양적형질간의 관련성을 검정할 수 있는 혼합효과모형(mixed-effect model)을 이용한 Allison(1999) 방법과 최대우도분산성분방법을 기초로 한 Fulker(1999) 방법을 소개하고 일(一) 병원 심장혈관유전체연구센터에서 수집된 가계도 자료에 적용하여 실제 양적형질의 유전적 관련성 분석

을 수행하고, 모의실험을 통하여 이에 대한 통계적 타당성에 관하여 검증한다.

Allison 방법은 통계패키지 SAS v8.1을 이용하여 분석하고 Fulker 방법은 QTDT v2.4.2a 패키지(by Gonçalo Abecasis, http://www.sph.umich.edu/statgen/abecasis/QTDT/index.html)를 이용하여 분석한다. 모의실험 자료는 SOLAR v1.7.3(by Blangero et al., http://www.sfbr.org/sfbr/public/software/solar/index.html)의 서브루틴 중 하나인 SIMQTL을 이용하여 생성한다.

제 2 장 고전적인 관련성 분석방법

2.1 개요

부모의 정보를 이용하여 관련성을 검정하기 위한 방법은 Spielman 등(1993)에 의해 처음 시도되었다. Spielman 등의 방법은 가족자료를 이용하여 질적형질에 관한 관련성 검정을 위한 좋은 방법이지만 양적형질에 대해서는 검정할 수 없다. 이 장에서는 양적형질을 이용한 관련성 검정방법들 중 이 논문에서 비교하고자하는 Allison 방법과 Fulker 방법이 소개되기 전에 선행된 대표적인 방법 두 가지에 대하여 소개하고자 한다. 첫째는 Allison(1997)의 다섯 가지 TDT모형에 대하여 소개하고, 둘째는 Rabinowitz(1997)의 비모수적 TDT모형에 대하여 소개한다.

2.2 Allison이 제시한 TDT

2.2.1 개요

양적형질의 분석에 대한 TDT 디자인의 첫 번째 적용은 Allison에 의해 시작되었다. Allison은 부모와 한 명의 자녀로 구성된 핵가족자료를 이용한 TDT_{Q1}-TDT_{Q5}의 다섯 가지 종류의 TDT방법에 대하여 소개하였다. 다섯 가지 검정법은 표식자가 두 개의 대립유전자를 갖는다는 것을 가정한다. 그러나 두 개 이상의 대립유전자를 갖는 경우에는 관심있는 대립유전자와 그렇지 않은 대립유전자로 나누어 검정할 수 있다. 임의교배, 하디-와인버그 균형(Hardy-Weinberg equilibrium, HWE)의 성립, 표식자가 곧 양적형질유전자라는 선행가정이 필요하다.

처음 네 가지 검정($TDT_{Q1} - TDT_{Q4}$)은 부모 중 한명이 이질적 대립유전자(heterozygous, 혹은 이형대립유전자)를 갖고 다른 한명은 동질적 대립유전자(homozygous, 혹은 동형대립유전자)를 가져야 한다는 조건이 필요하다. 이 조건을 만족하지 않는 경우에는 관찰치간의 독립성이 문제되어 검정의 타당성이 떨어지게 되고, 자료가 불완전한 경우 편의된 결과를 얻을 우려가 있다. 그럼, 이제 다섯 가지 형태의 TDT 검정방법에 대하여 소개한다.

2.2.2 TDT_{Q1}: 랜덤표본추출에 의한 t-검정

이 방법에서는 다음 일곱 가지의 가정을 필요로 한다.

1. 표본크기가 충분히 크다면 CLT(중심극한정리, central limit theorem)에 의해 정규성가정을 따른다.
2. 부모와 한명의 자녀로 구성된 핵가족에서 부모 중 한명은 이형대립유전자를 갖고, 다른 한 명은 동형대립유전자를 갖는다.
3. 표식자가 곧 형질유전자이다.
4. 부모와 자녀의 유전형(genotype)과 표현형(phenotype)은 오류가 없다고 알려져 있다.
5. 표식자는 두 개의 대립유전자를 갖는다.
6. 임의교배이다.
7. 하디-와인버그 균형이 성립한다.

두 개의 대립유전자 중에 이형대립유전자를 갖는 부모에게서 표현형에 영향을 줄 것이라고 기대되는 대립유전자를 전이(transmission)받은 경우와 그렇지 않은 경우로 나눈 후 두 군에 대하여 t-검정을 수행한다.

세 가지 유전형 AA, Aa, aa에 대한 표현형(Y) 평균을 μ_{AA}, μ_{Aa}, μ_{aa}, Y의

분산을 σ_e^2, 표현형에 영향을 줄 것이라고 기대되는 대립유전자 a의 빈도를 p라고 하면,

$$\mu_{Y1} = E(Y|\ T=1 \cap C) = \frac{(1-p)^2\mu_{\mathrm{AA}} + p^2\mu_{\mathrm{aa}}}{(1-p)^2 + p^2}$$

$$\mu_{Y0} = E(Y|\ T=0 \cap C) = \frac{(1-p)^2\mu_{\mathrm{AA}} + p^2\mu_{\mathrm{Aa}}}{(1-p)^2 + p^2}$$

$$\begin{aligned}\sigma_{y1}^2 &= Var(Y|\ T=1 \cap C) \\ &= \sigma_e^2 + \frac{(1-p)^2(\mu_{\mathrm{Aa}} - \mu_{y1})^2 + p^2(\mu_{\mathrm{aa}} - \mu_{y1})^2}{(1-p)^2 + p^2}\end{aligned}$$

$$\begin{aligned}\sigma_{y1}^2 &= Var(Y|\ T=1 \cap C) \\ &= \sigma_e^2 + \frac{(1-p)^2(\mu_{\mathrm{Aa}} - \mu_{y1})^2 + p^2(\mu_{\mathrm{aa}} - \mu_{y1})^2}{(1-p)^2 + p^2}\end{aligned}$$

$$\begin{aligned}\sigma_{y0}^2 &= Var(Y|\ T=0 \cap C) \\ &= \sigma_e^2 + \frac{p^2(\mu_{\mathrm{Aa}} - \mu_{y0})^2 + (1-p)^2(\mu_{\mathrm{AA}} - \mu_{y0})^2}{(1-p)^2 + p^2}\end{aligned}$$

가 된다. 여기서 C는 위에 설명한 두 번째 가정사항에 관한 조건을 의미한다. N은 정보를 줄 수 있는 두 번째 가정사항을 만족하는 부모-자식 트리오의 수라고 하면 $N = N_S P(C)$가 되고 이 때, N_S는 전체 부모-자식 트리오의 수, $P(C)$는 두 번째 가정사항을 만족하는 부모-자식 트리오가 랜덤하게 선택될 확률이다. 이 경우,

$$P(C) = P_{Hh} = 2\,[2p\,(1-p)][1-2p\,(1-p)]$$

이 되고, 여기서 Hh는 오직 한 명의 부모에 관한 표식자유전형이 이형대립유전자인 사건을 의미한다.

귀무가설 $\mu_{y1} = \mu_{y0}$를 검정하기 위한 검정통계량은

$$t_{nc} = \frac{\mu_{y1} - \mu_{y0}}{\sqrt{\frac{2\,(\sigma_{y1}^2 + \sigma_{y0}^2)}{N}}}$$

이 된다.

2.2.3 $\mathrm{TDT}_{\mathrm{Q2}}$: 극단값을 갖는 표본추출에 의한 χ^2-검정

이 방법에서는 다음 두 가지를 제외하고는 $\mathrm{TDT}_{\mathrm{Q1}}$의 가정사항을 동일하게 만족하여야 한다.

1. 표본이 충분히 클 필요가 없다.
2. 각 유전형내에서 잔차는 정규분포를 따른다.

전체 N_S개의 가족에서 자녀의 표현형이 극단값을 갖는 경우, 즉, $Y > Z_{U(\mathrm{upper})}$, $Y < Z_{L(\mathrm{lower})}$인 가족만을 표본추출하고 이렇게 추출된 가족을 [표 1]과 같이 관심있는 대립유전자의 전이여부로 구분한다. 그런 다음, 자유도가 1인 비중심 χ^2-검정을 수행한다. 검정통계량은

$$\chi^2_{nc} = \frac{N(\gamma_1\gamma_4 - \gamma 2\gamma_3)^2}{(\gamma_1+\gamma_2)(\gamma_3+\gamma_4)(\gamma_1+\gamma_3)(\gamma_2+\gamma_4)}$$

이고, 여기서

$$\begin{aligned} N &= \gamma_1+\gamma_2+\gamma_3+\gamma_4) \\ &= N_S[P(L\cap C\cap T=0)+P(L\cap C\cap T=1) \\ &\quad + P(U\cap C\cap T=0)+P(U\cap C\cap T=1)] \end{aligned}$$

이 된다.

