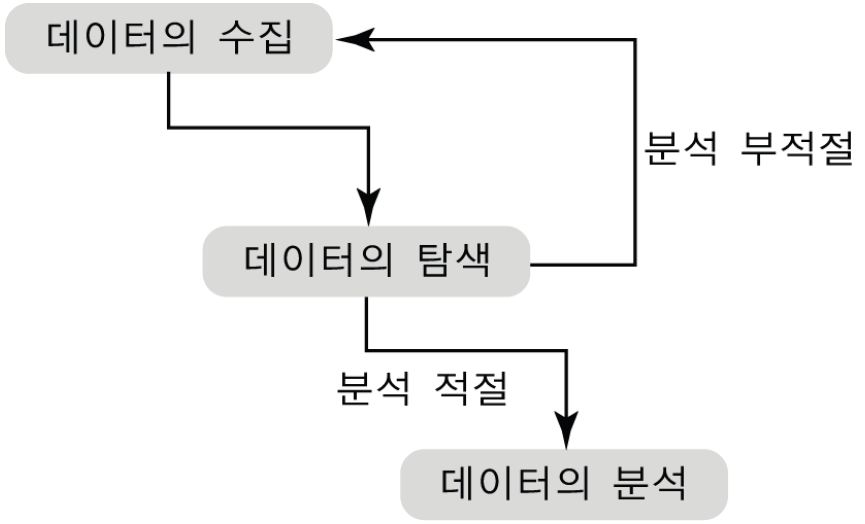
PART 2. 기초데이터분석

Chapter 4. 데이터탐색

1. 데이터탐색과 데이터분석

1.1. 데이터탐색과 데이터분석의 관계

- 데이터 수집, 데이터 탐색 데이터 분석



- 주요 데이터탐색 내용
  - 변수들에 대한 데이터탐색을 통한 정규성 검정
  - 대상변수들간의 상관관계 분석
  - 그룹별 대상변수의 데이터탐색을 통한 정규성 검정
  - 분석에 영향을 줄 수 있는 특이 관찰치(outlier)의 파악
  - 결측값(missing data)이 있는 경우 이에 대한 평가 및 결측값의 처리 및 활용

변수의 척도	변수의 개수	분석 방법
넌메트릭 데이터 (명목, 서열척도)	1	• 빈도분석 • 다중 응답 문항의 경우 다중응답에 대한 빈도분석 • 막대도표, 원도표 등을 통한 도표분석
	2	• 교차분석 • 다중 응답 문항의 경우 다중응답에 대한 교차분석 • 넌메트릭 척도의 상관관계분석(서열척도)
	3개 이상	• 넌메트릭 척도에 대한 신뢰도 검정
메트릭 데이터 (등간, 비율척도)	1	• 정규성 검정 • 평균과 표준편차 분석 • 넌메트릭 척도로 전환하여 빈도분석이나 도표분석 • 히스토그램을 포함한 정규분포 검정
	2	• 두 변수간 상관관계분석 • 두 변수간 산점도 분석
	3개 이상	• 메트릭 데이터 신뢰도 검정

1.2. 데이터분석의 구분

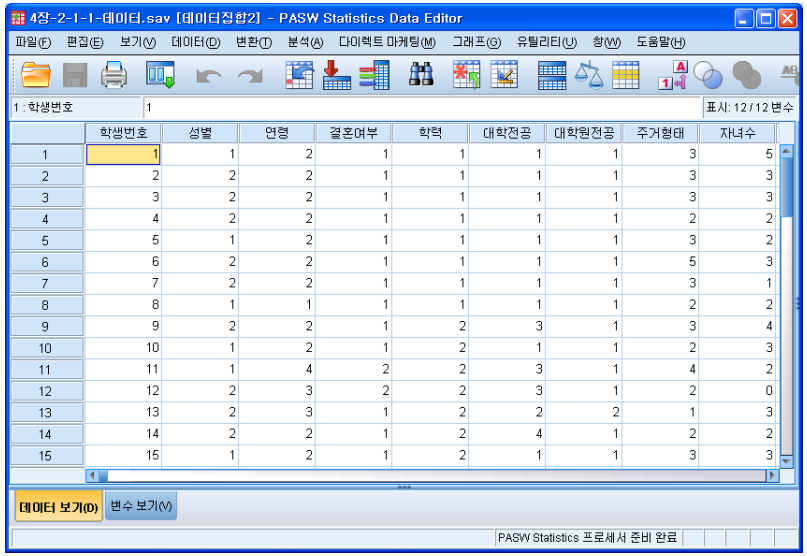
- 변수의 수(p.119)
  - 단일변량 데이터분석
  - 다변량 데이터분석
- 분석의 성격
  - 종속관계분석
  - 상호의존관계분석
- 척도의 종류
  - 넌메트릭 데이터: 비모수통계분석
  - 메트릭 데이터: 모수통계분석
- 표본집단의 수와 관계
  - 단일표본
  - 다표본

