8. χ^2분포에 의한 검정 (범주형 자료 분석=교차분석)

Q. 많은 실험 결과들의 측정은 양적(quantitative)보다 질적(qualitative) 또는 범주형(categorical) 즉,

㉠ 사람들을 5개 소득 계층으로 분류
㉡ 쥐가 3가지 자극 중에 하나의 자극에 반응
㉢ A 회사의 사탕은 6가지 색깔로 나누어짐
㉣ 공장 생산 품목을 ‘우량품, 보통, 불량품’으로 분류

등 품질(quality)/특성(characteristic)으로 측정값이 얻어지는데, 이럴 때에는 어떤 방법으로 정리하여 검정하나요?

A. 범주 또는 특성에 대한 리스트를 만들고 각 범주에 해당하는 측정값의 수를 세어 자료의 형태로 요약하여 검정하는 교차분석을 이용한다.

[예] 남녀 간 산아제한에 대한 찬반 검정 교차분석표

			성별		
			남자	여자	전체
찬/반	찬성	빈도	230	420	650
		성별의 %	46.0%	70.0%	59.1%
	반대	빈도	270	180	450
		성별의 %	54.0%	30.0%	40.9%
전체		빈도	500	600	1100
		성별의 %	100.0%	100.0%	100.0%

	값	자유도	유의확률
Pearson 카이제곱	64.985	1	0.000

8-1. 피어슨의 카이제곱(χ^2) 통계량

① χ^2분포를 사용하여 통계적 검정을 실시할 때 기본 가정을 확인해야 한다.

㉠ 첫째, 종속변수가 명명변수에 의한 질적변수이거나 최소한 범주변수여야 한다.

[예] 성별, 인종, 자동차 유형 등

또는 연속변수를 비연속변수로 변환한 범주변수여야 한다.

[예] 지능지수 ; 우수아, 보통아, 저능아
수입 ; 고소득자, 중산층, 저소득자

㉡ 둘째, 획득도수(획득빈도 obtained frequency ; 연구과정에서 얻은 각 범주에 포함되어 있는 도수)와 기대도수 (기대빈도 expected frequency ; 귀무가설 하에서 얻어질 것이라 기대되는 사례수)가 5보다 작은 칸(cell)이 전체 칸 수의 20% 이하여야 한다.

㉢ 셋째, 각 칸의 사례들은 서로 독립적 관계여야 한다.

[예] 인종별로 분류하고 눈동자의 색으로 분류할 때 동일인이 각기 중복되는 일이 없어야 함

⇒ 세 가지 조건을 만족하는 예

산아제한에 대한 남녀간 찬반 결과 획득도수				산아제한에 대한 남녀간 찬반의 기대도수			
	남	여			남	여	
찬성	230	420	650	찬성	295.45	354.55	650
반대	270	180	450	반대	204.55	245.45	450
	500	600	1,100		500	600	1,100

② χ^2분포에 의한 검정을 일반적으로 χ^2검정 즉, 교차분석이라 한다.