표 1. $\mathrm{TDT_{Q2}}$의 2×2 표

	전이상태	
표현형	0	1
L	$\gamma_1 = P(L\cap C\cap T=0)N_S$	$\gamma_2 = P(L\cap C\cap T=1)N_S$
U	$\gamma_3 = P(U\cap C\cap T=0)N_S$	$\gamma_4 = P(U\cap C\cap T=1)N_S$

* $\gamma_1, \gamma_2, \gamma_3, \gamma_4$: 셀 빈도

2.2.4 $\mathrm{TDT_{Q3}}$: 극단값을 갖는 표본추출에 의한 t-검정

CLT에 의해 정규성을 만족할 수 있는 정도의 표본크기여야하고 $\mathrm{TDT_{Q1}}$과 $\mathrm{TDT_{Q2}}$의 가정사항을 동일하게 만족하여야 한다. 이 방법은 특정 경계값(cutoff value) Z_U이상, Z_L이하의 값을 갖는 자녀들의 표현형을 표본추출하여 사용한다. $Z_L = Z_U$인 경우에는 $\mathrm{TDT_{Q1}}$과 동등한 검정방법이 된다. 여기서 조건 C는 부모

중 한명은 이형대립유전자를 갖고, 다른 한명은 동형대립유전자를 갖는다는 것과 위에 설명한 표현형이 극단값을 갖는 자녀들을 표본추출하는 것을 의미한다. 그러면,

$$\begin{aligned}\mu_{Y1} &= E(\,Y|\;T=1\,)\\ &= \frac{\begin{array}{r}P(\mathrm{Aa}|\;C\cap T=1\,)E(\,Y|\,\mathrm{Aa}\cap C)\\ +\,P(\mathrm{aa}|\;C\cap T=1\,)E(\,Y|\,\mathrm{aa}\cap C)\end{array}}{P(\mathrm{Aa}|\;C\cap T=1\,)+P(\mathrm{aa}|\;C\cap T=1\,)}\end{aligned}$$

이 되고, $P(\mathrm{Aa}|\;C\cap T=1\,)+P(\mathrm{aa}|\;C\cap T=1\,)\equiv 1$이 되기 때문에, 결국

$$\begin{aligned}\mu_{Y1} &= P(\mathrm{Aa}|\;C\cap T=1\,)E(\,Y|\,\mathrm{Aa}\cap C)\\ &\quad +P(\mathrm{aa}|\;C\cap T=1\,)E(\,Y|\,\mathrm{aa}\cap C)\end{aligned}$$

이 된다. 마찬가지로,

$$\begin{aligned}\mu_{Y1} &= E(\,Y|\;T=0\,)\\ &= P(\mathrm{AA}|\;C\cap T=0\,)E(\,Y|\,\mathrm{AA}\cap C)\\ &\quad +P(\mathrm{Aa}|\;C\cap T=0\,)E(\,Y|\,\mathrm{Aa}\cap C)\end{aligned}$$

이 되고, 분산은

$$\begin{aligned}\sigma^2_{Y1} &= Var(\,Y|\;T=1\,)\\ &= P(\mathrm{Aa}|\;C\cap T=1\,)\{\mathrm{Var}\,(\,Y|\,\mathrm{Aa}\cap C)+[E(\,Y|\,\mathrm{Aa}\cap C)-\mu_{y1}]^2\}\\ &\quad +P(\mathrm{aa}|\;C\cap T=1\,)\{\mathrm{Var}\,(\,Y|\,\mathrm{aa}\cap C)+[E(\,Y|\;\mathrm{aa}\cap\,C)-\mu_{y1}]^2\}\end{aligned}$$

$$\sigma^2_{Y0} = Var(Y|\ T=0)$$
$$= P(\mathrm{Aa}|\ C\cap T=0)\{\mathrm{Var}(Y|\ \mathrm{Aa}\cap C) + [E(Y|\ \mathrm{Aa}\cap C) - \mu_{y0}]^2\}$$
$$+ P(\mathrm{AA}|\ C\cap T=0)\{\mathrm{Var}(Y|\ \mathrm{AA}\cap C) + [E(Y|\ \mathrm{AA}\cap C) - \mu_{y0}]^2\}$$

이 된다. 여기서 $P(\mathrm{Aa}|\ C\cap T=1)$, $P(\mathrm{Aa}|\ C\cap T=0)$, $P(\mathrm{AA}|\ C\cap T=0)$, $P(\mathrm{aa}|\ C\cap T=1)$을 유도하기 위한 과정은 Allison(1997)의 논문에 잘 설명되어 있다. 따라서 검정통계량은

$$t_{nc} = \frac{\mu_{y1} - \mu_{y0}}{\sqrt{\frac{\sigma^2_{y1}}{P(T=1|\ C)N} + \frac{\sigma^2_{y0}}{[1-p(T=1|\ C)]N}}}$$

이 된다. N은 표본추출된 부모-자식 트리오의 수이고 $P(T=1|\ C)$는 표본추출된 자료에서 이형대립유전자인 부모가 자녀에게 관심있는 대립유전자를 전이시킬 확률이다.

2.2.5 $\mathrm{TDT}_{\mathrm{Q4}}$: $P(T=1|\ Y>Z_U) = P(T=1|\ Y<Z_L) = 1/2$의 검정

이 TDT방법도 극단값을 갖는 표본을 추출하여 사용한다. 그러나 여기서는 전이상태 여부에 관한 변수를 이용하는 것이 아니라 자녀의 표현형을 그대로 이용하고 양극단값을 갖는 자녀가 이형대립유전자를 갖는 부모에게서 관심있는 대립유전자를 전이받을 확률이 0.5인가를 검정한다.

$$P(T=1|\ Y<Z_L\cap C)$$
$$= \frac{P(T=1\cap L\cap C)}{P(T=1\cap L\cap\ ) + P(T=0\cap L\cap C)}$$

$$P(T=1 \mid Y < Z_L \cap C)$$
$$= \frac{P(T=1 \cap L \cap C)}{P(T=1 \cap L \cap) + P(T=0 \cap L \cap C)}$$

이고, 분산은

$$\sigma^2_{P(T=1 \mid Y > Z_U \cap C)}$$
$$= \frac{P(T=1 \mid Y > Z_U \cap C)[1 - P(T=1 \mid Y > Z_U \cap C)]}{N\left[\frac{P(Y > Z_U \cap C)}{P(Y > Z_U \cap C) + P(Y < Z_L \cap C)}\right]}$$

$$\sigma^2_{P(T=1 \mid Y < Z_L \cap C)}$$
$$= \frac{P(T=1 \mid Y < Z_L \cap C)[1 - P(T=1 \mid Y < Z_L \cap C)]}{N\left[\frac{P(Y < Z_L \cap C)}{P(Y > Z_U \cap C) + P(Y < Z_L \cap C)}\right]}$$

이 된다. 마찬가지로 N은 표본추출된 부모-자식 트리오의 수가 된다. 검정통계량은 $P(T=1 \mid Y > Z_U \cap C)$, $P(T=1 \mid Y < Z_L \cap C)$의 분산추정값에 따라 다른 가중치를 갖는 D를 이용하여 계산된다. D는 확률이 $1/2$이 된다는 기대값으로부터의 편차를 의미하고,

$$D = \frac{\begin{array}{c} P(T=1 \mid Y > Z_U \cap C)\frac{1}{\sigma^2_{P(T=1 \mid Y > Z_U \cap C)}} \\ + [1 - P(T=1 \mid Y < Z_L \cap C)]\frac{1}{\sigma^2_{P(T=1 \mid Y < Z_L \cap C)}} \end{array}}{\frac{1}{\sigma^2_{P(T=1 \mid Y > Z_U \cap C)}} + \frac{1}{\sigma^2_{P(T=1 \mid Y < Z_L \cap C)}}}$$

또는

$$D = \frac{\begin{array}{c} P(T=1 \mid Y > Z_U \cap C)\sigma^2_{P(T=1 \mid Y < Z_L \cap C)} \\ + [1 - P(T=1 \mid Y < Z_L \cap C)]\sigma^2_{P(T=1 \mid Y > Z_L \cap C)} \end{array}}{\frac{1}{\sigma^2_{P(T=1 \mid Y > Z_U \cap C)}} + \frac{1}{\sigma^2_{P(T=1 \mid Y < Z_L \cap C)}}}$$

이 된다. D의 분산추정값은

$$\sigma^2_{\hat{D}} = \frac{\sigma^2_{P(T=1 \mid Y > Z_U \cap C)}\sigma^2_{P(T=1 \mid Y < Z_L \cap C)}}{\sigma^2_{P(T=1 \mid Y > Z_U \cap C)} + \sigma^2_{P(T=1 \mid Y < Z_L \cap C)}}$$

이고 따라서 검정통계량은

$$Z = \frac{\hat{D} - \frac{1}{2}}{\sqrt{\sigma^2_{\hat{D}}}}$$

이 된다.