2. 빈도분석

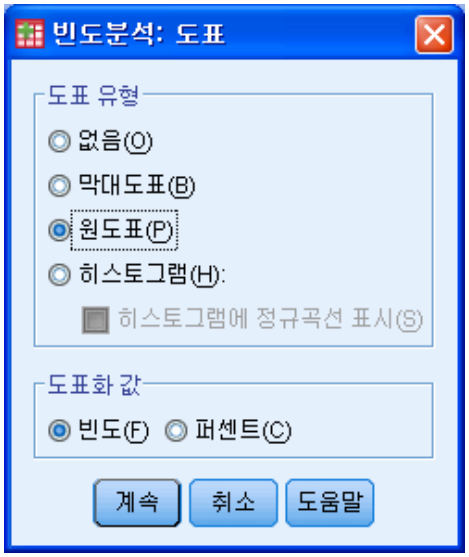
2.1. 넌메트릭 데이터의 빈도분석

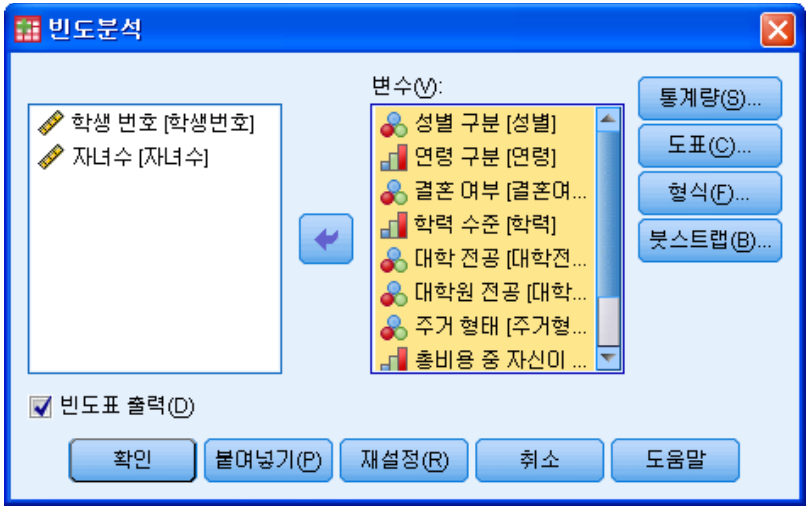
- 분석개요(p.121)
  - 38명의 응답자에 대한 빈도분석
- 분석데이터
  - 텍스트 데이터를 읽어 들여 변수 지정을 함
  - *.txt ⇒ *.sav 파일

01	12111135652	02	22111133422	03	22111133323	04	22111122323	05	12111132621
06	22111153631	07	22111131453	08	11111122641	09	22123134425	10	12121123625
11	14223142631	12	23223120613	13	23122213634	14	22124122634	15	12121133625
16	23122140621	17	22124133521	18	23221120431	19	22111132412	20	22111133422
21	12111133612	22	12111132224	23	12111125231	24	12111132121	25	21111123431
26	11111133112	27	21111122332	28	22114133411	29	14223142654	30	21111123224
31	12111124632	32	11111124211	33	11111124211	34	24221140423	35	23224121544
36	22221113645	37	22121126514	38	23324121654				



- 분석과정
  - STEP 01: [빈도분석] 메뉴를 선택
  - STEP 02-03: 빈도분석 변수를 선택
  - STEP 04: 원도표 선택

