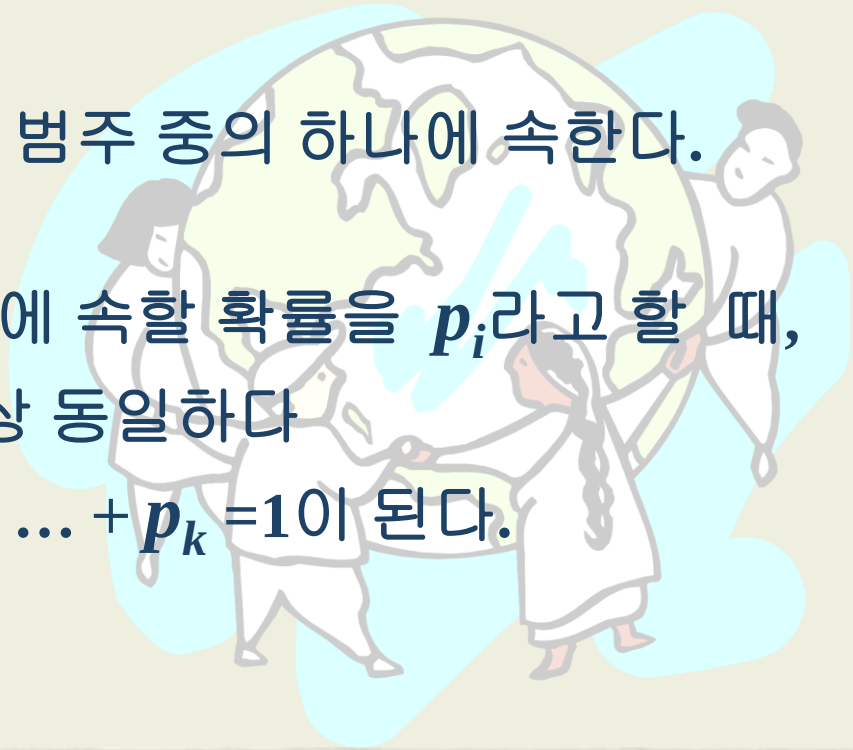
제 11장

범주자료에 대한 χ^2검정

고려대학교 경영대학 박 광태

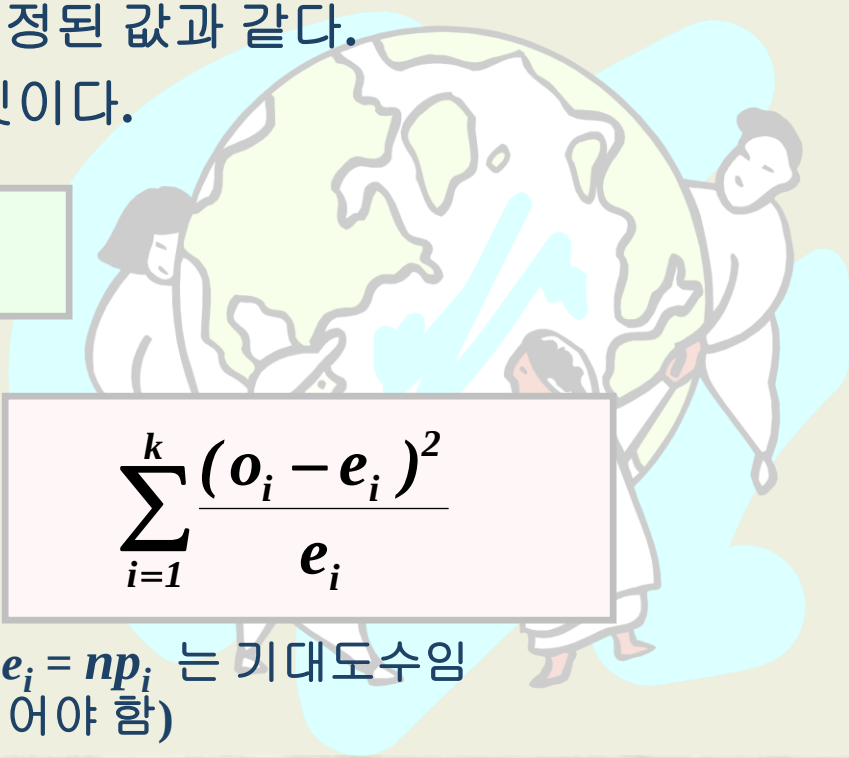
다항분포의 조건

- 다항분포의 조건
  - n번의 다항실험을 시행한다.
  - 각 시행의 결과는 k개의 범주 중의 하나에 속한다.
  - 실험결과가 i 번째 범주에 속할 확률을 p_i라고 할 때,
    매 시행마다 p_i 는 항상 동일하다
    여기서 $p_1 + p_2 + p_3 + \ldots + p_k = 1$이 된다.
  - 각 시행은 독립적이다.



적합성 검정

- 적합성 검정
  - 셋 이상의 모비율에 대한 가설검정.
  - $H_0 : P_1, P_2, \ldots, P_k$가 미리 설정된 값과 같다.
    $H_1 : H_0$중 적어도 하나가 거짓이다.
  - (기각치) $\chi^2_{k-1,\alpha}$
  - (피어슨 검정통계량) $\sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$



O_i는 관찰도수이고 $e_i = np_i$ 는 기대도수임
(기대도수는 5이상이어야 함)

만일 $k=2$이면, $n_2=n-n_1$이고 $p_2=1-p_1$이 된다. 그러므로 n이 충분히 클 때($np_1 \geq 5$, $np_2 \geq 5$)

$$\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}=\frac{(n_1-np_1)^2}{np_1}+\frac{(n_2-np_2)^2}{np_2}$$

$$=\frac{(n_1-np_1)^2}{np_1}+\frac{[(n-n_1)-n(1-p_1)]^2}{n(1-p_1)}$$