㉠ 범주형 자료를 분석

㉡ 획득도수(O_i)와 기대도수(E_i) 사이의 차이를 비교하여 가설 검정

i) 가설에 따른 기대도수의 값이 정확

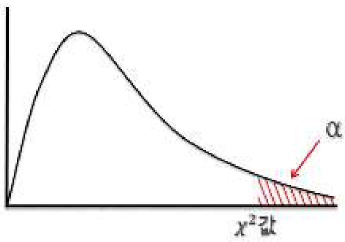
; $(O_i - E_i)$가 작게 되어 χ^2값이 0에 가깝게 됨

ii) 가설에 따른 기대도수의 값이 부정확

; $(O_i - E_i)$가 크게 되어 χ^2값이 큰 값을 갖게 됨

㉢ Pearson의 카이제곱(χ^2) 통계량

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (E_i = np_i)$$



㉣ n이 클 때 χ^2통계량은 반복 표본에서 근사 카이제곱확률분포를 이룸 ; 카이제곱검정

㉤ 오른쪽 꼬리 검정(right-tailed statistical test)을 사용하여 검정 통계량이 특히 큰 값을 갖는지를 확인

㉥ 교차분석의 유형

i) 적합성 검정 : 범주형 자료에 대해 얻어진 관찰값과 이론적으로 계산된 기댓값과의 차이 검정

H_0 : 관찰값은 기대치를 따름

H_1 : 관찰값은 기대치를 따르지 않음

ii) 독립성 검정 : 두 범주형 자료의 독립/종속 검정

〚예〛	두 종의 생물 ; A, B	종의 종류와 관찰 장소의 독립/종속 검정
	관찰된 장소 ; 풀밭, 모래밭, 경계	

H_0 : 종의 종류와 관찰 장소는 서로 독립적으로 분포

H_1 : 종의 종류와 관찰 장소는 서로 상호 관련

iii) 동질성 검정 : 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정

〚예〛	남학생 250명, 여학생 300명	남녀대학생의 생활환경 동일한 분포 검정
	생활환경 : 기숙사, 아파트, 부모님 집, 그 외	

H_0 : 남학생과 여학생의 생활환경 동일한 분포

H_1 : 남학생과 여학생의 생활환경 서로 다른 분포

8-2. 적합성 검정(goodness of fit test)

① 범주형 자료에 대해 얻어진 관찰값이 이론적으로 계산된 기대값과 얼마나 차이를 보이는지를 검정한다.

〖예1〗 부채꼴 5개로 나뉜 원 모양의 다트판에 화살을 던진다면, 이 다트판은 완벽하게 공정한가? 각 점수가 나올 확률이 모두 1/5인가?

〖예2〗 혈액형 A, B, AB, O의 비율이 각각 0.41, 0.10, 0.04, 0.45라고 할 때, 실제 모집단의 비율이 이 확률에 적합한가?

② 관찰도수/기대도수를 사용한 χ^2통계량을 이용한다.

③ k개의 범주 또는 칸(cell)으로 구성된 실험에 대한 χ^2통계량은 자유도 $df=(k-1)$인 χ^2분포를 따른다.

〖예〗 쥐가 녹색/적색/청색 중 어느 하나를 더 좋아하는지 검정하시오.

(쥐를 경사로의 끝에서 분리되는 3개의 다른 색깔의 문으로 유혹하는 실험을 디자인, 쥐를 경사로에 90번 올려 선택한 문 관찰)

	문의 색깔		
	녹색	적색	청색
관찰값(O_i)	20	39	31
	⇩		
기대값	30	30	30

[풀이]

㉠ 가설 설정 (만약 쥐가 더 좋아하는 색깔의 문이 없다면, 쥐가 선택한 문의 수는 동일한 값을 갖게 된다고 기대할 수 있다.)

H_0 : $p_1=p_2=p_3=\dfrac{1}{3}$

(쥐는 녹색, 적색, 청색을 동일하게 좋아한다)

H_1 : 적어도 하나는 $p_i \neq \dfrac{1}{3}$

(쥐가 더 좋아하는 색깔이 있다)

㉡ 카이제곱통계량 $\chi^2=\sum_{i=1}^{3}\dfrac{(O_i-E_i)^2}{E_i}=\dfrac{(20-30)^2}{30}+\dfrac{(39-30)^2}{30}+\dfrac{(31-30)^2}{30}=6.067$

㉢ 자유도 2(=3-1)에서

$$\chi^2_{0.050}=5.99 < \text{검정통계량}=6.067 < \chi^2_{0.025}=7.38$$

검정통계량 $\chi^2=6.067$이므로 $p-$값은 0.025와 0.050 사이의 값이 된다.

㉣ 연구원은 유의수준 0.05에서 유의적이라고 결론을 내리고 결과를 보고할 수 있다. 즉, 문을 선호하는 색깔이 없다는 가설은 기각된다는 의미이다. 즉, 쥐는 3가지 문 중에서 하나의 색깔을 선호한다는 충분한 증거를 가지고 있다.

〖예〗 미국, 백인의 모집단에서 혈액형 A, B, AB, O의 비율이 각각 0.41, 0.10, 0.04, 0.45라 한다. 실제 모집단의 비율이 이 확률에 적합한지 결정하기 위해 200명의 미국인을 무작위 표본으로 선정, 혈액형을 기록. 혈액형의 비율이 적합한지 검정?

