통계학 입문

2021-05-23

데이터과학융합스쿨

손 대 순 교수

앞의 강의에서 독립성 검정의 개념에 대해서 이야기 했습니다. 이는 카이제곱 검정을 이용하지요. 카이제곱 검정이 오늘의 주제입니다. 강의 교안을 통해서 한 번 더 이 검정법에 대해 이해해 보도록 합시다. 카이제곱 검정은 여러가지 종류가 있어요. 카이제곱 분포를 이용해서 확률계산을 하면 모두 카이제곱 검정이라고 하니, 종류가 참 많습니다. 그 중에서 독립성 검정, 동질성 검정, 그리고 적합도 검정을 설명하려고 합니다. 모두들 하는 방법은 간단한데, 그 차이를 명확하게 알고 사용해야 하죠.

우선 독립성 검정부터 살펴 봅시다.

성별	전공			
	A 전공	B전공	C전공	Total
남자	75	46	23	144
여자	30	32	24	86
Total	105	78	47	230

자, 위의 테이블을 한 번 봅시다. 전공이라는 것과 성별이라는 것에 연관성이 있는지에 관심이 있다고 하죠. 연관성이라는 것이 바로 독립성과 마찬가지 개념입니다. 연관이라는 것은 두 요인이 서로 영향을 주고 받는 것을 말합니다. 성별이라는 것이 전공선택이라는 요인과 어떤 관련성이 있는가에 해당하는 것이죠. 따라서, 이 두 요인은 병렬적입니다. 남녀라는 구분이 전공선택이라는 구분과는 선/후 관계가 없는 병렬적인 요인이라는 것입니다. 뒤에 동질성 검정과 비교해서 보면 더 이해가 쉬울거예요.

아무튼 연관성 검정(association test)는 곧 독립성 검정입니다. 앞에서 독립이라는 것의 정의가 무엇인지 장황하게 설명한 적이 있어요. 결국 $P(A전공 \cap 남자) = P(A전공)P(남자)$로 볼 수 있는지에 대한 검정이 바로 독립성 검정입니다. 이것이 귀무가설이 되는 것이구요.

귀무가설 : 성별과 전공선택과는 관계가 없다. (서로 독립이다)
대립가설 : 성별에 따라 전공선택에 차이가 있다.
(서로 독립이 아니다)

표에서 보면 P(A전공∩남자)=75/230입니다. P(A전공)P(남자)=105/230 X 144/230이죠. 이 둘을 같다고 볼 수 있는가? 라는 질문이죠. 위 표의 모든 칸(cell)에서 이와 같은 개념이 적용되는 것이예요. 검정이라는 개념이 귀무가설이 참이라는 가정하에서 진행하는 것이니 두 요인이 독립임을 가정하고, 계산한 기대값이 있을 겁니다.

$$E_{11} = \left(\frac{105}{230}\right)\left(\frac{144}{230}\right)230 = 65.7$$

$$\vdots$$

$$E_{23} = \left(\frac{86}{230}\right)\left(\frac{47}{230}\right)230 = 17.6$$

위와 같이 모든 cell에 대해서 기대값을 계산할 수 있지요. 카이제곱 검정통계량은

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_\alpha\left((r-1)(c-1)\right)$$

위와 같이 생겼습니다. 검정통계량의 구성요소는 관찰치와 기대값이 전부예요. 각 cell의 기대값을 구하고, 관찰값은 표에 있는 값이니 금방 구할 수 있겠죠?

$$\chi^2 = \frac{(75-65.7)^2}{65.7} + \cdots + \frac{(24-17.6)^2}{17.6} = 7.68$$

$$\sim \chi^2_\alpha(df=(2-1)(3-1)=2)$$

결국 이 말은, 관찰값과 기대값의 차이가 크면, 카이제곱 통계량 값이 커진다는 것을 의미합니다. 즉 독립임을 가정하고 계산한 기대값이, 실제 관찰값과 차이가 없다면 독립이라는 말이고, 그 차이가 크다면 서로 연관성이 있다는 뜻이지요.

그렇다면 동질성 검정을 살펴 봅시다. 동질성 검정은 독립성 검정과는 좀 다른 세팅입니다.

아래의 테이블을 한 번 살펴 보죠. 두 가지 식이요법의 효과를 비교하기 위해 150명을 대상으로

조사합니다. 임의로 추출된 80명에게는 식이요법 A를, 나머지 70명에게는 식이요법 B를 적용한 후 참여자의 건강상태를 체크하여 테이블을 얻었습니다.

		건강상태			합계
		양호	보통	불량	
식이요법	A	37	24	19	80
	B	17	33	20	70
합계		54	57	39	150

이러한 세팅에서 수행하는 검정을 동질성 검정이라고 합니다. 이는 건강상태와 식이요법의 독립성 검정이라고 하지 않아요. 식이요법과 건강상태라는 것은 시간적 선/후관계가 존재하죠. 식이요법에 따른 효과를 보는 것이니까요. 그렇다면 동질의 의미는 무엇일까요? 동일한 분포를 가진다는 뜻입니다. 여기에서는 공통확률이라고 표현할 수 있겠어요.

