CHAPTER 15

연관성에 대한 검정(1) : 카이제곱 검정

Tests for Association (1) Chi-square

- 연관성에 대한 검정과 카이제곱 검정은 무엇인가?
- 카이제곱 검정은 어떻게 사용하는가?
- 카이제곱 검정을 사용하는 데 있어 제한점은 무엇인가?

연관성에 대한 검정 Test for association

연구를 수행할 때 차이보다는 두 변수 사이에 연관성이 있는가에 관심이 있는 경우가 종종 있다. 예를 들어 특정 병리 소견(pathology)이 특정 인종과 연관성이 있는가에 관심이 있을 수 있다. 이러한 정보를 아는 것은 보건의료서비스(health care service) 계획과 공중 보건 발전에 도움을 줄 것이다. 이러한 연관성의 예는 낭포성 섬유증(cystic fibrosis)과 코카시안계 백인의 연관성과 같이, 아프리카계 캐리브인과 겸상 적혈구 빈혈(sickle cell anaemia) 사이에서 볼 수 있다. 명목척도를 이용하여 측정된 자료에서 연관성을 찾을 때 통상적으로 카이제곱 검정(chi-square test)을 사용한다.

다른 형태는 한 변수의 변화가 다른 변수의 변화와 연관성이 있다는 것을 밝히는 것이다. 예를 들어 연령과 유방암 발생과의 연관성이다. 때로는 한 변수가 감소하면 다른 변수가 증가하는가를 보고자 할 경우가 있다. 예를 들면 재산의 증가는 정신건강 문제의 발생률을 낮추는 것과 연관성이 있다. 이러한 형태의 연관성을 상관성(correlations)이라고 한다. 상관성은 보통 서열척도, 구간척도와 비율척도로 측정된 자료를 이용한다. 이 장에서는 카이제곱 검정을 살펴보고 16장에서는 상관성을 살펴볼 것이다.

그리스 문자 카이(chi)는 χ로 쓴다. 그러므로 카이제곱에 대한 표현은 χ^2이다. 여러 가지 형태의 χ^2검정이 있다. 그러나 모든 검정은 카이제곱 분포(χ^2 distribution)에 근거하고(연관성에 대한 검정을 사용하지만) 실제로 관찰된 값과 결과로 기대되는 값 사이의 차이를 찾는 것에 기초한다. 카이제곱 검정은 가장 널리 사용되는 통계 검정 중 하나이다. 카이제곱 검정은 언제나 사람 수(count of people)나 사물의 수로 측정된 자료를 사용한다. 이 장에서 두 가지 형태의 χ^2검정을 살펴볼 것이다. 연관성

검정을 위한 일원 카이제곱 검정(one-way χ^2 test)과 이원 카이제곱 검정(two-way χ^2 test)이다.

Box 15.1

χ^2검정은 언제 사용하나?

- 변수가 명목척도로 수집되었을 때
- 변수와 특정 현상 사이의 연관성을 보고 싶을 때

카이제곱 검정을 시행하기 전에 자료를 백분율로 전환하지 않는다.

일원 카이제곱 검정 One-way chi-square test

이 검정은 실행하기 가장 간단한 통계 검정 중 하나로 다음과 같이 계산한다.

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

O는 관찰빈도(observed frequency)이고 E는 기대빈도(expected frequency)이다. 기호 Σ은 '합(the sum of)'을 의미한다. 그러므로 χ^2은 모든 $(O-E)^2/E$를 합한 값이다. χ^2검정에 대하여 기억하여야 하는 중요한 것은 **빈도(frequencies)**를 사용한다는 것이다. 자료가 빈도로 기록될 수 없거나 빈도로 전환될 수 없다면 χ^2검정은 사용할 수 없다. 이 검정은 단지 하나의 변수만 포함되기 때문에(이 변수가 여러 범주로 나누어지긴 하지만) 일원 카이제곱 검정이라고 부른다.

표 15.1 간호사 교육에 참여한 코호트에서 개인의 인종(ethnicity of individuals)

	백인(UK와 EU)	캐리비안인	인도계인	계
관찰된 수	34	62	28	124

Box 15.2

눈 색(Eye colour)과 카이제곱 검정

- 개인을 대상으로 눈 색을 표본추출하였다. 눈 색과 표본 집단 사이에 유의한 연관성이 있는지 보기 위하여 검정을 시행한다. 이때 동일성(homogeneity)을 알아보기 위한 카이제곱 검정을 사용한다.
- 동일성을 알아보기 위한 카이제곱 검정은 적절한가? 기대 값(expected value)을 어떻게 결정할 것인가?