$$=\frac{(n_1-np_1)^2}{np_1}+\frac{(-n_1+np_1)^2}{n(1-p_1)}$$

$$=\frac{(1-p_1)(n_1-np_1)^2+p_1(-n_1+np_1)^2}{np_1(1-p_1)}$$

$$=\frac{(n_1-np_1)^2}{np_1(1-p_1)}=\left(\frac{n_1-np_1}{\sqrt{np_1(1-p_1)}}\right)^2\approx Z^2=x_1^2$$

적합성 검정

예제 11-2

시장에서 판매되는 네 가지 제품에 대한 시장점유율이 $p_1=0.2$, $p_2=0.3$, $p_3=0.4$, $p_4=0.1$이라는 것을 검정하기 위하여 시중에 판매된 200개의 제품을 관찰하였더니 다음과 같았다.

제 품	1	2	3	4
판매량(n_i)	55	65	72	8

유의수준 0.05에서 다음의 가설을 검정하시오.

H_0: $p_1=0.2$, $p_2=0.3$, $p_3=0.4$, $p_4=0.1$

H_A: 귀무가설 중 적어도 하나는 사실이 아니다.

적합성 검정

- 예제 11-2(계속)

적합성 검정

- 예제 11-2(계속)

예제 11-3

다음의 자료는 한 교차로에서 주중 발생하는 사고의 횟수(Y)를 $n=50$주 동안 관찰한 결과이다. 이를 이용해 확률변수 Y가 포아송분포를 갖는지를 유의수준 0.05에서 검정하시오.

주간사고 발생수(y)	0회	1회	2회	3회 이상
빈도(n_i)	32주	12주	6주	0

예제 11-4

통계학 교수는 학생들의 시험성적을 처리하는 데 있어서 학생들의 성적 X가 정규분포를 따른다는 가정을 전제로 한다. 표 11-1은 수강했던 학생들 50명의 성적이다. 이 자료로부터 변수 X가 정규분포를 따른다는 가정을 유의수준 0.05에서 검정하시오.