2.2.6 $\mathrm{TDT}_{\mathrm{Q5}}$: $E(Y)$와 G_i의 독립성 검정

이 방법은 이형대립유전자를 갖는 모든 부모쌍을 이용한다. 표식자가 대립유전자 A와 a를 갖는다고 하면, 모든 가능한 부모쌍의 유전형은 $\mathrm{Aa} \times \mathrm{AA}$, $\mathrm{Aa} \times \mathrm{Aa}$, $\mathrm{Aa} \times \mathrm{aa}$이 되고 이러한 부모의 자녀 중 $Y > Z_U$, $Y < Z_L$인 양극단값을 갖는 자녀만이 분석에 포함된다. 부모의 유전형 쌍에 따라 세 개의 집단으로

구분하여 가변수로 이용하고, 가변수를 자녀의 표현형값에 회귀시키고 이때의 R^2를 R_1^2으로 놓는다. 또 모형에 예측변수로서 X와 X^2를 추가하여 얻은 완전 R^2를 R_2^2라고 하면, 이 두 가지 R^2값을 이용하여 표식자 유전자의 가법효과(additive effect)와 우성효과(dominance effect)를 동시에 검정하는 F-검정을 수행한다. 여기서 X는 관심있는 대립유전자 a의 개수이다. 이 방법은 부모의 유전형쌍에 관한 조건에서 인구집단혼합성에 의한 영향을 제거할 수 있다. 두 가지 R^2는 각각

$$R_2^2 = \frac{\sum_{i=1}^{3} P(G_i|\, C)[E(Y'|\, G_i \cap C)]^2}{\mathrm{Var}\,(Y|\, G_i \cap C)}$$

$$R_1^2 = \frac{\sum_{j=1}^{3} P(W_j|\, C)[E(Y|\, C \cap W_j)]^2}{\mathrm{Var}\,(Y|\, C)}$$

이고 여기서 Y'는 Y에서 전체평균을 뺀 값이고, W_j는 부모의 세 가지 유전형쌍, G_i는 세 가지의 교배형태를 갖는 부모에게서 태어난 자녀의 세 가지 유전형이 된다. $P(G_i|\, C)$와 $P(W_j|\, C)$, $\mathrm{Var}\,(Y|\, G_i \cap C)$에 관한 유도과정은 Allison(1997)의 논문에 설명되어 있으므로 여기서는 생략하기로 하면,

$$E(Y'|\, G_i \cap C) = E(Y|\, G_i \cap C) - \sum_{i=1}^{3} P(G_i|\, C) E(Y|\, G_i \cap C)$$

$$E(Y|\, C \cap W_j) = \sum_{i=1}^{3} P(G_i|\, C \cap W_j)[E(Y'|\, G_i \cap C)]$$

이 된다. 우성효과와 가법효과의 결합 유의성을 검정하기 위한 검정통계량은

$$F_2(2, n-5) = \frac{(R_2^2 - R_1^2)/2}{(1-R_2^2)/(n-5)}$$

가 된다.

2.3 Rabinowitz가 제시한 비모수적 TDT

2.3.1 개요

Rabinowitz(1997)가 제안한 TDT 방법은 앞서 설명한 Allison의 TDT_{Q5}와 유사하다. 그러나 Rabinowitz의 방법은 형질의 분포에 관하여 가정하지 않는다는 점에서 비모수적 방법이라고 말할 수 있고, Allison의 TDT_{Q5}는 표식자가 두 개의 대립유전자를 갖는 경우로 제한되는 반면, 표식자가 다중대립유전자(multiallelic loci)를 갖는 경우에도 적용가능하다는 점에서 다르다고 할 수 있다. 또한 인구집단 층화나 혼합성에 의한 의사관련성을 보정할 수 있고 환경적 요인과 같은 공변량이 분석에 포함될 수 있다.

2.3.2 검정통계량

이 접근방법은 표식대립유전자와 형질간의 관련성에 관한 통계량을 구하는데 있다. 그런 다음 인구집단 혼합성으로 발생하는 의사관련성의 가능성을 배제하기 위해 부모정보를 이용하여 통계량을 수정한다. 먼저, 표식대립유전자와 양적형질간의 관련성에 대한 통계량을 소개하고 이 통계량을 수정하기위해 부모정보를 사용하는 방법을 소개한다. 그런 다음 표식자가 두개 이상의 대립유전자를 가졌을 때

수정된 통계량을 결합하기 위한 두 가지 방법에 대하여 소개한다. 마지막으로, 분석에 환경적·인구학적 공변량을 통합하기 위한 접근법에 대하여 설명한다.

이용 가능한 많은 핵가족이 있다고 가정한다. n은 가족수이고, i는 이들 가족에 관한 지표(index)이다. 가족은 각기 다른 자녀수를 갖는다. m_i는 i번째 가족의 자녀수이고, j는 자녀들에 대한 지표이다. 자녀의 양적형질 값은 Q, $Q_{i,j}$는 i번째 가족의 j번째 자녀에 대한 형질값이다.

이 방법은 대립유전자가 두 개 이상인 표식자에 적용할 수 있다. 표식자의 대립유전자 수는 k로 놓는다. 그러나 검정통계량은 한번에 하나의 주어진 표식대립유전자에 대해서 계산된다. 만약 대립유전자가 전이되었다면 지표변수(indicator variable)는 1값을 갖고, 전이되지 않았다면 0값을 갖는다. 지표변수를 Y라고 하면, $Y_{i,j,M}$은 i번째 가족의 j번째 자녀가 어머니쪽에서 대립유전자가 전이되었을 경우 1이고, 그렇지 않다면 0이 되고, 마찬가지로 $Y_{i,j,P}$는 대립유전자가 아버지에게서 전이되었을 경우 1이고, 그렇지 않은 경우 0이 된다.

양적형질과 주어진 대립유전자간의 관련성을 나타나는 통계량은 가족내에서의 모든 자녀들과 모든 가족들에 대한 합의 형태가 된다.

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Q_{i,j}-\overline{Q})(Y_{i,j,P}+Y_{i,j,M}).$$

여기서, $\overline{Q}$는 모든 가족내에서 모든 자녀들에 대한 양적형질의 평균이다. 통계량은 양적형질에 대한 지표변수의 로지스틱 회귀(logistic regression)의 점수통계량(score statistic)으로 생각할 수 있거나, 지표변수에 대한 양적형질의 회귀에서의 점수통계량으로 볼 수 있다.

이 통계량은 양적형질과 주어진 대립유전자의 존재에 대한 관련성을 반영한다. 만약 관련성이 존재하지 않는다면, 지표(indicator)는 양적형질에 대해 독립적이고, 이 통계량의 기대값은 0이 된다. 그러나 대립유전자가 형질과 관련되었다면, 형질

과 지표변수간에는 0이 아닌 공분산을 갖게 되고, 통계량의 기대값도 0이 아니다. 따라서 대립유전자와 형질간의 관련성은 0이 아닌 통계량값에 의해 암시된다.

인구집단 혼합성으로 인한 의사관련성이 나타나는 것을 방지하기 위해 부모정보를 이용하여 통계량을 수정한다. 수정은 두 가지 부분으로 이루어진다. 첫 번째 부분은 통계량에서 대립유전자에 대해 이형대립유전자가 아닌 부모와 관련되어있는 모든 지표변수를 제거하는 것이다. 두 번째 부분은 남아있는 지표변수에서 $1/2$을 빼는 것이다. $Y^*_{i,M}$, $Y^*_{i,P}$는 부모가 관심있는 대립유전자에 대해 이형인지에 관한 지표변수라고 하면, i번째 가족에서 어머니가 관심있는 대립유전자에 관하여 이형대립유전자를 갖는다면 $Y^*_{i,M}$은 1이 되고, 아버지가 관심있는 대립유전자에 대해 이형대립유전자를 갖는다면 $Y^*_{i,P}$가 1이 된다. 수정된 통계량은

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(Q_{i,j}-\overline{Q}\right)\left[Y^*_{i,M}\left(Y_{i,j,M}-\frac{1}{2}\right)+Y^*_{i,P}\left(Y_{i,j,P}-\frac{1}{2}\right)\right].$$

이 된다.

수정된 통계량은 인구집단 혼합성으로 인한 관련성에 영향을 받지 않는다. 비록 형질이 인구집단 혼합성을 통해 대립유전자와 관련되었을지라도, 형질유전자가 표식자에 연관되어있지 않는 한, 이형대립유전자를 갖는 부모의 대립유전자의 전이는 양적형질과는 독립적으로 $1/2$의 확률을 가지고 발생한다. 따라서, 연관성의 부재에서 형질과 전이상태를 나타내는 지표는 부모의 대립유전자 하에서, 조건부로 상호관련이 없게 되고, 통계량의 기대값은 0이 된다. 그러나 연관성의 존재에서는, 형질과 전이상태의 지표가 상호관련이 있고, 형질은 0이 아닌 수정된 통계량값에 의해 암시된다.