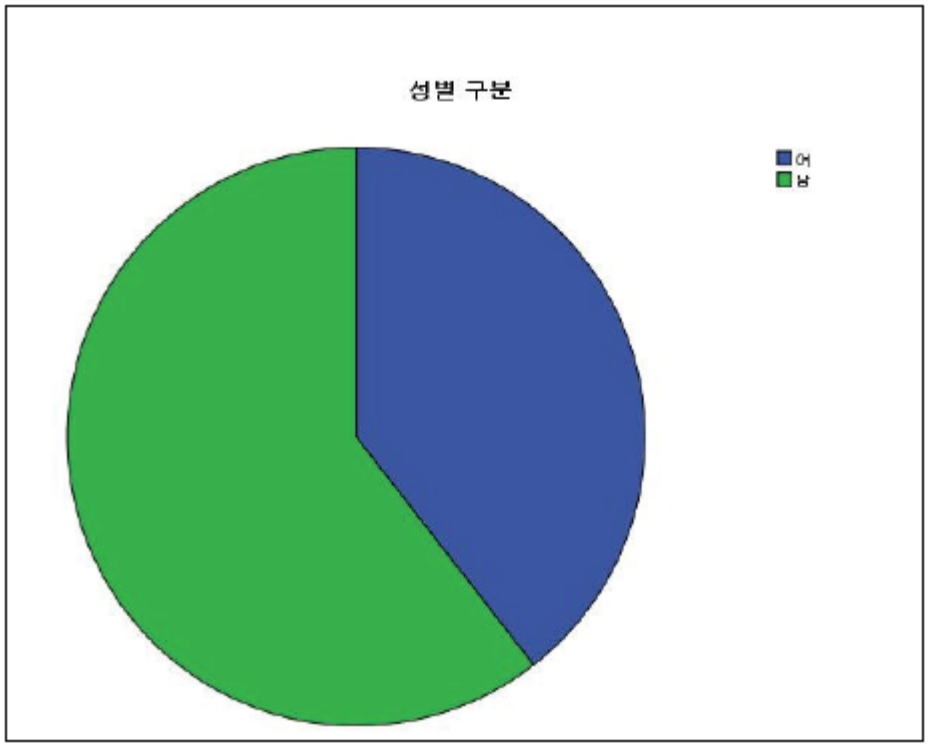




- 결과해석

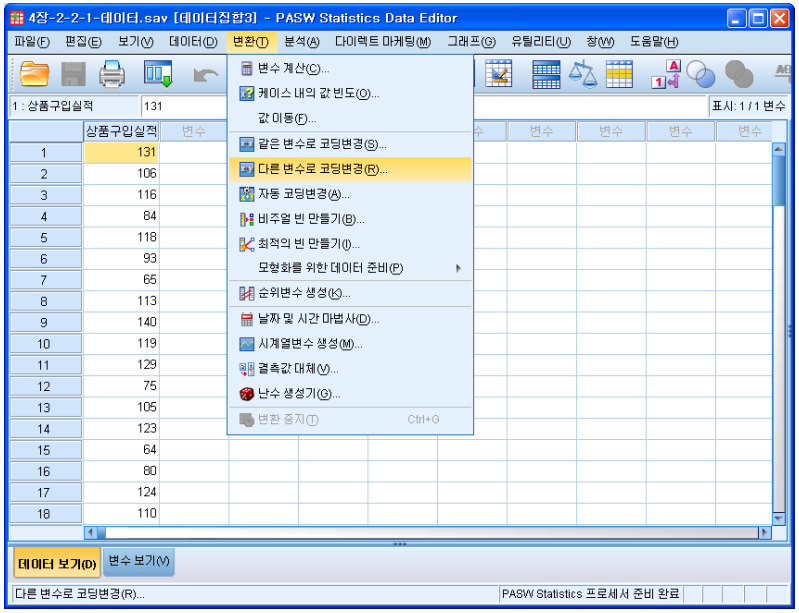
성별 구분

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	여	15	39.5	39.5	39.5
	남	23	60.5	60.5	100.0
	합계	38	100.0	100.0	

