통계 기반 데이터 분석 용어 이해하기 (상관분석)

목 차

실습 1. 상관분석 리서치

- (1) 상관분석(Correlation Analysis)이란 무엇인지 찾아 적어보자.
- (2) 상관계수란 무엇인지 찾아서 적어보자.
- (3) 상관계수가 -1이면 어떤 의미인지,
  1이면 어떤 의미인지 0이면 어떤 의미인지 찾아 적어보자.
- (4) 상관 계수 가설 검정에 사용하는 R에서의 함수는 무엇인가?

1-1 상관분석(Correlation Analysis)

▶ 상관분석(Correlation Analysis)?

가. 확률론과 통계학에서 두 변수 간의 어떤 선형적 관계를 갖고 있는지 분석하는 방법

나. 두 변수 간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다.

1-2 상관계수(Correlation Analysis)

▶ 상관계수(correlation coefficient)

상관 계수는 몇 가지 유형의 상관 관계를 수치로 측정한 것으로 두 변수 간의 통계적 관계를 의미한다.

▶ 상관계수의 개념

r = X와 Y가 함께 변하는 정도/X와 Y가 각각 변하는 정도 (피어슨의 상관계수)

일반적으로
r이 -1.0과 -0.7 사이이면, 강한 음적 선형관계,
r이 -0.7과 -0.3 사이이면, 뚜렷한 음적 선형관계,
r이 -0.3과 -0.1 사이이면, 약한 음적 선형관계,
r이 -0.1과 +0.1 사이이면, 거의 무시될 수 있는 선형관계,
r이 +0.1과 +0.3 사이이면, 약한 양적 선형관계,
r이 +0.3과 +0.7 사이이면, 뚜렷한 양적 선형관계,
r이 +0.7과 +1.0 사이이면, 강한 양적 선형관계
로 해석한다.

어떤 분석 방법이 있을까?

1-2 상관계수(Correlation Analysis)

▶ 피어슨 상관 계수(Pearson correlation coefficient)

두 변수의 관련성을 위해 보편적으로 이용된다.

r = X와 Y가 함께 변하는 정도/X와 Y가 각각 변하는 정도

▶ 스피어만 상관 계수(Spearman correlation coefficient)

데이터가 서열척도인 경우 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수이다.

▶ 크론바흐 알파 계수 신뢰도(Cronbach's alpha)

계수 α 는 검사의 내적 일관성을 나타내는 값이다. 변수들 간의 평균상관관계에 근거해 검사문항들이 동질적인 요소로 구성되어 있는지를 분석하는 것이다.

1-2 상관계수(Correlation Analysis)

▶ 양의 상관성, 음의 상관성

값의 범위는 -1 ~ 1 사이이다.

X의 변수 값이 커짐에 따라 Y의 변수 값도 커지는 경우, 양의 관련성이 있다.

X의 변수 값이 커짐에 따라 Y의 변수 값도 작아지는 경우, 음의 관련성이 있다.

▶ 상관 계수

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

▶ 검정 통계량

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

1-2 상관계수(Correlation Analysis)

▶ 상관계수의 가설검정(예)

상관계수 r은 모수 ρ(rho)에 대한 추정치이므로 r을 계산한 후에는 가설검정과정을 거쳐야 한다.

귀무가설(H_0)

H_0 : **ρ=0 (X, Y 사이에 관계가 없다)**

H_0 : **$\rho \leq 0$ (X, Y 사이에 관계가 양의 상관관계가 아니다.)**

H_0 : **$\rho \geq 0$ (X, Y 사이에 관계가 음의 상관관계가 아니다.)**

대립가설 (H_1)