통계적 추론을 위해 수정된 통계량을 사용하려면, 연관성이 없다는 귀무가설하에서 분산이 필요하다. 수정된 통계량은 아래와 같은 항들의 합이 된다.

$$\left(Q_{i,j}-\overline{Q}\right)Y_{i,M}^{*}\left(Y_{i,j,M}-\frac{1}{2}\right) \text{ or}$$

$$\left(Q_{i,j}-\overline{Q}\right)Y_{i,P}^{*}\left(Y_{i,j,P}-\frac{1}{2}\right).$$

연관성이 없다는 귀무가설 하에서, 조건부로, 형질값과 부모의 대립유전자가 주어졌을 때, 검정통계량의 분산에 대한 각 항의 기여도는

$$\frac{1}{4}\left(Q_{i,j}-\overline{Q}\right)^{2}Y_{i,M}^{*} \quad \text{또는} \quad \frac{1}{4}\left(Q_{i,j}-\overline{Q}\right)^{2}Y_{i,P}^{*}.$$

이 된다.

가족내의 다중 형제들에 대해서도 분산이 고려될 필요가 있다. 공분산은 연관성이 없다는 귀무가설하에서 조건부로 부모의 대립유전자가 주어졌을 때, 한 자녀로의 전이는 한 형제자료(sibling)에게 전이되는 것과 독립이기 때문에 고려될 필요가 없다. 따라서 통계량의 조건부 분산은 수정된 통계량에서 나타난 배수에 의해 가중된 각 항 $Q_{i,j}-\overline{Q}$의 제곱합의 4분의 1이 된다. 각 항은 부모의 대립유전자가 0, 1, 2개 이형대립유전자인지에 따라 0, 1, 2의 값으로 가중치가 부여된다. 이런 계산법은 양적형질의 분포에 대해 모수적 가정에 의존하지 않고, 또한 인구집단 혼합성이 없거나 표식자에서의 하디-와인버그 평형과 같은 가정을 필요로 하지 않는다.

분산에 제곱근을 취해 정규화시키면 수정된 검정통계량은 표식자가 두 개의 대립유전자를 가질 때 연관성을 검정하는 t-통계량으로 사용될 수 있다. 그러나 수정된 통계량은 다형표식자(polymorphic marker)에 대해 획일적인 검정절차로 결합되어야만 한다. 이분형형질(또는 질적형질)에 대한 관련성 방법에서, 다형표식자를 이용한 두 접근법이 사용된다. 첫 번째 접근법에서 열(rows)은 대립유전자, 행(coulmn)은 질병유무, entry는 염색체수인 $2\times k$ 분할표를 만드는 것이다. 그런 다음 일반적인 독립성 검정에 대한 관련성을 검정하기 위해 χ^{2}-검정이 사용된다.

두 번째 접근방법은 각 대립유전자에 대해, 열이 주어진 대립유전자의 존재 또는 부재를 나타내고, 행이 질병유무, entry는 염색체수인 붕괴(collapsed) 2×2 분할표를 만드는 것이다. 일반적인 독립성 검정에 대한 χ^2-검정은 각 k개의 분할표에 대해 적용되고 최대 χ^2 통계량이 추론에 사용된다.

두 개의 다른 대립유전자에 대한 수정된 검정통계량에서 l번째 대립유전자 각각에 대한 수정된 검정통계량을 $T_l(l=1,2,\cdots,k)$이라고 하자. 검정통계량간의 공분산은 검정통계량 각각에 대한 분산과 유사하다. 공분산은

$$\frac{1}{4}\left(Q_{i,j}-\overline{Q}\right)^2$$

이 된다. 그러나 공분산에 대한 가중치는 분산에 대한 가중치와는 다르다. 대립유전자쌍과 관련된 수정된 통계량들간의 공분산에서, 가중치는 두 개의 대립유전자가 모두 이형인 부모 수를 뺀 것이 된다. 이것은 대립유전자 l_1과 l_2에 대한 수정된 검정통계량간 공분산에서의 가중치는 대립유전자 l_1, l_2 모두를 갖는 i번째 가족에서의 부모수가 0이면 0, 1이면 -1, 2이면 -2의 값을 갖게 됨을 의미한다.

이제 수정된 검정통계량을 결합하는 두 접근법을 제시한다. 이분형형질에서의 두 접근법 중 첫째는 개개의 통계량을 하나의 χ^2 통계량으로 결합하는 것이다.

$$S=\sum_{l_1=1}^{k-1}\sum_{l_2=1}^{k-1}T_{l_1}T_{l_2}\sigma^{-1}_{l_1,l_2}.$$

여기서, $\sigma^{-1}_{l_1,l_2}$은 처음 $k-1$개의 수정된 검정통계량의 분산-공분산 행렬의 역행렬에 대한 i,j번째 구성요소이다. k개의 통계량이 선형의존(linearly dependent)이기 때문에 오직 처음 $k-1$개의 수정된 통계량이 고려된다. 이분형형질에 대한 두 번째 방법은 각 대립유전자에 관련된 χ^2 통계량의 최대값인

$$S_{\max} = \max_{l=1}^{k} \frac{T_l^2}{\sigma_l^2}$$

이 검정통계량으로 사용되는 것이다.

표본이 충분히 크다면, 대립유전자 l의 전이와 양적형질간의 관련성이 없다는 귀무가설하에서, 통계량 S는 자유도 $k-1$인 χ^2 분포를 따른다. 대립가설하에서, 통계량은 비중심(noncentral) χ^2_{k-1} 분포를 따른다. 결과에 대한 기본원리는 결합정규확률변수의 이차형태(quadratic forms)의 표준이론과 함께 수정된 통계량의 점근적(asymptotic) 결합정규성을 갖게 된다(Rao and Mitra[26, section 9.3]). 통계량 $S_{\max}$는 표준분포를 따르지 않기 때문에, p 값을 계산하기위해 시뮬레이션이 실행되어야 한다.

마지막으로, 환경적·인구학적 공변량이 분석에 포함되는 방법을 고려한다. 이 방법은 가능한 공변량에 대한 $Q_{i.j}$의 회귀로부터 $Q_{i,j}$에 대한 적합값(fitted value)으로 $\overline{Q}$를 대체하는 것이다. 따라서 통계량은

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Q_{i,j} - \hat{Q}_{i,j})\left[Y^*_{i,M}\left(Y_{i,j,M} - \frac{1}{2}\right) + Y^*_{i,P}\left(Y_{i,j,P} - \frac{1}{2}\right)\right]$$

이고, 여기서 $\hat{Q}_{i,j}$는 공변량에서 형질의 회귀로부터 j번째 가족에서 j번째 자녀의 형질에 대한 적합값이다. 회귀에서 잔차를 가진 $Q_{i,j} - \overline{Q}$ 항을 대체하는 것은 회귀모형에 의해 설명되는 형질에서 변동(variability)의 성분을 제거하는 것이다.

제 3 장 부모정보가 없는 경우의 관련성 분석방법

3.1 Allison 방법

3.1.1 개요

표현형과 표식자유전형(marker-locus genotype)간에 관련성(association)이 존재한다면, 이 관련성은 논리적으로 다음과 같은 이유 때문일 것이다.

1. 표현형이 표식자 유전형에서 변화(variation)를 일으킨다.
2. 표식자 유전형이 표현형에서 변화를 일으킨다.
3. 표식자의 유전자 위치는 물리적으로 연접해있고(linked) 또한 표현형에서 변화를 일으키는 또 다른 유전자 위치와 관련되어있다.
4. 표식자의 대립유전자(alleles)는 표현형에서 변화를 일으키는 어떤 다른 유전적인 요인과 관련이 있고, 물리적으로는 연접해있지 않다.

오직 2, 3번과 같은 관련성만이 형질에 영향을 주는 유전자의 위치화(localization)에서 유용하기 때문에 목표는 1, 4를 제외한 관련성 검정을 수행하는 것이다. 1번은 표현형을 조사하기 전에 유전형을 조사하는 것은 논리적으로 불가능하고, 원인은 결과에 앞서야한다는 인과관계의 기본적인 공리(axiom)에 의해 제외된다(Hume 17XX[1988]). Allison(1997)은 표식유전자에서 부모의 유전형에 관한 조건을 설정하는 것이 표식유전자의 유전형과 자손의 표현형간에 관찰된 관련성에 대한 가능한 설명으로서 4번이 제거되기에 충분하다는 것을 지적하였다. 이것은 연관성(linkage)이 없을 때 표식유전자에 대한 부모의 유전형이 있다는 조건하에서, 독립적분류법칙(independent assortment of law)에 의하여(물론, 관찰된 의미있는 표본 관련성이 우연히 발생할 수도 있다.) 자손의 표식자유전형과 다른 유

전적인 요인들간에는 관련성이 존재하지 않기(no population association) 때문이다.

게다가, 형제자료가 완전형제(full-sibs)로 구성되어있을 때, 형제자료 내 형제들의 유전형에 대한 확률은 전적으로 그들의 부모 유전형에 의존한다. 부모의 유전형을 알지 못할 때 즉, 특정한 형제자료 구성원들의 효과에 대하여 제어할 때, 형제자료에 관한 조건은 부모의 유전형을 제어하는 것과 동등하다. 왜냐하면, 한 형제자료 내의 모든 형제들은 동일한 부모에게서 태어났으므로 부모의 유전형이 모두 같기 때문이다. 이것은 인구집단 층화(population stratification)에 의한 중첩(confounding)의 가능성을 제거한다. Curtis(1997), Boehnke와 Langefeld(1998), Spielman과 Ewens(1998)은 형제자료에 관한 연관성검정방법을 개발하기위하여, 그리고 특정 자료구조에서는 연관성과 관련성 검정을 하기 위하여 이러한 아이디어를 사용하였다.