예제를 살펴보자. 간호사 교육에 참여한 서로 다른 인종 배경을 갖는 사람들의 비율에 관심이 있다고 하자. 분명히 인종(ethnicity)은 명목 수준의 자료이고 간호사 교육과 특정 인종 간의 연관성을 찾고자 한다. 그러므로 χ^2검정을 사용하기 원한다. 검정하고자 하는 가설은 특정 인종과 간호사 교육 사이에는 연관성이 있다는 것이다. 물론 영가설(null hypothesis)은 연관성이 없다는 것이다.

가설을 설정한 후, 다음 단계는 간호사 교육에 참여한 각 인종의 학생 수를 기록하는 것이다(모집단은 아마도 지리적 지역에 의하여 정의되었을 것이다(제3장). 여기서는 런던의 동부지역에 관심이 있다고 하자. 그때 이러한 자료들이 관찰된다. 표 15.1에 가상적인 값이 있다. 자료를 비율(proportions)이나 백분율(percentages)로 전환하지 않는 것이 중요하다.

단순화를 위하여 세 인종을 이용하고 성인 환자를 보살피기 위한 교육을 받은 학생의 첫해 코호트의 인종을 살펴보았다고 하자. 이러한 자료가 표 15.1과 같이 관찰되었다. 계산한 기대값은 이론적인 기대 값에 의존한다. 이 예제에서 간호사 교육과 특정 인종 간에 연관성이 없다면 각 인종이 지역사회를 반영해서 표현되었다는 것이 기대된다.

기대빈도(expected frequencies)

- 남성에 대하여 62×0.5=31.
- 여성에 대하여 62×0.5=31.

인종 구성이 London Borough of Hackey의 구성과 유사한 런던 동부지역을 선택했다면 전체의 48%가 백인, 33%가 캐리비안인이고 19%가 인도계인임을 기대할 수 있다. 전체 학생 수에 각 집단(백인, 캐리비안인, 인도계인)의 비율을 곱해서 기대 빈도(expected frequency)를 계산할 수 있다. 표 15.2에 각 범주에 대한 관찰빈도와 기대빈도를 산출하였다.

표 15.2 간호 학생의 교실에서 세 인종 집단의 관찰빈도와 기대빈도

빈도	백인(UK와 EU)	캐리비안인	인도계인	계
관찰	34	62	28	124
기대	59.5	41	23.5	124

기대빈도를 계산한 후 χ^2검정을 실행할 수 있다.

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = \frac{(34-59.5)^2}{59.5} + \frac{(62-41)^2}{41} + \frac{(28-23.5)^2}{23.5}$$

$$= \frac{(-25.5)^2}{59.5} + \frac{(21)^2}{41} + \frac{(4.5)^2}{23.5}$$

$$= \frac{650.25}{59.5} + \frac{441}{41} + \frac{20.25}{23.5}$$

$$= 10.9 + 10.7 + 0.86 = 22.46$$

따라서 χ^2=22.46이다. 이제 χ^2분포에서 이 값을 찾아보아야 한다. 이때 얼마나 큰 자유도(degree of freedom)를 갖는가를 알 필요가 있다. 동일성 검정을 위한 χ^2검정에서 자유도는 범주의 수−1로

주어진다. 이 경우에 세 범주(백인, 캐리비안인, 인도계인)가 있으므로 자유도는 2가 된다. 자유도 2를 갖는 χ^2분포표에서 p=0.05에 해당하는 값은 5.99이다(부록 2). 앞에서 구한 χ^2값이 5.99보다 크기 때문에 유의수준 0.05하에서 유의한 차이가 있다고 말할 수 있다($p\langle 0.05$). 사실 앞에서 구한 χ^2값은 p=0.01에서의 임계점(critical value)보다 크다. 따라서 유의수준 0.01하에서도 유의한 차이가 있다고 말할 수 있다($p\langle 0.01$). 따라서 간호사 교육과 인종 간에는 유의적인 연관성이 있다는 결론을 내릴 수 있다. 이 특정 기관에서 간호사 훈련에 참여한 백인은 기대되는 수보다 적었으며, 아프리카계 캐리비안인은 기대되는 수보다 더 많았다.