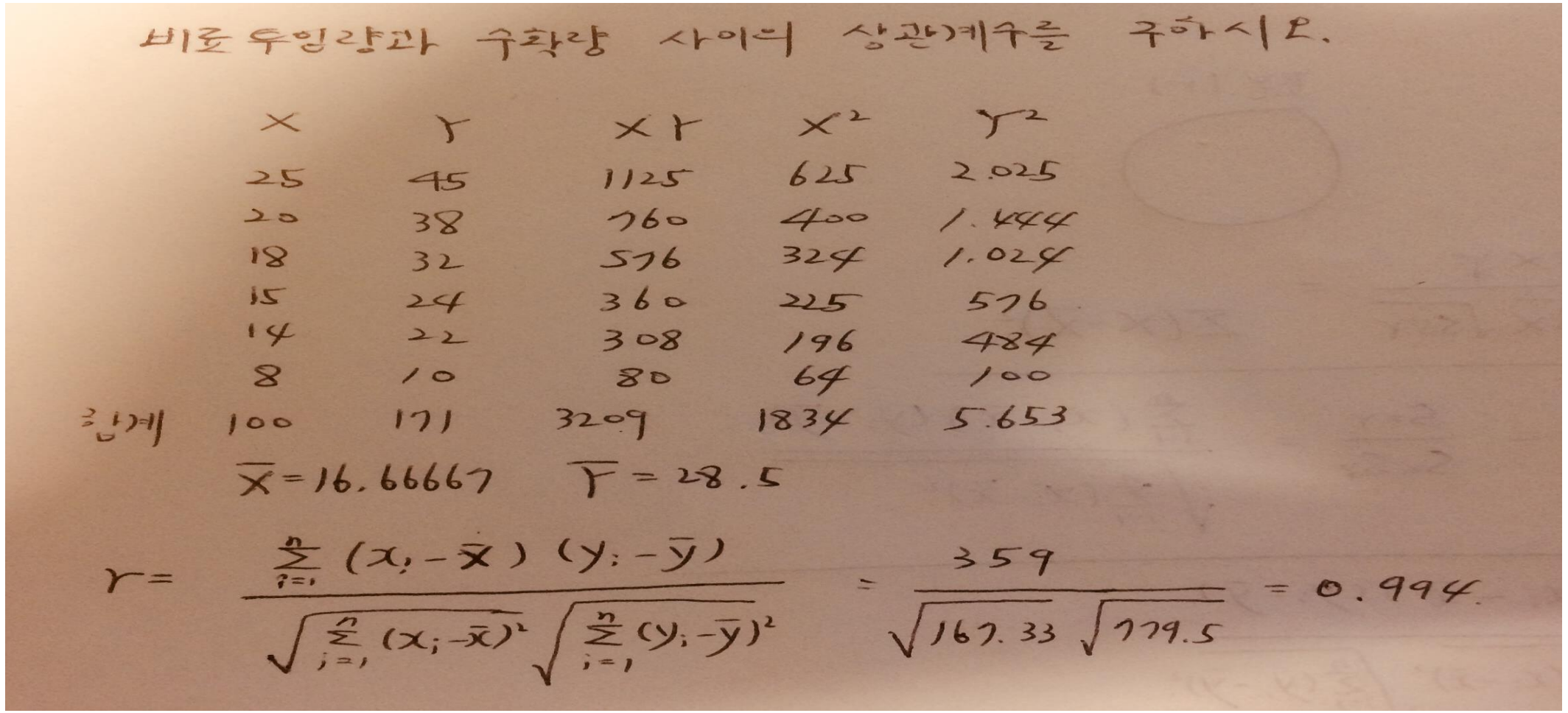
H_1 : **ρ=0 (X, Y 사이에 관계가 있다)**

H_1 : **$\rho > 0$ (X, Y 사이에 관계가 양의 상관관계이다.)**

H_1 : **$\rho < 0$ (X, Y 사이에 관계가 음의 상관관계이다.)**

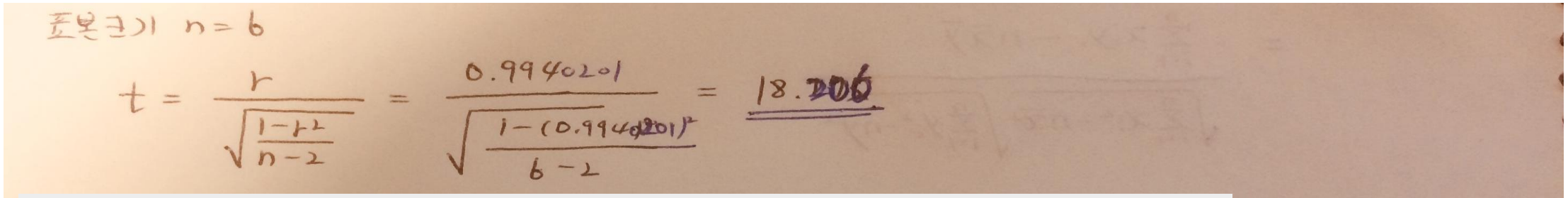
1-2 상관계수(Correlation Analysis)