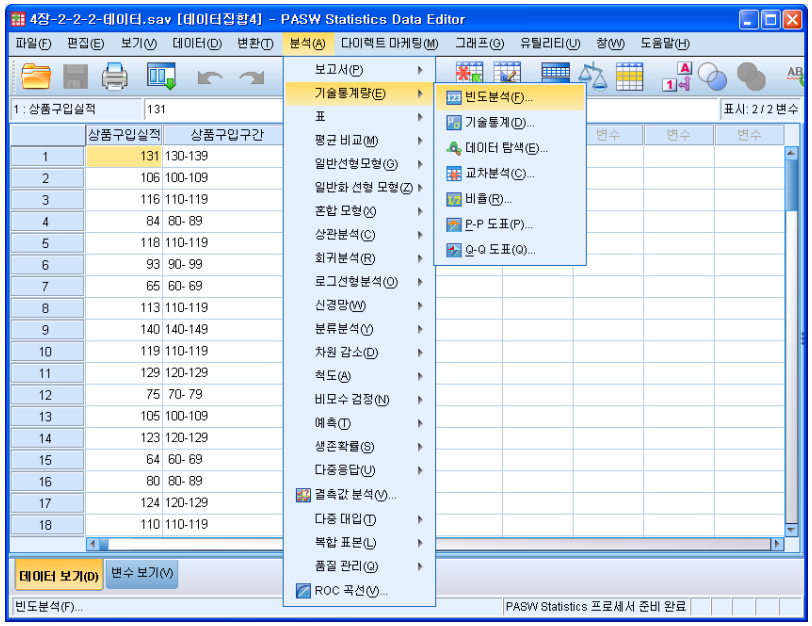


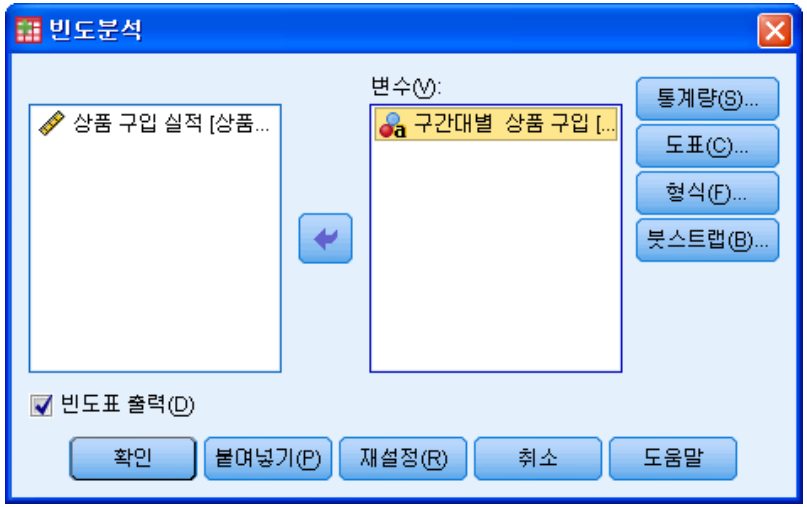
2.2. 메트릭 데이터의 빈도분석

- 분석개요
  - 적절한 계급수를 지정
- 분석데이터
  - STEP 01: 분석데이터 입력
  - STEP 02-06: 코딩변경을 통해 코딩변경



- 분석과정
  - STEP 01: [빈도분석] 메뉴를 클릭
  - STEP 02-03: 변수 지정
  - STEP 04: 도표 버튼을 클릭해 막대도표 지정