Box 15.3

적합도(Goodness of fit)와 카이제곱 검정

적합도를 살펴보기 위하여 사용한 검정과 여기서 기술된 것이 무엇인가를 스스로 질문해 보자. 답은 간단하다. 중요한 차이는 적합도 검정에서 기댓값은 실제 자료가 아니라 수학적 모형을 이용하여 계산된 값이라는 것이다. 기댓값을 산출하기 위하여 사용한 수학적 모형은 가상적이기 때문에 기대 값이 실제 자료를 이용하여 계산된 것보다 더 많은 자유도를 잃을 수 있다.

기대 범주가 모두 같은 경우 χ^2의 특정 형태가 존재한다. 예를 들어 성에 기초한 연관성을 찾을 때 기대빈도는 남성 50%, 여성 50%를 제시하는 것은 아마도 합리적일 것이다.

이처럼 기대빈도가 같은 경우의 χ^2검정은 동일성 검정을 위한 χ^2검정으로 알려졌다. 이 검정을 위한 계산은 위에서 기술한 것과 정확하게 같다.

주의할 점 Thinks to look out for

에이츠 수정: 제한된 집단의 수에 대한 수정 Yates correction: a correction for a limited number of groups

두 범주를 가질 때(특정 성별과의 연관성을 찾을 때) 자유도 1을 사용하게 된다. 이러면 수정이 필요하다. 이 수정을 에이츠 수정(Yates' correction)이라고 부른다. 에이츠 수정은 x^2 통계량을 계산하기 위한 공식을 약간 변형하면 된다. 새로운 공식은 다음과 같다.

$$\chi^2 = \Sigma \frac{(|O-E|-0.5)^2}{E}$$

관찰 값-기댓값, 즉 분자(분수의 위쪽)의 $O-E$의 양쪽에 부호인 수직선(|)을 볼 수 있다. 방정식에서 이 부분을 먼저 계산한 후 0.5를 뺀다. $O-E$의 양쪽에 있는 수직 막대는 부호를 무시한 $O-E$ 값에서 0.5를 뺀다는 것을 의미한다. 이 과정이 끝난 후에 χ^2검정을 진행한다.

예를 들어 $(7-12-0.5)$의 합을 계산하면 -5.5이다. 그러나 $(|7-12|-0.5)$의 합을 계산하면

4.5이다. 에이츠 수정 χ^2값을 계산할 때 분자는 $(|O-E|-0.5)^2$을 사용한다.

Box 15.4

연습(Practice)

개인을 대상으로 성별을 표본추출하였다. 특정 성과 표본 사이에 연관성이 있는가를 보기 위하여 x^2검정을 사용한다.

카이제곱 검정을 사용하는 데 있어서의 제한점 Restrictions on the use of the chi-square

기대빈도가 5보다 작은 값이 하나라도 있다면 χ^2검정을 사용해서는 안 된다. 예를 들어 150쪽에서 사용한 예제에서 인종을 좀 더 정밀한 범주로 나누었다. 예를 들면 '인도계인'을 방글라데시인, 파키스탄인, 시크교인, 이슬람교계 인도인과 힌두교계 인도인으로 나누었다. 이럴 때 각 범주는 5보다 작은 기대빈도를 갖는다는 것을 알 수 있다. 이 예제가 보여주는 것처럼 이 문제를 해결하는 하나의 방법은 범주를 합하는 것('인도계인'의 형태)이다. 그러나 이러한 접근에 대한 문제는 합해진 범주가 그 의미를 잃을 수 있다는 것이다. 범주 '인도아대륙인'은 너무 다양해서 많이 사용되지 않는다는 논쟁이 쉽게 일어난다.

범주를 합할 수 없다면, 5보다 작은 기대빈도를 가질 것이고 행할 수 있는 유일한 방법은 다른 검정을 이용하는 것이다. 이때 사용할 수 있는 검정은 G 검정(G test) 또는 피셔의 정확 검정(Fisher's exact test)이다(Sokal and Rohlf, 1966).