A요법에서 P(양호), P(보통), P(불량) 세 개의 확률이 있군요. 이를 모두 더하면 1이 되구요. 결국 동질성 검정은 A의 P(양호)와 B의 P(양호)가 동일한가, P(보통)이나 P(불량)의 비율도 A, B 상관없이 동일한가를 확인 하는 검정입니다. 이것이 곧 식이요법의 효과로 해석되겠지요. 앞의 독립성 검정과는 가설이 완전히 다르죠?

그러니, 가설에 따라 기대값을 먼저 구합니다.

식이요법 A, B 상관없이 양호, 보통, 불량의 비율이 동일하다는 것이 귀무가설이므로, P(양호)=54/150, P(보통)=57/150, P(불량)=39/150이라고 추정합니다. 이를 토대로 각 cell의 기대값을 구해 봅시다.

첫번째 cell을 예로 들면, 전체가 80명인데, P(양호)=54/150이니

$$80 \times \frac{54}{150}$$

가 되겠네요. 식이요법 B에서 건강상태가 보통이라면 기대값은,

$$70 \times \frac{57}{150}$$

이 될겁니다. 이를 이용해서 동일하게 카이제곱 통계량에 대입하여 통계량을 구하면 됩니다.

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_\alpha\left((r-1)(c-1)\right)$$

카이제곱 통계량은 앞에서와 마찬가지죠. 카이제곱 분포를 따르는 통계량이니, 어떠한 검정방법이든 카이제곱통계량이 바뀌지는 않습니다. 다만 귀무가설에 따라 기대값의 계산법이 달라질 뿐이

죠.

자, 그런데 여기서 마법 같은 일이 일어납니다. 위에서 보듯 독립성 검정과 동질성 검정은 귀무가설이 완전히 다르죠. 그래서 기대값의 계산 방식도 전혀 달랐습니다. 독립성 검정에서는 결합확률이 주변확률의 곱으로 표현된다는 원리를 이용하였고, 동질성 검정에서는 공통확률을 추정하여 두 집단의 총합을 곱해주는 방식으로 각cell의 기대값을 계산했습니다.

동질성 검정에서 예로 들었던 두 cell을 살펴 봅시다. 그리고, 위의 동질성 검정을 위한 표를 이용해 독립성 검정을 한다고 가정해 봅시다.

$$80 \times \frac{54}{150} = \frac{80}{150} \times \frac{54}{150} \times 150$$

$$70 \times \frac{57}{150} = \frac{70}{150} \times \frac{57}{150} \times 150$$

위 식의 등호 왼쪽은 동질성 검정의 기대값 계산이고, 오른쪽은 독립성 검정의 기대값 계산입니다. 서로 다른 원리를 가지고 계산했지만 정확히 동일하죠? 같은 값을 가집니다. 때문에 독립성검정이나 동질성 검정은 그 검정통계량이 동일한, 참으로 묘한 경우입니다. 실제 R에서 구현하는 방법도 동일하죠. 하지만, 같은 검정이 아닙니다. 가설이 완전히 다르니까요. 때문에 결과의 해석도 달라져야 합니다. 두 요인이 독립이라는 것과 두 집단이 동일 분포에서 추출된 표본이라는 것은 서로 다른 의미인 것이죠.

마지막으로 적합도 검정을 살펴 봅시다. 이 검정방법도 동일한 카이제곱 통계량을 사용합니다. 다만 기대값을 계산하는 방법이 조금 다를 뿐이죠.

멘델의 유전법칙을 알고 있죠? 기억 저편에 있는 것을 끄집어 내 봅시다. A:B:C:D=9:3:3:1 !!

어떤 잡종 식물을 교배하여 240개체를 관찰하였더니, A:B:C:D=120:40:55:25로 관찰되었다고 합시다. 멘델의 유전법칙을 따르고 있는 것인가요? 아니면 이 법칙을 벗어난 것일까요?

이 처럼 알고 있는 모비율에 데이터가 적합한지를 검정하는 방법이 적합도 검정입니다. 귀무가설을 생각해 봅시다.

기대값을 계산하는 방법이 다른 것이죠.

$$E(A) = 240 \times \frac{9}{16} = 135$$

$$E(B) = 240 \times \frac{3}{16} = 45$$

이와 같은 방식으로 계산 되겠지요. 그렇다면 카이제곱 통계량을 계산할 수 있겠죠?

동일한 카이제곱 통계량을 이용하지만 그 의미가 다른 세 가지 검정법을 알아 보았습니다.

R을 가지고 실습을 좀 해 보죠.

```
obs <- c(120, 40, 55, 25)
null.prob <- c(9/16, 3/16, 3/16, 1/16)
chisq.test(obs, p=null.prob)
```

적합도 검정은 위와 같이 주어진 확률에 대한 검정인 것이죠.