여기서는 서로 독립적인 형제자료에 관한 표본으로 혼합효과모형(mixed-effects model)에 관한 검정방법을 평가한다. 혼합효과모형은 형제자료 내 유전형으로 인한 변동(variation)을 평가함으로서 인구집단 층화에 대하여 통제한다. 이것은 형제자료의 조건하에서 형제들이 임의적으로(randomly) 유전형을 할당받는 것을 의미한다. 그리고 둘 또는 그 이상의 유전자에서 유전형의 확률은 그들이 물리적으로 연접해있는 것에 의존하게 된다. 조건부 형제자료로 수행된 유전적 표식자(genetic marker)와 표현형간의 관련성을 검정하는 것은 타당한 연관성검정이 된다.

이러한 접근방법은 형제자료에 관한 조건없이 사용되는 가족자료를 포함하는 관련성 검정과는 기본적으로 다르다. 예를 들면, Trégouët 등(1997)은 몇몇 자료가 상호연관되어 있는 경우에 관련성에 대하여 분석하는 방법으로 GEE (generalized estimating equations)방법을 제안하였다. 이 방법은 인구집단 층화로 인해 발생하는 의사관련성을 보정하지 못한다. 이것은 George-Elston(1987)에 의해 개발되어 SAGE 패키지에서 실행가능한 프로시져(procedure)와 매우 유사하다.

3.1.2 혼합효과모형을 이용한 Allison 방법

일란성쌍둥이(monozygotic trin, MZ)가 아닌 둘 이상의 완전형제들(full sibling)로 구성된 형제자료가 J개라고 하자. j번째 형제자료의 형제 수는 K_j이고, n_{ij}는 i번째 유전형을 가진 j번째 형제자료에서의 형제 수를 나타낸다($\Sigma_i n_{ij} = K_j$). j번째 형제자료에서 i번째 유전형의 k번째 형제에 대한 표현형은 Y_{ijk}로 나타낸다. 표식자 M은 m개의 대립유전자를 갖는다고 가정하고, $M_1, M_2, \cdots, M_m$으로 표현하면, 전체 $\frac{m(m+1)}{2}$개의 유전형이 가능하다. 이런 상황에서, 유전형을 $\frac{m(m+1)}{2}$개의 수준을 갖는 고정효과(fixed factor) A라 하고, 형제자료를 J개의 수준을 갖는 임의효과(random factor) B라고 하자. 따라서 표현형에 대한 이요인 혼합효과모형(two-factor mixed-effects model)을 다음과 같이 쓸 수 있다(e.g., Burdick and Graybill 1992; Neter et al. 1996).

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \ , \tag{1}$$

$$\begin{cases} i = 1, \cdots, I\left[= \frac{m(m+1)}{2}\right] \\ j = 1, \cdots, J \\ k = 1, \cdots, n_{ij}(\geq 0); \Sigma_i \, n_{ij} = K_j \end{cases}$$

여기서, μ는 상수이고 α_i는 고정효과(유전형) A에 대한 주 효과(effect sizes), β_j는 임의효과(sibship) B에 대한 주 효과(effect sizes), 교호작용효과인 $(\alpha\beta)_{ij}$는 임의효과이다.

이 공식에서, 형제자료 효과는 랜덤하게 모형에 포함되기 때문에 표현형에 대한 유전형의 영향을 검정하는 것은 형제자료에 대하여 조건부가 된다. 그러나 대부분의 표본추출설계(sampling scheme)하에서, 각각의 유전형을 갖는 각 형제자료

내의 개체수가 모든 형제자료에 대하여 일정하지 않게 된다. 이것은 식 (1)에서 모형이 불균형(unbalanced)이라는 것을 의미하고 따라서 n_{ij}가 모두 같지 않음을 나타낸다. 혼합불균형모형(mixed unbalanced model)에서의 통계적 추론은 복잡하다(Burdick and Graybill 1992). 비록 분산성분을 추정하고, 그것의 통계적 유의성을 검정하기 위한 많은 접근방법들이 있지만, 여기서는 분산분석(ANOVA-based F-test)(Burdick and Graybill 1992, chap.6)에 의한 혼합효과모형을 검정하는 것으로 제한한다.

3.2 Fulker 방법

3.2.1 우도를 이용한 접근방법

원래 형제자료에 관한 최대-우도 모형을 포함한다. 자료에 대한 다음과 같은 우도에 자연로그를 취하여 최대화시킴으로써 완전형제자료 공분산구조를 모형화 한다.

$$L = \prod_{i=1}^{M} (2\pi)^{-k_i/2} |\Sigma_i|^{-1/2} e^{-1/2 |(y_i - \mu_i)\Sigma_i^{-1}(y_i - \mu_i)|} , \qquad (2)$$

k가 i번째 가족에서 측정된 변수의 개수일 때, Σ_i는 i번째 가족에서 형제들간의 기대공분산행렬(expected covariance matrix)이고, y_i는 i번째 가족에서 형제들에 관해 얻은 관찰된 점수벡터(a vector of observed scores)이며, μ_i는 i번째 가족에 대한 기대평균벡터, M은 가족수이다. 다변량정규성 자료의 우도에 관한 일반적인 표현에서 기대공분산행렬, 평균벡터의 원소들은 직접 추정될 수 있다. 또한 더 유용하게 이 원소들이 관심있는 이론적 모수에 관한 함수로 만들어질 수 있다. 이러한 이론적 모수들은 관심있는 모수에 관한 모형에 적합시킨 후 모형에 자연로그를 취하여 얻은 우도 $\ln(L_1)$와 특정 모수가 0인지에 관한 귀무가설 하에서 얻은 우도 $\ln(L_0)$에 의하여 통계적 유의성을 검정할 수 있다. 표본크기가 충분히 크다면, $2[\ln(L_1) - \ln(L_0)]$은 점근적으로 χ^2분포를 따르게 된다. 이 때, 자유도는 검정될 모수의 개수가 된다. 이러한 접근방법은 형제자료 또는 확장된 가족자료에서 얻은 양적형질에 관한 일반적인 모형이다.

어떤 염색체상에 위치한 특정한 위치에 대한 연관성 검정의 경우에 형제쌍들의 기대공분산행렬은

$$\sum_i = \begin{pmatrix} \sigma_q^2 + \sigma_c^2 + \sigma_e^2 & \hat{\pi}_i\sigma_q^2 + \sigma_c^2 \\ \hat{\pi}_i\sigma_q^2 + \sigma_c^2 & \sigma_q^2 + \sigma_c^2 + \sigma_e^2 \end{pmatrix}$$

가 된다. 여기서 σ_q^2은 추정상의 QTL(quantitative trait locus)이 설명하는 분산의 추정값, σ_c^2은 형제들간에 공유하고 있는 환경적 분산과 가법적다유전자분산(the additive polygenic variance)의 1/2을 포함하는 잔여형제유사분산(the residual sibling resemblance)의 추정값, σ_e^2은 가족 내에 있는 형제들간에 공유되지 않은 분산 추정값을 의미한다. 이 접근방법은 형제쌍들의 IBD(identical by descent)비율[1]의 추정값인 $\hat{\pi}$을 이용한다. Kruglyak과 Lander(1995)는 형제쌍이 0, 1, 2개의 대립유전자를 IBD할 확률을 직접 이용하는 접근법에 대하여 소개하였다. 이는 0, 1 또는 2개의 대립유전자를 공유할 때 각각의 가중치를 주어 계산된 우도를 합하여 계산된다. Fulker와 Cherny(1996)는 이러한 방법을 완전 형제쌍의 공분산구조에 이용하는 방법으로 확장하였다. 이와 같은 우도함수를 이용하는 것은 후보유전자 또는 형질유전자와 불균형상태(disequilibrium)일 것이라고 추측되는 유전자의 대립유전자 효과를 모형화하기위해 쉽게 확장될 수 있다.

3.2.2 최대우도분산성분방법을 이용한 Fulker 방법

추정상의 QTL이 대립유전자빈도가 각각 p와 q인 대립유전자 A_1과 A_2를 갖는다고 가정하면, 세 가지 유전형 A_2A_2, A_1A_2, A_1A_1에 관한 효과는 각각 $-a$, 0, a가 된다. 이 때 임의교배(random mating)의 가정하에서 형제쌍의 유전형 조합은 총 9가지가 되고 유전형 조합에 따른 형제쌍평균(sibpair means)과 형제쌍차

1) 형제쌍이 대립유전자를 0, 1, 2개 공유할 확률을 의미하며, Haseman과 Elston(1972)에 의해 처음 소개되었고, Kruglyak 등(1995)과 Kruglyak과 Lander(1995)에 의해 다중표식자인 경우로 확장되었다. IBD비율을 추정하는 방법은 가장 간단한 가표식자(pseudomarker)의 사후확률을 이용하는 방법에서 MCMC(Markov Chain Monte Carlo) 알고리듬을 이용한 방법까지 다양하게 확장되고 있다.