독립성 independence

χ^2검정을 이용할 때 각 자료는 서로 독립이어야 한다. 예를 들어 서로 다른 민족의 병원 이용을 살펴본다면 같은 사람이 재방문한 것은 배제할 필요가 있다. 또한, 각 범주가 서로 배타적으로 되도록 하여야 한다. 개인이 하나 이상의 범주에 들어가는 일은 일어나지 않아야 한다.

연관성에 대한 이원 카이제곱 검정 Two-way chi-square test for association

동시에 하나 이상의 명목척도 변수에 관심이 있을 때가 있다. 예를 들면 HIV에 대한 노출 여부가 특정 지역과 연관이 있는가에 관심이 있을 수 있다. 여기에는 두 명목척도 변수인 HIV 노출과 지역이 있다. 이럴 때 이원 χ^2검정(two-way χ^2 test)을 사용한다. 이 검정은 연관성 검정(test for association)을 위한 χ^2검정이라고 부른다. 다른 형태의 χ^2검정이지만 기본 공식은 같다.

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

이 연관성 검정을 위한 χ^2검정과 다른 점은 기대빈도를 어떻게 계산하는가이다. 기대빈도를 계산하는 방법은 예제를 통하여 아주 잘 설명되어 있다.

Box 15.5

연관성이 있는가를 알아볼 필요가 있는 세 쌍의 변수를 적어보자.

다음의 예제에서 세 유럽 국가에서 MDR 결핵(multi-drug-resistant tuberculosis) 사례에 대한 빈도를 살펴보았다. 이 예제에서 MDR 사례의 비율에 대하여 실제 자료(http://www.eurotb.org)를 사용하였지만, 결핵 사례의 전체 수에 대해서는 가상 자료를 사용하였다. 크로아티아, 체코와 리투아니아 등 세 국가를 살펴볼 것이다. MDR 결핵 발생과 지역 사이에 유의한 연관성을 보이는가? 물론 영가설은 연관성이 없다는 것이고 관찰된 차이는 단순히 우연히 일어났다는 것이다.

표 15.3 세 동유럽 국가에서 MDR 결핵의 발생률을 보여주는 3×2 분할표

국가	MDR 결핵	Non-MDR 결핵	계
크로아티아	칸 13	299	행의 합 312
체코	6	184	190
리투아니아	123	170	293
계	열의 합 142	653	전체 합 795

가설을 설정하고 자료를 수집한 후, 다음 단계는 **분할표(contingency table)**를 만드는 것이다. 분할표는 단순히 표준화된 표의 형태로 자료를 정리한 것이다. 분할표를 만들기 위하여 한 변수의 범주를 첫 번째 행에 적고 다른 변수의 범주를 첫 번째 열에 적는다. 각 첫 번째 행과 열에 대하여 각 행과 각 열의 합계를 적는다(표 15.3). 행에 세 범주가 있고 열에 두 범주가 있기 때문에 이러한 형태의 표를 3×2 분할표라고 부른다. 행에 두 범주만 있다면, 2×2 분할표라고 부른다.

Box 15.6

출산 경험과 사용된 진통제의 관련성을 보기 위한 다음의 자료를 이용하여 두 변수 사이에 연관성이 있는지를 보기 위하여 χ^2검정을 시행한다.

경험	진통제 선택		
	가스 (gas and air)	페티딘 (pethidine)	경막의 마취제 (epidural)
첫 번째 출산	66	56	36
두 번째 출산 이후	102	22	12

독립성과 관련하여 이 자료가 가지고 있는 문제는 무엇인가?

다음 단계는 기댓값을 계산하는 것이다. χ^2검정을 이용할 때 기댓값이 무엇인가는 이미 알고 있다. 인종을 살펴보기 위한 앞 절(previous section)의 예제에서 지역사회 내의 인종의 빈도를 사용하였다. 성별에 대하여 살펴볼 때는 남성과 여성의 비율을 사용하였다. 적합도 검정을 위한 χ^2검정에서는 수학적 모형을 이용하였다. 연관성 검정을 위한 χ^2검정은 기대빈도를 예측하기 위하여 실제 자료를 이용한다.

