범주형 자료의 분석

범주형 자료의 관련성

- **범주로 측정된 두 변수 사이의 관련성을 어떻게 입증할 수 있을까?**
- **두 변수가 독립이 아니라는 것을 어떻게 보일 수 있을까?**
- **범주별로 '실제로 관찰된 빈도'와 '독립일 때 기대되는 빈도'의 차이를 살피는 것이 분석의 시작**

연구 문제 9-1

- **흡연은 폐암과 관련이 있을까?**
- **흡연이라는 위험인자에 대한 노출은 폐암의 발생을 증가시키는가?**

		질병 발생 여부		
		발생(+)	미발생(–)	합 계
위험 인자 노출 유무	예(+)	A=30	B=70	A+B=100
	아니오(–)	C=10	D=90	C+D=100
합 계		A+C=40	B+D=160	A+B+C+D=200

- **위험인자에 대한 노출 유무'가 '질병의 발생 여부'와 관련이 있다고 말할 수 있는가? 유의수준 0.05에서 검정하라.**

교차분석표

◆ **교차분석표(cross-tabulation) : 각 범주와 범주에 해당하는 관찰빈도를 나타낸 표**

		Y		
		Y=0	Y=1	계
X	X=0	$A=N(X=0 \cap Y=0)$	$B=N(X=0 \cap Y=1)$	$(A+B)=N(X=0)$
	X=1	$C=N(X=1 \cap Y=0)$	$D=N(X=1 \cap Y=1)$	$(C+D)=N(X=1)$
	계	$(A+C)=N(Y=0)$	$(B+D)=N(Y=1)$	A+B+C+D

교차분석표를 통한 두 변수간 독립성 검정

◆ 만일 X와 Y가 독립일 경우 빈도 (기대빈도)는?

		Y		
		Y=0	Y=1	계
X	X=0	A′	B′	*N*(X=0)
	X=1	C′	D′	*N*(X=1)
	계	*N*(Y=0)	*N*(Y=1)	

● A′=P(X=0 ∩Y=0)×(전체 빈도합)= P(X=0) × P(Y=0) ×(전체 빈도합)
=*N*(X=0)/(*N*(X=0)+*N*(X=1)) × *N*(Y=0)/(*N*(Y=0)+*N*(Y=1)) × (전체 빈도합)

관찰빈도(O_{ij})	폐암발생	폐암 미발생	계
흡연	30	70	100
비흡연	20	80	100
계	50	150	200

기대빈도(E_{ij})	폐암발생	폐암미발생	계
흡연	25	75	100
비흡연	25	75	100
계	50	150	200

● A′=P(흡연 ∩폐암발생) ×200= P(흡연) × P (폐암발생) ×200
=100/200 ×50/200 ×200=25

교차분석표를 통한 두 변수간 독립성 검정

- **관찰빈도와 기대빈도가 비슷하다면?**
  - 관찰빈도가 독립을 가정한 기대빈도와 비슷하므로 독립이라고 결론
- **관찰빈도와 기대빈도가 다르다면(차이가 많이 난다면)?**
  - 관찰빈도가 독립을 가정한 기대빈도와 차이가 많이 나므로 독립이 아니라고 결론
- **귀무가설?**
  - 두 변수는 독립이다.
- **그럼 차이의 정도는 어떻게 보나?**
  - Chi-square 통계량
- **식은?**
  - $\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum\frac{(\text{관찰빈도}-\text{기대빈도})^2}{\text{기대빈도}}$
- **폐암 발생에 관한 교차분석표의 χ^2 값은?**
  - $(30-25)^2/25 + (70-75)^2/75 + (20-25)^2/25 + (80-75)^2/75 = 1 + 1/3 + 1 + 1/3 \simeq 2.6667$ <- 차이의 정도

교차분석표를 통한 두 변수간 독립성 검정

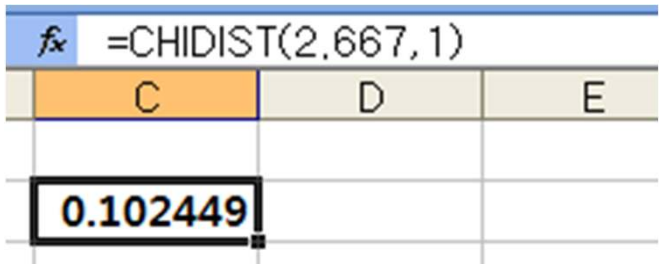
◆ **폐암 발생에 관한 교차분석표의 χ^2 값은?**

- $(30-25)^2/25 + (70-75)^2/75 + (20-25)^2/25 + (80-75)^2/75 = 1 + 1/3 + 1 + 1/3 \simeq 2.6667$ ← 차이의 정도