표 11-1 50명 학생의 통계학 성적

71	66	61	65	54
93	60	86	70	70
73	73	55	63	56
62	76	54	82	79
76	68	53	58	85
80	56	61	61	64
65	62	90	69	76
79	77	54	64	74
65	65	61	56	63
80	56	71	79	84

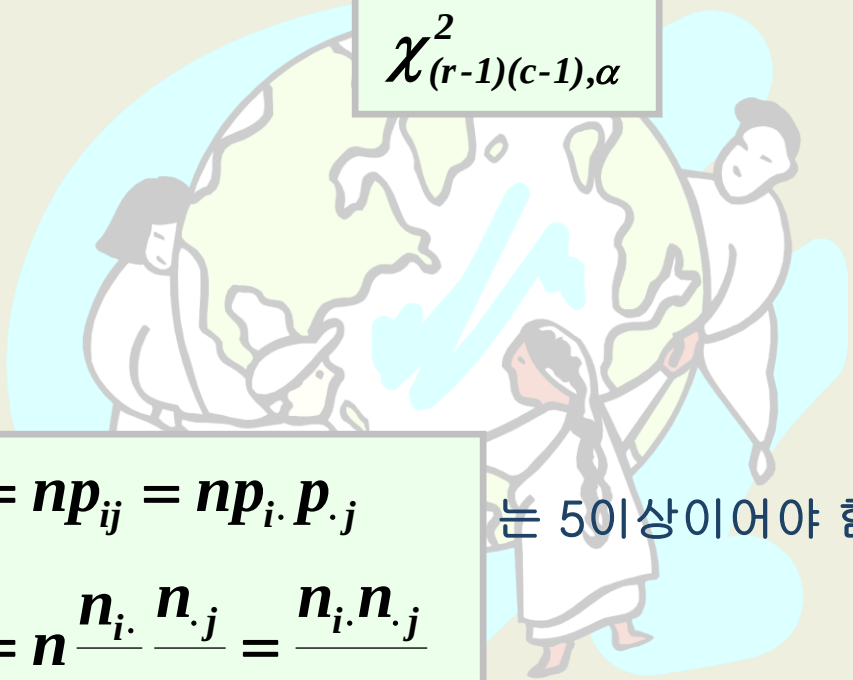
독립성 검정

- 독립성 검정
  - 두 분류변수의 독립성 검정에 이용
  - H_0 : A 와 B는 상호독립이다.
    H_1 : H_0는 사실이 아님
  - (검정통계량)

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}$$

  - (기각치)

$$\chi^2_{(r-1)(c-1),\alpha}$$



(여기서 기대도수

$$E(n_{ij}) = np_{ij} = np_{i\cdot}p_{\cdot j}$$
$$= n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{i\cdot}n_{\cdot j}}{n}$$

는 5이상이어야 함)

독립성 검정

예제 11-5

개인의 소득과 정치적 성향 사이에 관계가 있는지를 조사하려고 한다. 표 11-4는 150명을 임의로 추출하여 정치적 성향(친여 또는 친야)과 소득수준(저소득층, 중간층 또는 고소득층)에 따라 분류한 분할표이다. 유의수준 0.05에서

H_0: 정치적 성향과 소득수준은 무관하다.

H_A: H_0는 사실이 아니다.

를 검정하시오.

표 11-4 예제 5에 대한 분할표

정치적 성향 \ 소득수준	고소득층	중간층	저소득층	계
친여	45	30	15	90
친야	5	20	35	60
계	50	50	50	150

그림 11-7에서와 같이 예제 5의 자료에 대해 제목, 열의 이름과 열총합, 행의 이름 및 행총합을 입력한다. 관찰도수표의 열총합을 산출하기 위해서는 그림의 메뉴바상에 있는 Σ 아이콘을 클릭하여 계산하면 된다.