	A	B	AB	O
관찰도수	89	18	12	81
		⇩		
기대도수 $(200\times p_i)$	82	20	8	90

[풀이]

㉠ 가설설정

$H_0 : p_1 = 0.41,\ p_2 = 0.10,\ p_3 = 0.04,\ p_4 = 0.45$

H_1 : 네 가지 확률 중에 적어도 하나는 특정한 확률값과 다르다.

㉡ 카이제곱량 $\chi^2 = \dfrac{(89-82)^2}{82} + \dfrac{(18-20)^2}{20} + \dfrac{(12-8)^2}{8} + \dfrac{(81-90)^2}{90} = 3.70$

㉢ $df = 3$에서 $\chi^2_{0.100} = 6.25 > 3.70$이므로, $p-$값은 0.10보다 더 크다.

㉣ H_0를 기각하기 위한 충분한 증거를 가지고 있지 않다. 즉, 미국의 백인에 대한 혈액형 비율이 앞에 보고된 비율과 다르다고 할 수 없다.

8-3. 독립성 검정(test for independence)

① 두 가지 질적 변수(범주형 변수)가 서로 독립인지 종속인지 판단한다.

② 두 개의 범주형 변수에 따라 실험 단위를 분류하고 이변량 자료를 만들어 분할표 (contingency table)에 기록한 후, 한 가지 분류 방법이 다른 분류 방법에 대해 종속적인지 독립적인지를 결정한다.

③ 두 가지 분류 방법의 독립성에 대한 질문은 카이제곱통계량에 의한 가설검정을 이용하여 검정할 수 있으며 가설은 다음과 같다:

H_0 : 두 가지 분류 방법은 독립적이다.

H_1 : 두 가지 분류 방법은 종속적이다.

④ 만약 적합성 검정처럼 분할표의 관찰도수에 대한 확률 p_{ij}를 안다면, 기대도수 $E_{ij} = np_{ij}$를 구하여 카이제곱통계량을 구할 수 있다. 그러나 그렇지 못하면 기대도수를 추정해야 한다. 기대도수를 추정하기 위한 방법을 설명하려면 독립사건의 개념을 이용한다. 즉, 분할표의 i행 j열의 관찰값의 확률을 p_{ij}라 할 때, 행들과 열들이 서로 독립이면

$$\begin{aligned} p_{ij} &= P(i\text{행 } j\text{열의 관찰값}) \\ &= P(i\text{행의 관찰값}) \times P(j\text{행의 관찰값}) \\ &= p_i p_j \end{aligned}$$

이다. 여기서 p_i와 p_j는 i행과 j열 각각의 주변확률(marginal probability)이 된다. 만약 주변확률의 적당한 추정값을 얻을 수 있다면 기대도수를 계산하는 공식에서 추정된 기대도수를 사용할 수 있는데, 다음과 같다.

⑤ 행확률을 계산하기 위하여

$$\hat{p_i} = \frac{i\text{행의 관찰도수 합}}{\text{관찰도수의 총합}} = \frac{r_i}{n}$$

⑥ 열확률을 계산하기 위하여

$$\hat{p_j} = \frac{j\text{열의 관찰도수 합}}{\text{관찰도수의 총합}} = \frac{c_j}{n}$$

⑦ i행과 j열의 기대도수의 추정값은 독립성의 가정에 따라

$$\widehat{E_{ij}}(\text{추정된 기대도수}_{ij}) = n \times p_i \times p_j = n\left(\frac{r_i}{n}\right)\left(\frac{c_j}{n}\right) = \frac{r_i c_j}{n}$$

이다. r행과 c열로 만들어진 분할표에 대한 카이제곱 검정통계량

$$\chi^2 = \sum \frac{(O_{ij} - \widehat{E_{ij}})^2}{\widehat{E_{ij}}}$$

이며 검정통계량은 $df = (r-1)(c-1)$인 자유도를 갖는 근사 카이제곱분포를 따른다.

[예] 총 $n = 309$개의 가구에서 결점을 확인하였다. 생산자는 결점 형태를 A, B, C, D로 분류하고 가구의 결점 형태를 생산라인에 의해서 분류표로 정리하였다.