이(sibpair differences)에 의해 구분될 수 있다. 이는 [표 2]와 같다.

표 2. 대립유전자가 2개인 단일가법유전자에 관한 기대형제쌍평균과 차이(expected sib-pair means and differences)

유전형		가법적 효과			
형제 1	형제 2	형제 1	형제 2	평균	차이/2
A_1A_1	A_1A_1	a	a	a	0
A_1A_1	A_1A_2	a	0	$a/2$	$a/2$
A_1A_1	A_2A_2	a	$-a$	0	a
A_1A_2	A_1A_1	0	a	$a/2$	$-a/2$
A_1A_2	A_1A_2	0	0	0	0
A_1A_2	A_2A_2	0	$-a$	$-a/2$	$a/2$
A_2A_2	A_1A_1	$-a$	a	0	$-a$
A_2A_2	A_1A_2	$-a$	0	$-a/2$	$-a/2$
A_2A_2	A_2A_2	$-a$	$-a$	$-a$	0

형제쌍에 대하여, 식 (2)의 우도함수에서 총 평균(overall mean) m과 형제쌍평균 s_m, 형제쌍차이 s_d의 함수로 기대평균벡터를 모형화할 수 있다. [표 1]에 나타나 있는 형제쌍평균과 차이에 대한 기대값은

$$\mu_1 = m + s_m + (s_d/2)$$
$$\mu_2 = m + s_m - (s_d/2)$$

이 된다. 그런 다음 자유도가 1인 χ^2 검정으로 a에 대한 관련성을 검정할 수 있다. 그러나 이 검정법은 인구집단 층화로 인한 의사관련성을 보정하지 못한다. 인구집단 층화는 형제쌍평균에 영향을 주지만 형제쌍차이에는 영향을 주지 않기 때문이다. 의사관련성을 보정하기 위한 방법은 형제쌍평균과 형제쌍차이에 대한 차

이가 되는 유전자효과(gene effect) a를 허용하는 것이다. 형제쌍평균의 유전자효과를 a_b라 하고 형제쌍차이의 유전자효과를 a_w라고 하면, 형제 1과 형제 2에 관한 모형은 아래 [표 3]과 같다.

표 3. 가법적 효과의 형제간(between-pairs), 형제내(within-pairs)성분 분할

유전형		가법적 효과			
형제 1	형제 2	형제 1	형제 2	평균	차이/2
A_1A_1	A_1A_1	a_b	a_b	a_b	0
A_1A_1	A_1A_2	$(a_b/2)+(a_w/2)$	$(a_b/2)-(a_w/2)$	$a_b/2$	$a_w/2$
A_1A_1	A_2A_2	a_w	$-a_w$	0	a_w
A_1A_2	A_1A_1	$(a_b/2)-(a_w/2)$	$(a_b/2)+(a_w/2)$	$a_b/2$	$-a_w/2$
A_1A_2	A_1A_2	0	0	0	0
A_1A_2	A_2A_2	$(-a_b/2)+(a_w/2)$	$(-a_b/2)-(a_w/2)$	$-a_b/2$	$a_w/2$
A_2A_2	A_1A_1	$-a_w$	a_w	0	$-a_w$
A_2A_2	A_1A_2	$(-a_b/2)-(a_w/2)$	$(-a_b/2)+(a_w/2)$	$-a_b/2$	$-a_w/2$
A_2A_2	A_2A_2	$-a_b$	$-a_b$	$-a_b$	0

[표 3]를 이용하여 식 (2)의 우도함수의 지수부분을 일반적인 형태로 나타내면,

$$-\frac{1}{2}(\boldsymbol{y_i}-\boldsymbol{m}-\boldsymbol{X_{b_i}a_b}-\boldsymbol{X_{w_i}a_w})\Sigma_i^{-1}(\boldsymbol{y_i}-\boldsymbol{m}-\boldsymbol{X_{b_i}a_b}-\boldsymbol{X_{w_i}a_w}) \quad (3)$$

이 된다. 여기서 $\boldsymbol{X_{b_i}}$와 $\boldsymbol{X_{w_i}}$는 [표 3]의 세 번째, 네 번째 열에 나타나 있는 계수를 이용하여 구한 i번째 형제쌍에 대한 대각행렬(diagonal matrices)이 되고, $\boldsymbol{a_b}$와 $\boldsymbol{a_w}$는 형제자료의 구성원 개개인에 대한 모수 a_b와 a_w를 각각 포함하는 벡터가 된다. $\boldsymbol{m}$은 하나의 형제자료에 대한 총 표현형평균벡터이고 이는 한 형제자료의 모든 형제들에 대하여 동일하다.

a_w를 0으로 놓은 모형과 그렇지 않은 모형간의 우도비는 자유도가 1인 χ^2 검정을 통하여 통계적 유의성에 관하여 검정할 수 있고 이는 인구집단층화를 제어하는 로버스트한 관련성 검정방법이 된다(Fulker et al. 1999). 우성유전자인 경우에 식 (3)은 쉽게 확장될 수 있고 다중대립유전자(multiple alleles)인 경우에 대립유전자의 수가 증가할수록 관련성 검정을 하기 위한 검정통계량의 계산 시 자유도가 극적으로 증가하기는 하지만 손쉽게 확장될 수 있다.

제 4 장 실제자료에 적용

일(一) 병원 심혈관계질환유전체연구센터(Cardiovascular Genome Center)에서 수집된 가계도 자료(CGC자료)를 이용하여 Allison 방법과 Fulker 방법을 통한 양적형질과 표식자간의 관련성 분석을 하였다. 전체 CGC자료 중에서 부모를 제외한 형제자료만을 이용하였고, 모두 정상인으로 구성된 형제자료를 이용하는 것은 관련성 검정에 의미가 없기 때문에 형제 중에 심혈관 질환이 있는 환자를 포함하는 형제자료만을 분석에 사용하였다. 총 구성원 수는 190명, 가족당 평균 구성원 수는 약 3명이고, 69개의 가족으로 구성되어 있다. [표 4]는 분석에 사용된 CGC자료의 분포를 보여준다.

표 4. 분석에 사용된 CGC자료의 분포

형제수	가족수	%
2	40	57.97%
3	13	18.84%
4	12	17.39%
5	3	4.35%
8	1	1.45%

관련성 분석의 대상이 되는 양적형질은 심장병, 뇌일혈, 동맥경화증 발생에 기여하는 것으로 알려진 LDL(low-density lipoprotein, 저밀도 지단백)값이다. 심혈관 질환을 일으키는 질병유전자는 여러 종류가 있는데 혈중지질대사에 관여하는 ApoE(apolipoprotein E)를 표식자로 사용하였다. ApoE는 19번 염색체에 위치하고, E2, E3, E4의 세 가지 대립유전자를 가지며 유전형은 E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, E4/E4로 6가지이다. ApoE는 LDL 콜레스테롤 수치를 상승시켜 심

혈관 질환을 일으키는 대표적인 유전자로 알려져 있다(Lahoz et al, 2001).

이 중에서 E4의 대립유전자를 갖고 있는 사람은 그렇지 않은 사람에 비해 동맥경화, 치매 등에 걸릴 위험이 높은 것으로 알려져 있다. ApoE의 각 대립형질에 따른 빈도는 인종에 따라 약간 다르지만, 대규모 Framingham 연구에 따르면 E2=0.079, E3=0.802, E4=0.119로 나타난다(Lahoz et al, 2001).

Allison 방법과 Fulker 방법을 적용하여 ApoE와 LDL간의 관련성 검정을 수행하는데 나이, 성별, 흡연유무, BMI(body mass index)를 공변량으로 포함하였다.

ApoE와 LDL의 관련성 분석을 위해 Allison과 Fulker 방법을 적용한 결과는 [표 5], [표 6]와 같다. 유의수준 0.05에서 ApoE 유전형에 따른 LDL의 차이에 관하여 Allison 방법에서는 p-value=0.1815, Fulker 방법에서는 p-value=0.2998로 두 방법의 결론은 모두 관련성이 없다고 나타났다. 단, Fulker 방법에서는 나이, 성별, 흡연유무, BMI의 공변량에 대하여 검정할 수 없었지만, SAS를 이용하여 분석가능한 Allison 방법에서는 각 공변량에 관한 유의성 검정결과도 알 수 있다. Allison 결과[표 5]에서 성별이나 흡연유무, BMI는 LDL에 영향을 주지 않는 것으로 나왔으나 나이는 p-value=0.0001로 LDL과 매우 높은 관련이 있음을 볼 수 있다.