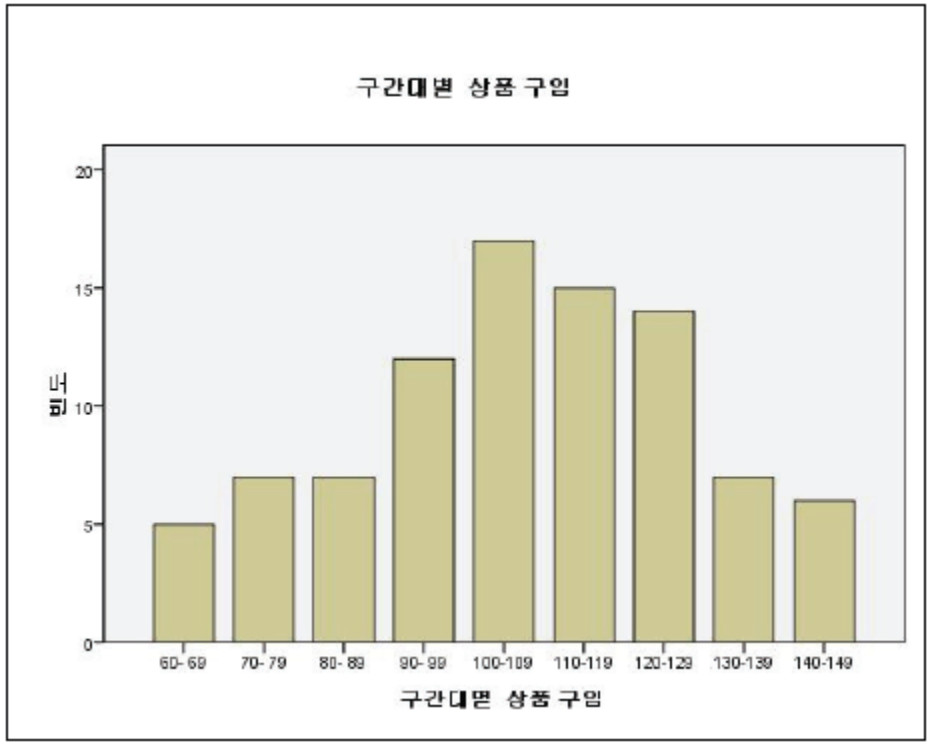




- 결과해석

구간대별 상품 구입

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	60- 69	5	5.6	5.6	5.6
	70- 79	7	7.8		
	80- 89	7	7.8		
	90- 99	12	13.3		
	100-109	17	18.9		
	110-119	15	16.7		
	120-129	14	15.6		
	130-139	7	7.8		
	140-149	6	6.7		
	합계	90	100.0		



2.3. 교차분석

<note> 교차 분석

1. 목적

⇒ 교차분석은 명목 및 서열(순위)척도를 통하여 수집된 범주형 변수들(categorical variables)을 분석한다.

→ 즉, 한 변수의 범주를 다른 변수의 범주에 따라 빈도를 교차 분류하는 교차표 또는 분할표(cross tabulation))를 작성하여, 두 변수들 사이의 독립성과 관련성을 분석할 수 있다.

< 참고 > 독립성과 관련성

① 관련성(Association) ⇒ 한 변수의 값이 변함에 따라 다른 변수의 값이 변하면, 두 변수 사이에 관련성이 있다고 한다.

② 상호 독립적(independent) ⇒ 두 변수들 사이에 관련성이 없다면 한 변수의 값이 변해도 다른 변수의 값에는 영향을 미치지 못한다. 이러한 경우 두 변수는 상호 독립적이라 한다.

2. 가설 검정

⇒ 교차표를 이용한 두 범주의 독립성 검정에 대한 검정절차는 다음과 같다.

1) 가설설정

H_0 : 두 개의 범주는 서로 독립이다.
(두개의 범주는 아무런 관련성이 없어 서로에게 영향을 미치지 않는다.)

2) 검정통계량

$\chi^2 = \sum\sum (O_{ij} - E_{ij})^2 / E_{ij}$ 이며,
이것은 자유도가 (r-1)(c-1)인 χ^2 분포를 따른다.

여기서, r = 행의 수, c = 열의 수
O_{ij} = 실제 빈도
E_{ij} = 기대 빈도

3. 교차(분할)표

⇒ 교차표 또는 교차분할표는 두 종류의 특성 X, Y를 각각 구분하고, 그 특성의 구분을 조합한 이원분류를 생각하여, 그 각 구분마다에 해당하는 단위 수를 계산한 이차원의 도수분포표에 의해 두 변수 간의 관계를 나타내는 방법이다.

→ 즉, 두 변수 간의 관련성 분석을 위해 한 변수는 열(column)에, 다른 변수는 행(row)에 배치하여 행과 열이 교차하는 칸(cell)에 필요한 빈도(frequency), 백분율(percentage) 등을 나타낸 표로써, 이원교차표라고도 한다.

■ 교차표의 이해

→ 다음은 Lenski(1963)의 연구 결과로서, 인종과 선호하는 정당에 대한 교차표이다.

	선호하는 정당			
인종-종교 그룹	민주당	공화당	기타	주변합
백인 개신교	235	249	140	624
백인 천주교	274	92	113	479
유태교	32	1	15	48

① 백분율이 아닌 관찰치를 직접 나타내는 경우

⇒ 각 칸의 빈도를 직접 비교할 수 없다.

(예) 민주당을 지지하는 백인 개신교도가 235명인데 반해 유태인은 32명이라고 해서 백인 개신교도가 유태인에 비해 민주당을 더 선호한다고 주장할 수 없다.