	생산라인 (독립변수)			
결점 형태	1	2	3	합계
A	15 (22.51)	26 (22.99)	33 (28.50)	74
B	21 (20.99)	31 (21.44)	17 (26.57)	69
C	45 (38.94)	34 (39.77)	49 (49.29)	128
D	13 (11.56)	5 (11.81)	20 (14.63)	38
합계	94	96	119	309

생산라인에 따라 가구의 결점의 형태가 다르다는 충분한 증거를 자료는 설명하는가?

[풀이]

㉠ 기대도수 구하기 ; 생산라인 2에서 생산된 결점 C의 기대도수의 추정값

$$\widehat{E_{32}} = \frac{r_3 c_2}{n} = \frac{(128)(96)}{309} = 39.77$$

㉡ 가설검정

H_0 : 생산라인에 따라 가구의 결점 형태가 같다.

H_1 : 생산라인에 따라 가구의 결점 형태가 다르다.

㉢ 검정통계량 $\chi^2 = \frac{(15-22.51)^2}{22.51} + \cdots + \frac{(20-14.63)^2}{14.63} = 19.18$

㉣ $df = (r-1)(c-1) = (4-1)(3-1) = 6$인 카이제곱분포에서

$$\text{검정통계량} = 19.18 > \chi^2_{0.005} = 18.5476$$

이고 $p-$값은 0.005보다 작으므로 H_o를 기각하고 매우 유의하다고 결론 내린다. 즉, 결점 형태의 비율은 생산라인에 따라 변한다는 사실에 충분한 증거가 된다.

[예] 새로운 독감 면역주사의 효과를 평가하는 조사가 실시되었다. 면역주사는 2주 기간 동안에 계속 두 번 접종을 하는데 무료로 제공되었다. 1000명의 지역 주민 조사 결과는 아래의 표와 같다. 이 자료는 면역주사가 지역 사회에서 독감 감염자 수를 감소시키는데 성공적이었다는 것을 나타내기 위하여 충분한 증거를 보이는지 설명하고 있는가?

	비접종	한번 접종	두 번 접종	합계
독감 걸림	24	9	13	46
독감 걸리지 않음	289	100	565	954
합계	313	109	578	1000

[풀이]

㉠ 가설설정

H_0 : 면역주사의 회수와 독감 감염여부 사이에 관계가 없다.

H_1 : 면역주사의 횟수와 독감 감염여부 사이에 관계가 있다.

㉡ 관측도수와 기대도수를 계산하여 정리하면 아래와 같다.

관측도수	비접종	한번 접종	두 번 접종	합계
독감 걸림	24	9	13	46
독감 걸리지 않음	289	100	565	954
합계	313	109	578	1000

⇓

기대도수	비접종	한번 접종	두 번 접종	합계
독감 걸림	14.40	5.01	26.59	46
독감 걸리지 않음	298.60	103.99	551.41	954
합계	313	109	578	1000

㉢ 검정통계량 $\chi^2 = 17.313$이므로 $p-$값=0.000이다.

㉣ 따라서 매우 높은 유의성을 가진다고 결론지을 수 있다. 즉, 독감 감염여부는 면역 주사의 횟수에 의존한다는 충분한 증거를 보인다. 그리고

비접종	한번 접종	두 번 접종
24/313=0.08	9/109=0.08	13/578=0.02

두 번 접종을 한 그룹은 독감에 덜 감염되며 한번 접종을 한 그룹은 감염률이 감소하는 것으로 보이지 않는다.

8-4. 동질성 검정(test for homogeneity)

① 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정하는 방법이다.

[예] 남·녀 대학생 생활환경의 동일한 분포는 동일한가?

		기숙사	아파트	부모집	그 외
독립변수	남학생	72	84	49	45
	여학생	91	86	88	35

H_0 : 남학생과 여학생의 생활환경은 동일한 분포를 보인다.

H_1 : 남학생과 여학생의 생활환경은 서로 다른 분포를 보인다.

② 두 개 또는 두 개 이상의 모집단이 어떤 특성을 갖는 분포에 대하여 서로 비슷한 경향을 띠는지를 알아보는 것이다.