◆ **그러면 이 정도로 차이날 확률은 얼마나 되나 (p-value)?**

- CHIDIST (χ^2 값, 자유도)
- CHIDIST (2.6667, 1)
- m ×n 교차분석표의 자유도는 (m-1) ×(n-1)

◆ **유의수준 5%에서 가설 검정 결과는?**



독립성 검정 - 연습

연구 문제 9–1

◆ **위험인자에 대한 노출 유무'가 '질병의 발생 여부'와 관련이 있다고 말할 수 있는가? 유의수준 0.05에서 검정하라.**

		질병 발생 여부		
		발생(+)	미발생(–)	합 계
위험 인자 노출 유무	예(+)	A=30	B=70	A+B=100
	아니오(–)	C=10	D=90	C+D=100
합 계		A+C=40	B+D=160	A+B+C+D=200

독립성 검정 - 풀이

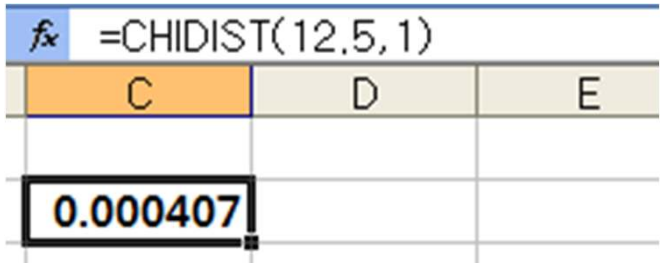
- **귀무가설: '위험인자에 대한 노출 유무'는 '질병의 발생 여부'독립이다(관련이 없다).**
- **대립가설: '위험인자에 대한 노출 유무'는 '질병의 발생 여부'독립이아니다(관련이 있다).**

		질병 발생 여부		
		발생(+)	미발생(–)	합 계
위험 인자 노출 유무	예(+)	0.5×0.2 = 0.1	0.5×0.8 = 0.4	100/200 = 0.5
	아니오(–)	0.5×0.2 = 0.1	0.5×0.8 = 0.4	100/200 = 0.5
합 계		40/200 = 0.2	160/200 = 0.8	200/200 = 1

		질병 발생 여부		
		발생(+)	미발생(–)	합 계
위험 인자 노출 유무	예(+)	0.1×200 = 20	0.4×200 = 80	0.5×200 = 100
	아니오(–)	0.1×200 = 20	0.4×200 = 80	0.5×200 = 100
합 계		0.2×200 = 40	0.8×200 = 160	1×200 = 200

교차분석표 연습1 풀이

- 교차분석표의 $\chi^2 = (30-20)^2/20 + (70-80)^2/80 + (10-20)^2/20 + (90-80)^2/80 = 5 + 1.25 + 5 + 1.25 = 12.5$
- P-value = 0.000407



FISHER의 정확한 검정(FISHER'S EXACT TEST)

◆ χ^2 검정 통계량은 기대빈도가 작아질수록 급격하게 값이 증가됨

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum\frac{(\text{관찰빈도}-\text{기대빈도})^2}{\text{기대빈도}}$$

Cochran's Rule
범주형 자료의 $m \times n$ 분할표에서, 모든 셀의 기대빈도가 5보다 커야 한다. 이를 흔히 5의 법칙이라고 한다. **관찰빈도가 아니라 기대빈도라는 것에 유의**하여라.

◆ 그러면 기대빈도가 5이하인 경우는 어떻게 해야 하나?

- 통계분석에 'exact test' 옵션을 선택

◆ 범주를 합할 수 있는 경우 (합해도 의미의 변화가 미미한 경우)

- 범주를 합하여 기대 빈도가 5를 초과하게 조정

적합성 검정

◆ 하나의 변수가 갖는 범주별 빈도가 동일한지 검정

- ●귀무가설: 범주별로 동일
- ●대립가설: 범주별로 다름

◆ 한국 의료관광을 이용한 외국인의 의료이용 분야

- ●분야별로 이용에 차이가 있는가?

C2 fx =CHITEST(D5:G5,D6:G6)

	A	B	C	D	E	F	G	H
1								
2			0.00030					
3								
4			건강/치료	피부미용	건강검진	한방	기타	행합계
5			관찰빈도	3	10	1	16	30
6			기대빈도	7.50	7.50	7.50	7.50	30