기대빈도를 계산할 때 자료가 두 변수(국가와 MDR 발생) 모두에 대하여 분포한다는 사실을 설명할 필요가 있다. 각 칸에 대한 기대빈도는 그 칸이 속한 행의 합과 열의 합을 곱한 후 전체 합으로 나누어 계산할 수 있다. 표 15.3에서 맨 위 칸에 대하여 기대빈도는 (312×142)/795=55.7이 된다. 이러한 과정을 모든 칸에 대하여 반복한다. 표 15.3에 대한 기대빈도의 결과를 표 15.4에서 보여주고 있다.

각 칸에 대하여 다음의 공식을 이용하여 값을 계산한다.

$$\frac{(O-E)^2}{E}$$

이 값이 표 15.5에 나타나 있다.

표 15.4 표 15.3에서 각 칸에 대한 기대빈도

국가	MDR 결핵 사례	Non-MDR 결핵 사례
크로아티아	55.7	256.3
체코 공화국	33.9	156.0
리투아니아	52.3	240.7

표 15.5 표 15.3과 15.4에서 각 칸에 대한 카이제곱 계산

국가	MDR 결핵 사례	Non-MDR 결핵 사례
크로아티아	32.7	7.1
체코 공화국	23.0	5.0
리투아니아	95.6	20.8

Box 15.7

- 표가 2×2 분할표이면 자유도는 얼마인가?
- 이와 같은 경우에 무엇을 하여야 하는가? 152쪽을 참조하라.

다음 단계는 χ^2값을 구하기 위하여 각 칸의 계산 결과를 모두 합하는 것이다. 이 예제에서 χ^2 =184이다.

χ^2분포표에서 이 값을 찾기 전에 우선 자유도가 얼마인가를 알 필요가 있다. 연관성에 대한 χ^2 검정에서 자유도는 $(r-1)\times(c-1)$로 주어진다. 여기에서 r은 행의 범주의 수이고 c는 열의 범주의 수이다. 이 예제에서 자유도는 $(3-1)\times(2-1)=2$이다. χ^2표로부터 $p=0.01$에 해당하는 임계점은 9.21이다. 계산된 χ^2값이 9.21보다 크다. 그러므로 지역과 MDR 결핵 발생 간에 연관성이 없다는 영가설을 기각할 수 있다. 이러한 결과에 대하여 어떤 생각을 할 수 있는가?

χ^2검정에 대한 대안이 있다. 가끔 교재에서 인용되는 것 중 하나가 G 검정(G test)이다. 실제로 여러 통계학자는 G 검정이 χ^2검정보다 우월하다고 생각한다. 그러나 χ^2검정을 더 많이 이해하고 있고 G 검정을 지원하는 통계 패키지(statistical package)가 거의 없어서 전통적인 접근 방법을 사용하고 있다. G 검정에 대한 더 많은 것을 알고 싶으면 Sokal and Rohlf(1966) 책을 참조하시오. 이 책에서 G 검정에 대하여 자세히 설명하고 있다.

이번 장을 읽고 연습문제를 풀면 아래와 같은 생각들과 단어에 익숙해질 것이다:

- 연관성에 대한 검정
- 명목척도로 측정된 변수의 검정
- χ^2검정을 사용하는 데 있어 제한점
- 기대 값의 계산
- 에이츠 수정

연 습 문 제

1. 병원 연구(5장)에 대하여 병원 방문이 특정 증상과 연관성이 있는가를 보기 위한 적절한 검정을 시행하시오.
   (a) 이러한 증상이 성병과 강한 연관성을 가지고 있는가에 관한 결과는 무엇인가?
   (b) 이러한 관계에 대하여 어떤 설명을 할 수 있는가?
   (c) 피임약의 선택과 인종 간에 관련성이 있는가?
   (d) 이러한 관계에 대하여 어떤 설명을 할 수 있는가?
2. 병원 방문에 대하여 인종과 성별 사이에 연관성이 있는가? 이러한 차이에 대하여 어떤 설명을 할 수 있는가?
3. 개인으로부터 눈 색을 표본추출하시오. 이때 눈 색과 표본 집단 사이에 유의한 연관성이 있는가를 보기 위한 검정을 시행하라. 동일성 검정을 위한 χ^2검정을 시행하라. 동일성 검정을 위한 χ^2 검정은 적절한가? 기대 값을 어떻게 결정할 수 있는가?