다음 예제를 한 번 봅시다.

	LN(+)	LN(-)	계	비율
생존	100	50	150	100/150 = 67%
사망	200	100	300	200/300 = 67%
계	300	150	450	

폐암 환자 450명을 대상으로 임파절 전이(lymph node metastasis) 여부를 검사했더니, 위와 같이 양성 300명, 음성 150명이 나왔다고 합시다. 가상의 데이터입니다. 폐암에서 임파절 전이는 곧 다른 장기(organ)로의 전이 가능성이 높다는 것을 의미합니다. 추적 관찰하여, 1년내 사망한 경우와 생존한 경우로 분류하였다고 해 봅시다.

이는 사실 동질성 검정이 되겠죠. 해석도 동질성검정으로 해야 합니다. 즉 생존한 사람의 임파절 전이여부 비율이나 사망한 사람의 임파절 전이여부 비율이 차이가 없다, 혹은 있다 라고 해석해야 하는 것이죠. 임파절 전이와 생존/사망 사이에 연관성이 있다/없다는 정확하지 않은 해석입니다. 그러나, 검정통계량이 같으니 R에서 분석방법은 독립성, 동질성 모두 동일합니다.

```
ln.meta <- matrix(c(100, 50, 200, 100), ncol=2, byrow=T)
ln.meta
chisq.test(ln.meta)
```

R에서 테이블의 형태로 데이터를 입력하는 방법에 대해 익혀두면 좋습니다. matrix 함수를 사용하고, 그 안에 데이터를 입력하지요. 열이 2개라는 정보를 ncol이라는 매개변수로 전달하고, byrow라는 매개변수를 이용해서, 행 우선으로 데이터를 처리하라는 명령을 줍니다. byrow=F로 주고 ln.meta를 쳐 보면 테이블이 어떻게 만들어지는지 비교해 보세요. 이해가 될겁니다. 자, 결과가 어떻게 나오나요? p-value는 1이죠? 생존의 경우나 사망의 경우에 임파절 전이에 대한 분포가 다르다고 볼 수 없다는 결론을 지을 수 있겠습니다.

그런데, 폐암에는 여러가지 세부 유형이 있지만 크게 ADC와 SQC type으로 나누어 보겠습니다. 위의 표를 두 type으로 나누어서 만들어 보죠. 마치 성별에 따라 나누었다고 생각해도 좋겠어요.

<ADC>

	LN(+)	LN(-)	계	비율
생존	80	40	120	80/120 = 67%
사망	100	80	180	100/180 = 56%
계	180	120	300	

<SQC>

	LN(+)	LN(-)	계	비율
생존	20	10	30	20/30 = 67%
사망	100	20	120	100/120 = 83%
계	120	30	150	

두 type으로 나누어 표를 구성해 보았더니 위와 같이 되었습니다. 그렇다면 폐암의 type별로 동질성 검정을 수행해 볼 수 있겠네요. 한 번 해 봅시다.

```
adc <- matrix(c(80, 40, 100, 80), ncol=2, byrow=T)
sqc <- matrix(c(20, 10, 100, 20), ncol=2, byrow=T)
chisq.test(adc)
chisq.test(sqc)
```

결과가 어떤가요? 유의수준 5%에서 유의하다고 판단할 수는 없겠지만, 조금 전 전체를 대상으로 검정할 때와는 사뭇 p-value가 다르죠? p-value가 매우 떨어진 것을 볼 수 있습니다.

이러한 현상을 Simpson's Paradox라고 합니다. 전체의 결과가 세분화된 결과와 다른 양상을 보인다는 것이죠. 우리가 상관계수를 배울 때(다음 시간에 다시 한 번 자세히 설명합니다), 절대로 인과관계로 해석해서는 안된다는 주의사항이 있었죠. 연관성, 동질성 등의 분석을 할 때에도 세분화하면 다른 결론이 나오는지, 제3의 변수(무엇인지 알지 못하는 경우가 많음)에 의해 왜곡된 결론을 한 것은 아닌지 면밀히 살펴야 합니다.

오늘은 지난주까지와는 많이 다른 방식의 테스트를 한 번 경험해 보았네요. 아직은 이해가 어렵고 생소한 느낌인 것이 당연합니다. 통계학은 이러한 뭉게구름 같은 상황이 계속 벌어지는 학문이예요. 모집단을 알고자 하는 것인데, 모집단이라는 것이 뭉게구름이니까요. 어떠한 이론적 체계를 세우고 있는지는 고급통계학에서 다루어야 할 문제들입니다. 다만 여러분에게 전달하고 싶은 것은 데이터의 유형과 특징에 따라 다양한 방법들을 적용할 수 있어야 한다는 것이죠. 이것이 앞서 기술통계학(descriptive statistics) 단계에서 데이터를 유형별로 분류한 이유입니다. 그에 따라 분석 방법이 결정되는 경우가 많으니까요.

오늘도 수고 많았습니다.