그림 11-7 관찰도수표의 작성

그리고 기대도수표의 작성을 위해서는 그림 11-8과 같이 기대도수표를 계산하기 위한 표를 우선 만든다. 방법은 관찰도수표를 만드는 것과 동일하다. 열총합과 행총합은 관찰도수표에서 산출된 수치를 그대로 입력하면 된다.

이제 기대도수표에 들어갈 수치는 다음 수식을 입력하여 얻는다.

기대도수 = 열총합 * 행총합 / 전체총합

예를 들어 B13셀에 들어갈 기대도수는 『=B8 * E6/150』식을 입력하면 된다. 참고로 여기서 $표시는 엑셀의 셀에서 절대참조를 의미하고 $표시가 없는 경우는 상대참조를 의미하는데 위의 표현은 B13셀의 수식을 다른 셀에 복사하기 위함이다.

이제 B13의 수식을 B13:D14에 복사하면 되는데 요령은 B13셀을 지정한 뒤 마우스를 그 셀의 오른쪽 아래에 갖다 놓으면 +가 생기게 된다. 이 +를 복사할 나머지 셀범위에 드래그하면 자동복사되어 기대도수가 자동생성된다. 그 결과는 그림 11-8과 같다.

그림 11-8 기대도수표의 작성

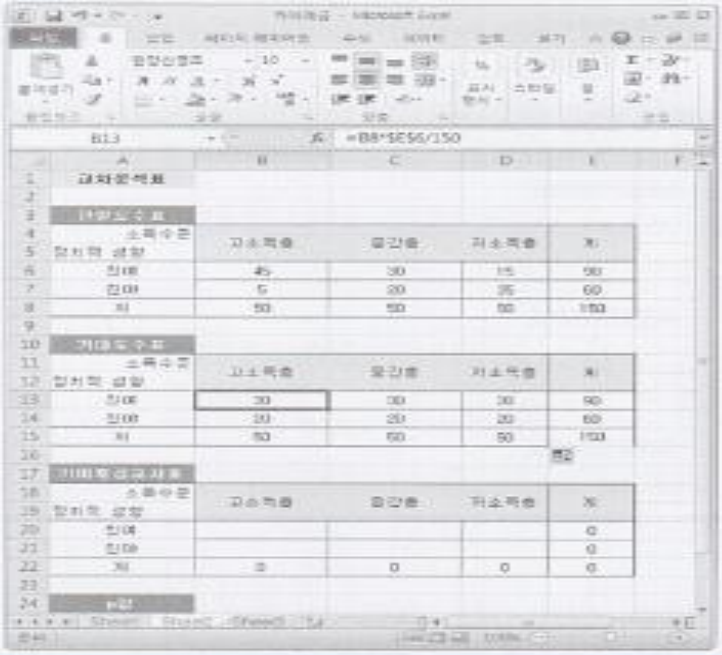
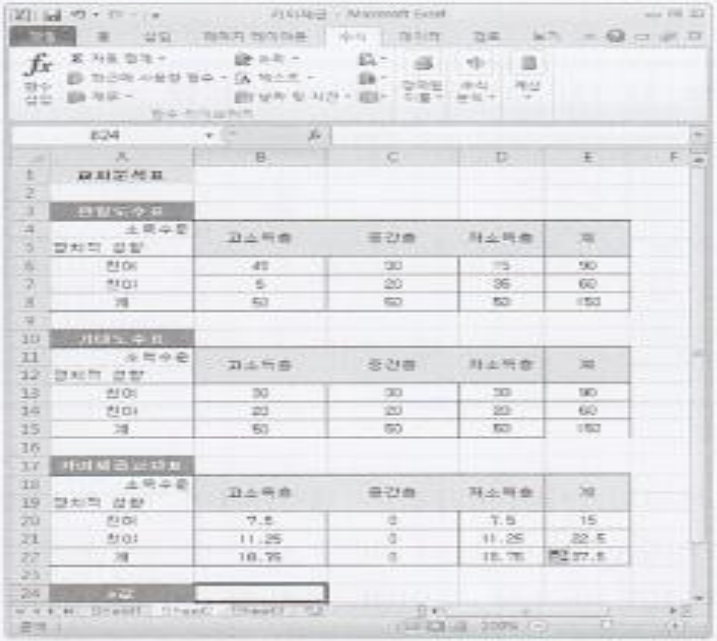