	선호하는 정당			
인종-종교 그룹	민주당	공화당	기타	주변합
백인 개신교	(235)	(249)	(140)	(624)
	37.7%	39.9%	22.4%	100%
백인 천주교	(274)	(92)	(113)	(479)
	57.2%	19.2%	23.6%	100%
유태교	(32)	(1)	(15)	(48)
	66.7%	2.1%	31.2%	

100% → 행 방향으로 계산

② 이처럼 각 그룹별 표본 수가 같지 않은 경우, 불균형을 극복하기 위해 교차표를 백분율로 나타내는 방법을 택한다.

<참고> 고려해야 할 사항 ⇒ 백분율이 계산된 방향이

① 행으로 계산 된 것인지
② 열로 계산된 것인지
③ 전체를 대상으로 계산 된 것인지

에 따라 교차표를 읽는 방향이 달라진다.

(예) 위의 표를 보면, 선택된 표본에서

① 백인 개신교도의 37.7%가 민주당을, 39.9%가 공화당을
22.4%가 기타 정당을 선호한 것으로 나타났다.

② 경우 백인 개신교도의 33.7%가, 백인 천주교도의 57.2%가,
유태교도의 66.7%가 민주당을 지지하는 것으로 나타났다.

4. 예제 :

1) 교차표 작성

⇒ OO사는 자사의 직무성과에 관한 보상방법에 대하여 420명을 대상으로 조사를 실시하였다. 즉, 보상 방법에 있어 기존의 방식과 새로운 방식에 대해 응답자의 선호도에 대한 반응이 어떻게 다른지 조사하였다.

→ 아래 표는 두 개의 행(row)과 네 개의 열(column)로 구성되어 있다.

보상방식 지역	서울	부산	대구	인천	합계
기존 방식을 선호	68	75	57	79	279
새로운 방식 선호	32	45	33	31	141
합 계	100	120	90	110	420

2) 실제빈도와 기대빈도 값의 계산

(1) 실제빈도 ⇒ 교차표의 각 cell 내에 기록 된 실제 관측 값.

(2) 기대빈도

⇒ 각 cell의 기대빈도 값은 다음과 같이 구할 수 있다.

① '서울'과 '기존 방식을 선호'에 대한 기대빈도

→ 전체 인원이 420명이고,

'서울'이 100명, '기존 방식 선호'가 279명이므로,

이 cell (1행1열)의 기대빈도는 다음과 같다.

$E_{11} = (100 \times 279) / 420 \fallingdotseq 66$

② 같은 방법으로,

'부산'과 '기존 방식을 선호'에 대한 기대빈도는

$E_{12} = (120 \times 279) / 420 \fallingdotseq 80$

③ 같은 방법으로,
'서울'과 '새로운 방식을 선호'에 대한 기대빈도는

$E_{21} = (100 \times 141) / 420 \fallingdotseq 34$

와 같은 방법으로 나머지 cell의 기대빈도를 계산한다.

보상방식 지역	서울	부산	대구	인천	합계
기존 방식을 선호	68(66)	75(80)	57(60)	79(73)	279
새로운 방식 선호	32(34)	45(40)	33(30)	31(37)	141
합 계	100	120	90	110	420

( )안은 기대빈도

3) 귀무가설의 채택과 기각의 원리

① 실제빈도와 기대빈도의 차이가 크면
→ 검정 통계량인 χ^2 값이 커지므로
→ 귀무가설이 기각될 가능성이 높아 진다.

따라서, 두개의 범주가 서로 밀접한 관계가 있을 가능성이 높아진다.

② 실제빈도와 기대빈도의 차이가 작으면
→ 검정 통계량인 χ^2 값이 작아지므로
→ 귀무가설이 기각될 가능성이 낮아 진다.

따라서, 두개의 범주가 서로 밀접한 관계가 있을 가능성이 낮아진다.

4) χ^2 통계량의 계산

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

이므로,

$$\chi^2 = (68-66)^2 / 66 + (75-80)^2 / 80 + \ldots + (31-37)^2 / 37$$
$$= 3.032$$

이다.