표 5. Allison 방법 적용결과: SAS v8.1

	고정효과에 관한 Type 3 검정결과			
효과	분자자유도	분모자유도	F값	p-value
성별	1	169	0.03	0.8633
나이	1	123	15.49	0.0001
BMI	1	176	0.37	0.5449
흡연유무1	1	164	3.07	0.0815
흡연유무2	1	175	0.17	0.6811
ApoE유전형	4	69.7	1.61	0.1815

표 6. Fulker 방법 적용결과: QTDT

	귀무가설		대립가설			
	df(0)	Ln(0)	df(1)	Ln(1)	χ^2 값	p-value
ApoE유전형	195	823.81	193	822.61	2.41	0.2998

제 5 장 모의실험을 통한 비교

5.1 자료와 방법

Allison 방법과 Fulker 방법의 검정력과 제1종 오류율을 평가하기 위해 모의실험 자료를 생성하였다. 모의실험 자료는 하나의 양적형질과 단일표식자 자료로 QTL(quantitative trait locus, 양적형질유전자)과 표식자는 두 개의 대립유전자를 갖는다고 가정하였다. 표식자의 대립유전자는 M, m의 대립유전자를 갖는다고 하면, 각 대립유전자의 빈도는 0.3과 0.7, 총 유전율은 100%로 설정하였다. 양적형질은 평균이 0, 분산이 1인 정규분포를 따른다고 가정한다[표 7].

표 7. 모의실험자료생성에 관한 가정사항

가정사항	검정력	제1종 오류율
표식대립유전자	M, m	M, m
M의 빈도	0.3	0.3
m의 빈도	0.7	0.7
양적형질의 평균		
μ_{QQ}	1	0
μ_{Qq}	0	0
μ_{qq}	-1	0
θ	0	0
유전율	100%	100%

모의실험 자료는 QTL과 단일표식자 간의 재조합률(recombination fraction) θ 값에 따라 구별할 수 있다. 여기서는 완전 연관성을 갖는다는 것을 가정하여 $\theta = 0$으로 설정하여 모의실험자료를 생성하였다. 형제가 2명인 경우와 3명인 경

우, 그리고 5명인 경우 각각에 관하여 $n = 500$으로 총 형제수는 위의 세 가지 경우 모두 같도록 하였다. 세 가지 자료유형에 대하여 각각 100개의 반복표본을 생성하여 검정력과 제1종 오류율을 평가하였다[표 8].

검정력을 평가하기 위해 세 가지 유형의 자료에 대한 100개의 반복표본은 $\theta = 0$이고 QTL의 유전형에 따른 양적형질의 평균은 $\{\mu_{qq}, \mu_{Qq}, \mu_{QQ}\} = \{-1, 0, 1\}$에서 생성되었다. 이 때 유의수준 $\alpha = 0.05$에서 '관련성이 없다'는 귀무가설을 기각하게 되는 비율을 검정력으로 정의한다.

표 8. 모의실험자료의 구성

자료유형	총 형제수	가족수	반복수
Ⅰ. 형제2명	500	250	100
Ⅱ. 형제3명	501	167	100
Ⅲ. 형제5명	500	100	100

제1종 오류율을 평가하기 위해 세 가지 유형의 자료는 $\theta = 0$, QTL의 유전형에 따른 양적형질의 평균은 $\mu_{qq} = \mu_{Qq} = \mu_{QQ} = 0$으로 차이가 없다는 가정하에서 세 가지 자료유형에 대한 100개의 반복표본을 생성하였다. 마찬가지로 제1종 오류율은 관련성이 없는 QTL과 표식자 자료에 두 방법을 적용했을 때, 유의수준 $\alpha = 0.05$에서 '관련성이 없다'는 귀무가설을 기각하게 되는 비율을 제1종 오류율로 정의한다.

세 가지 유형의 자료는 검정력과 제1종 오류율을 평가하기 위해 각각 100개의 반복표본으로 구성되었고, SOLAR의 SIMQTL로 생성되었다. SIMQTL은 SOLAR의 서브루틴 중 하나로, QTL과 표식자의 대립유전자 빈도와 QTL과 표식자간 재조합률, QTL의 유전형에 따른 양적형질의 평균이 주어졌을 때 QTL과 양적형질, 단일표식자를 모의생성한다.

Allison 방법을 이용한 관련성 검정은 통계패키지 SAS v8.1의 'mixed' 프로시져를 이용하여 수행하였다. 각 형제자료들을 구분할 수 있는 가족ID와 표식자의

유전형, QTL값이 분석모형에 포함되었고 QTL과 표식자간의 관련성을 평가하기 위해 모형에 포함된 모든 다른 효과에 대하여 조정하였을 때의 효과에 대한 제곱합을 계산한 제3종 제곱합이 계산되고 이에 대한 F값에 의한 p-value가 계산된다.

Fulker 방법을 이용한 관련성 검정은 QTDT(Quantitative Transmission Disequilibrium Test) 패키지에서 수행하였다. QTDT는 관련성을 검정하기 위한 방법들 중 Allison(1997)의 TDT_{Q5}, Rabinowitz(1997), Fulker et al.(1999), Monks et al.(2000), Abecasis et al.(2000)이 제시한 방법에 대한 관련성 분석을 수행할 수 있다. Fulker 방법을 이용하여 관련성 검정을 하기 위해서는 IBD비율에 대한 계산이 선행되어야 한다. QTDT에서는 GENEHUNTER2 또는 Simwalk2 프로그램을 이용하여 IBD비율을 추정하는데, GENEHUNTER2는 유닉스 운영체제에서 사용가능하며 Simwalk2는 유닉스, 리눅스, 윈도우즈에서도 사용가능하다. 여기서는 Simwalk2를 이용하여 IBD비율을 추정하였다. Simwalk2는 MCMC(Markov Chain Monte Carlo) 방법을 이용하여 IBD비율을 추정한다.

5.2 검정력

Allison 방법과 Fulker 방법의 검정력을 비교하기 위해 QTL과 단일표식자간의 $\theta = 0$, $\{\mu_{qq}, \mu_{Qq}, \mu_{QQ}\} = \{-1, 0, 1\}$인 자료에 두 방법을 적용한 결과는 [표 9]와 같다. 검정력은 모두 유의수준 0.05에서 계산되었다. Allison 방법에서의 검정력은 38%~60%로 나타났고, Fulker 방법의 검정력은 32%~36%로 나타났다. 두 방법 모두 한 가족 당 형제수가 증가할수록 검정력이 증가하였는데, Allison 방법에서는 형제가 5명인 경우에는 3명인 경우에 비해 8% 증가하였고, 형제가 2명인 경우에 비해 형제가 3명인 경우에는 38%에서 52%, 14%로 형제가 3명인 경우에서 5명인 경우에 대한 증가분의 약 2배 정도 증가하였다. Fulker 방법에서도 마찬가지로 형제수가 증가할수록 검정력이 증가하였지만, 전반적으로 증가폭이 너무 작아 미미

한 차이를 보였다. 형제가 3명일 때 2명인 경우에 비해 3% 증가하였으나 형제가 5명인 경우에는 3명인 경우보다 1%밖에 증가하지 않았다. 이 결과에서 알 수 있듯이, Allison 방법이 Fulker 방법에 비해 모든 경우에 대하여 검정력이 높았고, 형제수가 증가할수록 검정력의 증가폭도 더 크게 나타났다.

표 9. Allison 방법과 Fulker 방법의 검정력 비교

	Allison 방법	Fulker 방법
형제 2명	38%	32%
형제 3명	52%	35%
형제 5명	60%	36%

* $\theta = 0$, $\{\mu_{qq}, \mu_{Qq}, \mu_{QQ}\} = \{-1, 0, 1\}$

5.3 제1종 오류율

Allison 방법과 Fulker 방법의 제1종 오류율은 유의수준 0.05에서 잘못된 귀무가설을 기각하는 비율로 계산된다. 제1종 오류율을 비교하기 위해, 검정력을 계산하기 위해 얻은 모의실험 자료와 마찬가지로 완전 연관성이 있다는 가정하에서 QTL과 단일표식자간의 $\theta = 0$으로 설정하고, 표현형에 따른 평균은 $\mu_{qq} = \mu_{Qq} = \mu_{QQ} = 0$으로 주었다. 두 방법을 적용한 결과는 [표 10]과 같다. Allison 방법에서의 제1종 오류율은 4%~10%로 나타났고, Fulker 방법의 검정력은 5%~8%로 나타났다. 두 방법 모두 한 가족 당 형제수가 증가할수록 제1종 오류율이 감소하였는데, Allison 방법에서는 형제가 5명인 경우에는 3명인 경우에 비해 1% 감소한 반면, 형제가 2명인 경우에 비해 형제가 3명인 경우 제1종 오류율의 감소율이 5%로 매우 컸다. Fulker 방법에서도 마찬가지로 형제수가 증가할수

록 제1종 오류율이 감소하였지만, 전반적으로 감소폭에서 많은 차이를 보이지는 않았다. 형제가 2명인 경우를 제외하고는 Allison 방법이 Fulker 방법보다 제1종 오류율이 더 낮았다. 그러나 형제가 두 명인 경우에는 Allison 방법이 Fulker 방법보다 제1종 오류율이 더 높게 나타났으나 2%로 거의 차이가 없다고 할 수 있다. 전반적으로 두 가지 방법에서 각 자료유형에 관한 제1종 오류율의 차이는 거의 없다.