그림 11-9에서와 같이 카이제곱값을 산출하기 위해 다음의 식을 이용하기로 하자.

$$\chi^2 = \sum_i \sum_j \frac{(\text{실제빈도}_{ij} - \text{기대빈도}_{ij})^2}{\text{기대빈도}_{ij}}$$

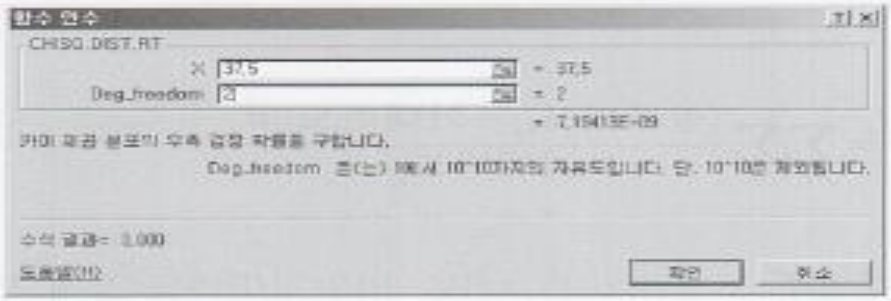
위 식에 각각의 셀을 대입해 보면 B20은 =(B6-B13)^2/B13을 통해 구할 수 있으며, 이것을 다른 셀에 복사하여 전체 37.5의 카이제곱값을 구할 수 있다.

그림 11-9 카이제곱값의 산출



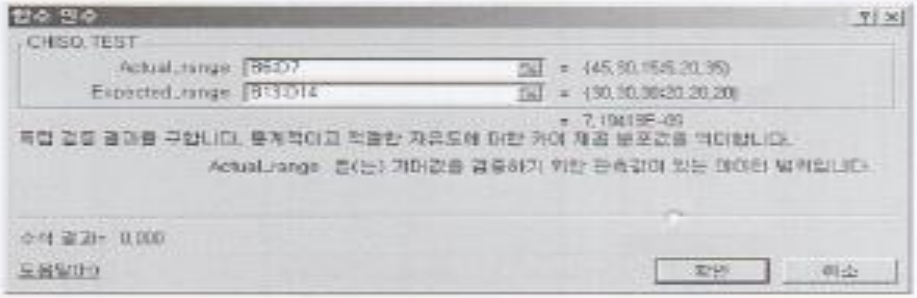
그리고 그림 11-10과 같이 유의확률값인 p값을 구하기 위해 카이제곱값 37.5를 이용하여 =CHISQ.DIST.RT(37.5,2)를 계산하면 p값은 0.000임을 알 수 있다.

그림 11-10 유의확률 p값의 산출



또는 기대도수표 작성 후 카이제곱교차표를 작성하지 않고 그림 11-11과 같이 =CHISQ.TEST(B6:D7, B13:D14)를 이용하여 p값 0.000을 구할 수 있다.

그림 11-11 CHITSQ.TEST를 이용한 p값의 산출



교차분석결과 카이제곱값은 37.5이며 이에 대한 유의확률 p값은 0.000이다. 이는 유의수준 0.05보다 작기 때문에 두 분류기준이 서로 독립이라는 귀무가설은 기각되고 따라서 개인의 정치적 성향과 소득수준은 독립이 아니라고 결론내릴 수 있다.