5) 자유도 계산

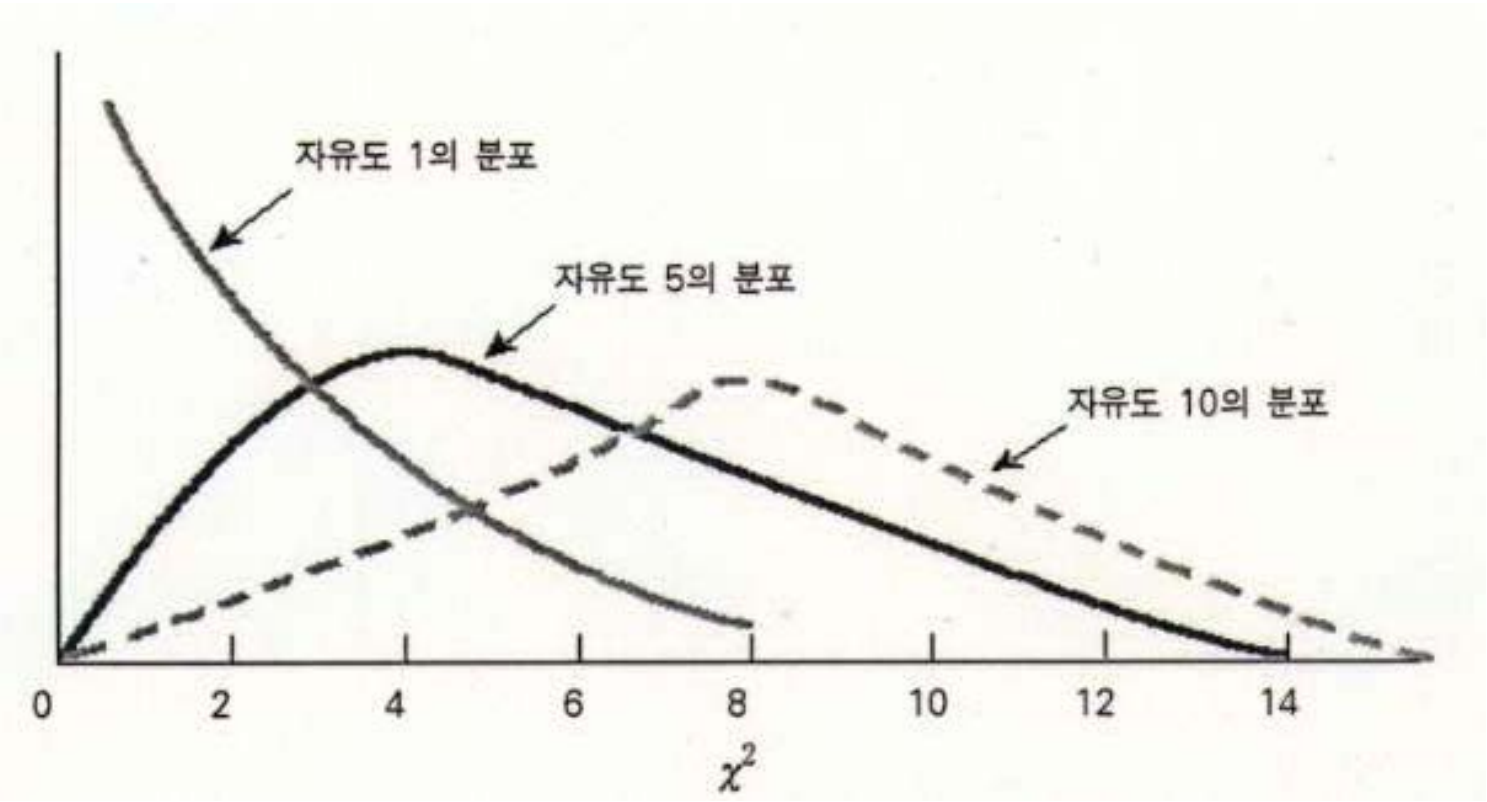
자유도 = (r-1)(c-1) 이므로, 이 경우

(2-1)×(4-1) = 3이다.

6) χ^2 분포

① 자유도의 수가 증가할 수록 곡선은 더 대칭적이 된다.

② χ^2 분포는 확률분포 이므로 각 χ^2 분포 내의 곡선 아래 부분의 전체 면적의 합은 1이다.



7) χ^2 검정

⇒ 자유도 3인 χ^2 분포에서 유의수준 10%일 때 기각역은 검정통계량이 6.251보다 클 때이다.

→ 이 예에서는 검정통계량의 값이 3.032이므로, 귀무가설은 기각되지 않는다.

→따라서, 유의수준 10% 하에서 '지역'과 '보상방식'은 서로 독립이어서 아무런 연관관계가 없다.