표 10. Allison 방법과 Fulker 방법의 제1종 오류율 비교

	Allison 방법	Fulker 방법
형제 2명	10%	8%
형제 3명	5%	6%
형제 5명	4%	5%

* $\theta = 0,\ \mu_{qq} = \mu_{Qq} = \mu_{QQ} = 0$

제 6 장 토의 및 결론

지금까지 양적형질의 유전자 관련성 분석을 검정하기 위한 Allison 방법과 Fulker 방법에 대하여 알아보고 검정력과 제1종 오류율을 비교하였다. 두 방법 모두 부모정보를 알 수 없을 때 형제자료만으로도 표식자와 양적형질간의 관련성을 검정할 수 있는 방법이고, 이러한 관련성 검정 시 인구집단 층화나 혼합성으로 인한 의사관련성을 보정할 수 있는 방법이다.

실제자료에 두 방법을 적용한 결과, 두 방법에서 모두 표식자(ApoE)와 양적형질(LDL)간의 유의한 관련성이 있다는 결과를 얻지 못하였다. 이는 양적형질 LDL이 표식자 외에도 많은 환경적·인구학적 요인들에 의해 영향을 받는 형질이고, 여러 유전자에 의해 영향받을 가능성이 있기때문이라고 할 수 있다.

모의실험을 통해 분석한 결과에서 알 수 있듯이 Allison 방법은 Fulker 방법에 비해 모든 자료유형에 대하여 검정력이 높았고, 제1종 오류율은 대체로 낮았다. 비록, 형제가 2명인 경우에 Fulker 방법보다 2% 높은 제1종 오류율을 나타냈지만, 이는 모의실험의 개수를 증가시키면 개선될 수 있을 것이다. 따라서 모의실험 결과에서, 검정력은 Allison 방법이 더 좋았고 제1종 오류율은 두 방법에서 별 차이를 보이지 않았다. Fulker 방법을 적용하기 위해서는 IBD비율을 별도로 추정해야하는 과정이 필요한 반면, Allison 방법은 비교적 간단하게 적용시킬 수 있는데, Fulker 방법이 Allison에 비해 검정력이 낮은 결과를 보이는 이유는 부모정보가 없는 형제자료에서 IBD비율을 추정하는데 있다고 볼 수 있다. 즉, 자녀들의 유전형만으로 부모의 유전형을 추정하고 그에 따라 IBD비율을 추정해야하기 때문에 왜곡된 결과를 도출할 수 있는 것이다. 이러한 점에서, 복잡한 IBD비율을 고려하지 않고 접근할 수 있는 Allison 방법은 더 효율적이라고 할 수 있고, 또한 공변량의 효과도 함께 검정할 수 있는 장점을 갖고 있다고 할 수 있다. 물론, 핵가족뿐만 아니라 확장된 가계자료를 이용하여 관련성 분석을 하기 위한 방법들이 제시되고 있기는 하지만 심혈관 질환처럼 고연령에 발생하는 질병인 경우에는 환자의 부모에 관한 정보를 얻기가 어렵기 때문에 그러한 방법을 적용할 수 없게 된다. 따라

서 형제자료만으로도 관련성 검정을 할 수 있는 방법은 현실적으로 그 중요성을 갖고 있고, 이 논문에서는 그러한 방법 중 대표적인 Allison 방법과 Fulker 방법에 관하여 논의하였다.

이 본문에서는 표식자가 하나인 경우에 관하여 두 방법에 대해 논의하였지만, 실제로 여러 개의 유전자가 질병에 관여하는 경우가 많기 때문에 앞으로 두 방법을 이용하여 다중표식자(multiple marker)인 경우까지 확장하는 방법에 관하여 논의되어야 할 것이다. 또한, 부모정보를 얻을 수 있는 경우에는 그 정보를 최대한 분석에 이용하고, 불가피하게 부모정보를 얻을 수 없는 경우에도 적용할 수 있는 관련성 검정에 관한 방법이 연구되어야 할 것이다.

참 고 문 헌

Brown H, Prescott R (1999) Applied mixed models in medicine. John Wiley & Sons Ltd, England

Camp NJ, Cox A (2002) Quantitative trait loci: methods and protocols. Human Press Inc, New Jersey

Elston RC, Olson JM, Palmer L (2002) Biostatistical genetics and genetic epidemiology. John Wiley & Sons Ltd, England

Littell RC, Milliken GA, Stroup WW, Wolfinger RD (1996) SAS system for mixed models. Cary, NC, SAS Institute Inc

Abecasis GR, Cardon LR, Cookson WO (2000) A general Test of association for quantitative trsits in nuclear families. Am J Hum Genet 66:279-292

Allison DB (1997) Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 60:676-690

Allison DB, Heo M, Kaplan N, Martin ER (1999) Sibling-based tests of linkage and association for quantitative trait. Am J Hum Genet 64:1754-1763

Allison DB, Schork NJ (1997) Selected methodological issues in meiotic mapping of obesity genes in humans: issue of power and efficiency. Behav Genet 27:401-421

Blackwelder WC, Elston RC (1982) Power and robustness of sib-pair linkage studies. Am J Hum Genet 61:423-429

Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950-961

Burdock RK, Graybill FA (1992) Confidence intervals on variance components. Marcel Dekker, New York

Collins FS, Guyer MS, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580-1581

Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319-333

Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 56:811-812

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455-464

Fulker DW, Cherny SS (1996) An improved multipoint sib-pair analysis of quantitative traits. Behav Genet 26:527-532

Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association analysis for quantitative traits. Am J Hum Genet 64:259-267

George VT, Elston RC (1987) Testing the association between polymorphic markers and quantitative traits in pedigrees. Genet Epidemiol 4:193-201

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3-19

Hume D (1748[1988]) An enquiry concerning human understanding. Prometheus Books, Amherst, MA

Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56:519-527

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439-454

Lahoz C, Schaefer EJ, Cupples LA, Wilson PW, Levy D, Osgood D, Parpos S, Pedro-Botet J, Daly JA, Ordovas JM (2001) Apolipoprotein E genotype and cardiovascular disease in the Framinham Heart Study.

Atherosclerosis 154:529-537

Monks SA, Kaplan NL (2000) Removing the sampling restrictions from family-based tests of association for a quantitative trait locus. Am J Hum Genet 66:576-592

Morris AP, Whittaker JC, Curnow RN (1997) A likelihood ratio test for detecting patterns of disease-marker association. Ann Hum Genet 61:335-350

Neter J, Kutner MH, Nachtshem CJ, Wasserman W (1996) Applied linear statistical models, 4th ed. Irwin, Chicago

Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Hum Hered 47:342-350

Rao RC, Mitra SK (1971) Generalized inverses of matrices and its applications. New York, Wiley

Risch N, Merikangas K (1996) The future of genetic studies of complex human disease. Science 273:1516-1517

Rosenberg IH (1997) Sarcopenia: origins and clinical relevance. J Nutr 127 Suppl 5:990S-991S

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test(TDT) for multi-allele marker loci. Ann Hum Genet 59:323-336

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450-458

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus(IDDM). Am J Hum Genet 52:506-516

Trégouët D-A, Ducimetière P, Tiret L (1997) Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. Am J Hum Genet 61:189-199

Wilson PW, Myers RH, Larson MG, Ordovas JM, Wolf PA, Schaefer EJ (1994) Apolipoprotein E alleles, dyslipidemia, and coronary heart disease. The Framingham Offspring Study. JAMA 272:1666-1671

ABSTRACT

A comparison of sibling-based association tests for quantitative traits

Kim, Minji
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

The transmission/disequilibrium test(TDT) developed by Spielman et al.(1993) is a powerful family-based association test of qualitative trait. It has recently been extended to the allowance for implementation with quantitative traits as well as qualitative traits by Allison(1997) et al.

In this thesis, we compare the Allison's mixed-effects model with Fulker's maximum-likelihood variance component method that these include allowance for multiple alleles, allowance for testing the absence of parental data. In addition to that, the spurious associations due to population stratification and admixture can be adjusted by these methods.

Allison's method considers the marker genotype as a fixed factor, the siblings in each sibship as a random factor and covariates can easily be accommodated. Fulker's method partitions the mean effect of a locus into

between- and within-sibship components.

In the analysis of real data, there is no statistically significant results in both methods. The power and type I error rate of these methods are illustrated through simulation in the generated data(2 , 3 and 5 siblings in each sibship, respectively). In the Allison's method, the power improves when the number of the siblings in each sibship increases and is higher than Fulker's with respect to the power. In the type I error rate, Allison's method is somewhat smaller than Fulker's however, the differences of theirs are low.

Key words : quantitative trait, population stratification, sibling, association test, Allison's method, Fulker's method