③ 카이제곱값을 구하는 방법은 독립성 검정과 동일하고 확률값도 동일하지만, 가설이 다르고 의미가 다르다.

독립성 검정	동질성 검정
표본을 일정량만큼 추출한 다음 행기준과 열기준에 따라 나누어 분석 (예) 300명의 성인들을 랜덤하게 추출한 후 남녀로 나누고 찬성/반대/무응답으로 나눔	행의 합 또는 열의 합, 둘 중 하나에 대해 추출할 표본 수를 각각 정한 후에 추출하고, 나머지 고정 안한 기준에 의해 분류 (예) 행의 인원수를 남자 150명, 여자 150명으로 고정해서 뽑은 후, 이를 찬성/반대/무응답에 따라 나눔

[예] 산아제한에 대한 성인 남·녀간의 찬반 여부에 차이가 있는지를 유의수준을 0.05에서 검정하시오. (단, 남자 모집단 500명과 성인 여자 모집단 600명)

		종속변수		
		남	여	
독립 변수	찬성	230	420	650
	반대	270	180	450
		500	600	1,100

획득도수 : 연구과정에서 얻은 도수
230, 420, 270, 180

주변도수(marginal frequency)
: 650, 450, 500, 600

[풀이]

㉠ 가설 세우기

H_0 : 산아제한에 대해 찬성하는 남녀의 비율은 같다. $p_{남} = p_{여}$.

H_1 : 산아제한에 대해 찬성하는 남녀의 비율은 다르다. $p_{남} \neq p_{여}$

ⓛ 기대도수 구하기

	남	여	
찬성	230	420	650
반대	270	180	450
	500	600	1,100

⇨

	남	여	
찬성	295.45	354.55	650
반대	204.55	245.45	450
	500	600	1,100

- (500×650)/1,100=295.45　　(600×650)/1,100=354.55
- (500×450)/1,100=204.55　　(600×450)/1,100=245.45

ⓒ χ^2통계값 구하기

χ^2통계값

$$=\frac{(230-295.45)^2}{295.45}+\frac{(420-354.55)^2}{354.55}+\frac{(270-204.55)^2}{204.55}+\frac{(180-245.45)^2}{245.45}=64.98$$

ⓓ 결과 해석하기

자유도=(2-1)(2-1)=1, 유의수준 0.05인 χ^2임계값=3.84

따라서 χ^2통계값 64.98은 기각값 3.84보다 크므로 귀무가설을 기각하게 된다. '유의수준 0.05에서 산아제한에 대한 남녀집단의 찬성비율은 같지 않다'

【참고】 Karl Pearson이 제안한 χ^2검정의 기본 원리는?

① 남녀 전체 산아제한에 대한 찬성비율은 650/1,100=59.09%, 반대비율은 450/1,100=40.9%이다. 따라서 남녀집단의 산아제한에 대한 찬성비율이 차이가 없으려면, 즉 귀무가설이 사실이려면

ⓐ 성인 남자 표본에서도 59.09%($=500\times\frac{650}{1,100}$)가 산아제한에 대하여 찬성

ⓛ 성인여자 표본의 59.09%($=600\times\frac{650}{1,100}$)에 해당하는 사람들이 산아제한에 대하여 찬성

ⓒ 나머지는 반대비율

② 귀무가설이 사실이라면 각 칸에서의 획득도수와 기대도수의 차이가 없다. 반대로 획득도수가 귀무가설 하에서 추정된 기대도수와 차이가 심할수록 귀무가설과 멀어지는 결과를 가져오게 된다.

③ Karl Pearson이 이 원리를 이용하여 χ^2통계값을 계산하는 공식을 제안하였다.

$$\chi^2=\sum\frac{(\text{획득도수}-\text{기대도수})^2}{\text{기대도수}}$$

즉, 모든 칸에서 획득도수에서 기대도수를 뺀 다음 제곱하여 각 칸에 해당되는 기대도수로 나눈 후 모든 칸의 값을 더한 값이다.

④ χ^2통계값이 클수록 H_0을 부정하게 되어 집단간의 차이가 있음을 알 수 있다.