▶ 상관계수의 가설검정(예)

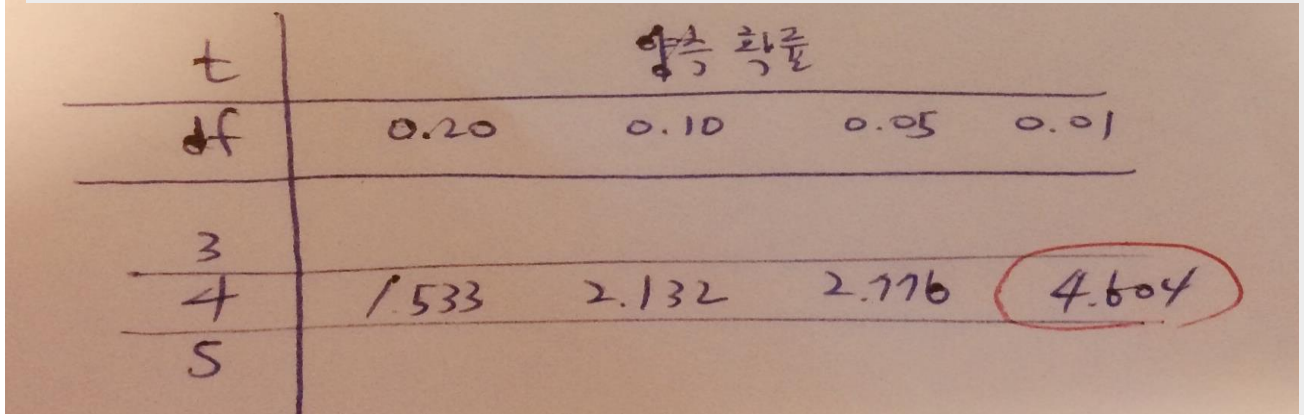


1-2 상관계수(Correlation Analysis)

▶ 상관계수의 가설검정(예)



양측 검정 하에서 유의수준이 α=0.01이고 자유도 df=n-2=4일 때의 t 임계값은 4.064이다. 귀무 가설의 채택역은 -4.604<= t <= 4.604가 된다.



T 통계량 값은 18.206으로 범위를 벗어나 있다.

ρ=0의 귀무가설을 기각하고 대립가설을 채택한다.

그렇다면 범주형 변수의 관계도
를 구하는 것은 없을까?

1-2 교차분석(카이제곱 검정)

▶ 교차분석(cross-tabulation analysis)?

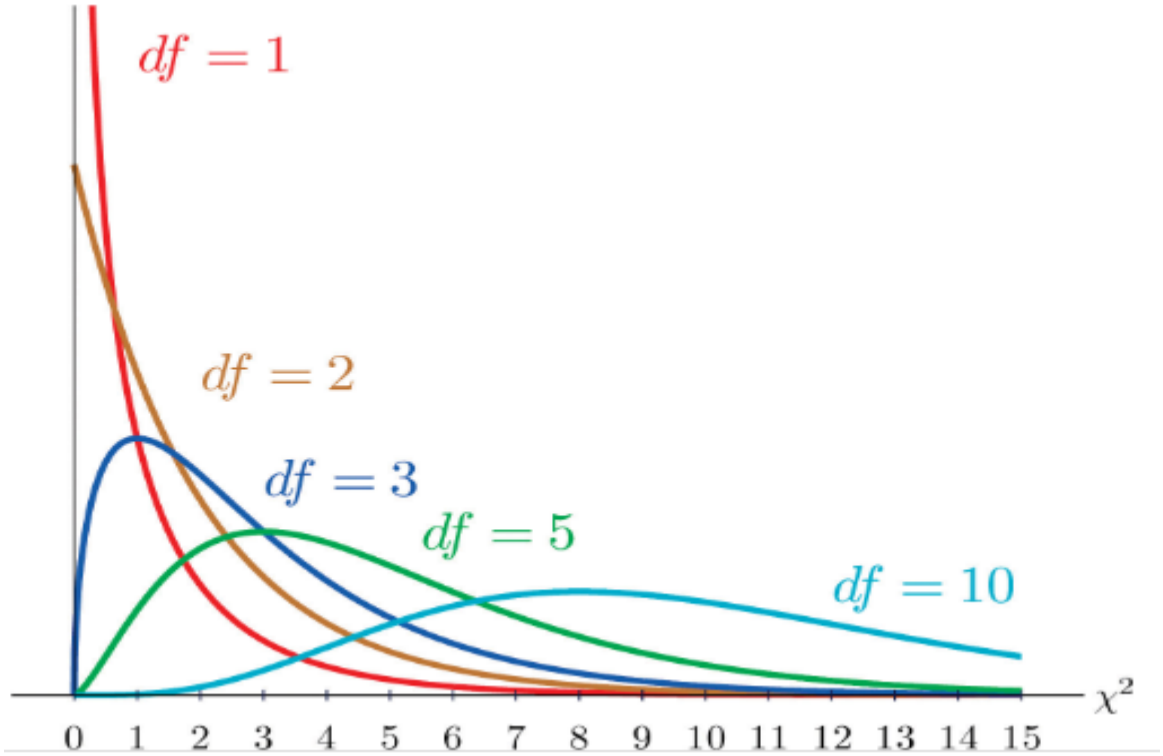
2개의 조사 요인에 대한 자료값을 각각 행과 열로 배열하여 교차되는 항목에 대한 빈도를 나타낸 표를 교차표(cross-tabulation)라 한다.

	열	구매의사		행의 합계
행		있음(1)	없음(2)	
지역	1	n_{11}	n_{12}	n_{1j}
	2	n_{21}	n_{22}	n_{21}
열의 합계		n_{i1}	n_{i2}	n

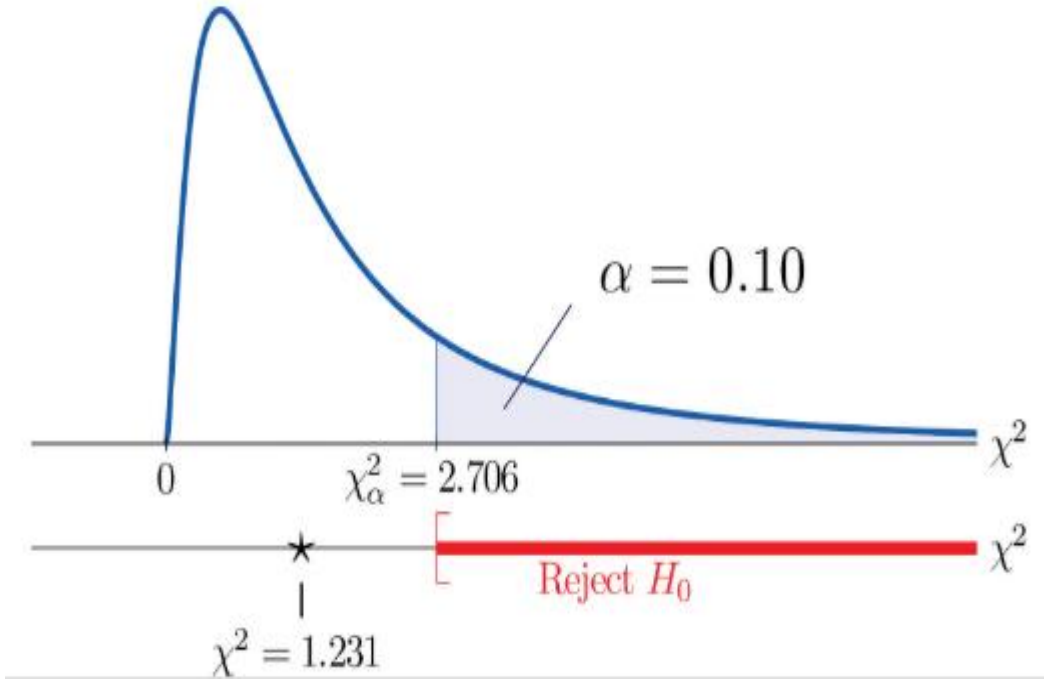
1-2 교차분석(카이 제곱 검정)

▶ 카이 제곱 분포

▶ 카이 제곱 분포와 자유도



▶ 유의수준과 귀무가설 기각



1-2 교차분석(카이 제곱 검정)

▶ 카이 제곱 검정의 귀무가설과 독립가설

귀무가설(H_0)

H_0 : 두 **요소가 독립적이다.**

대립가설 (H_1)

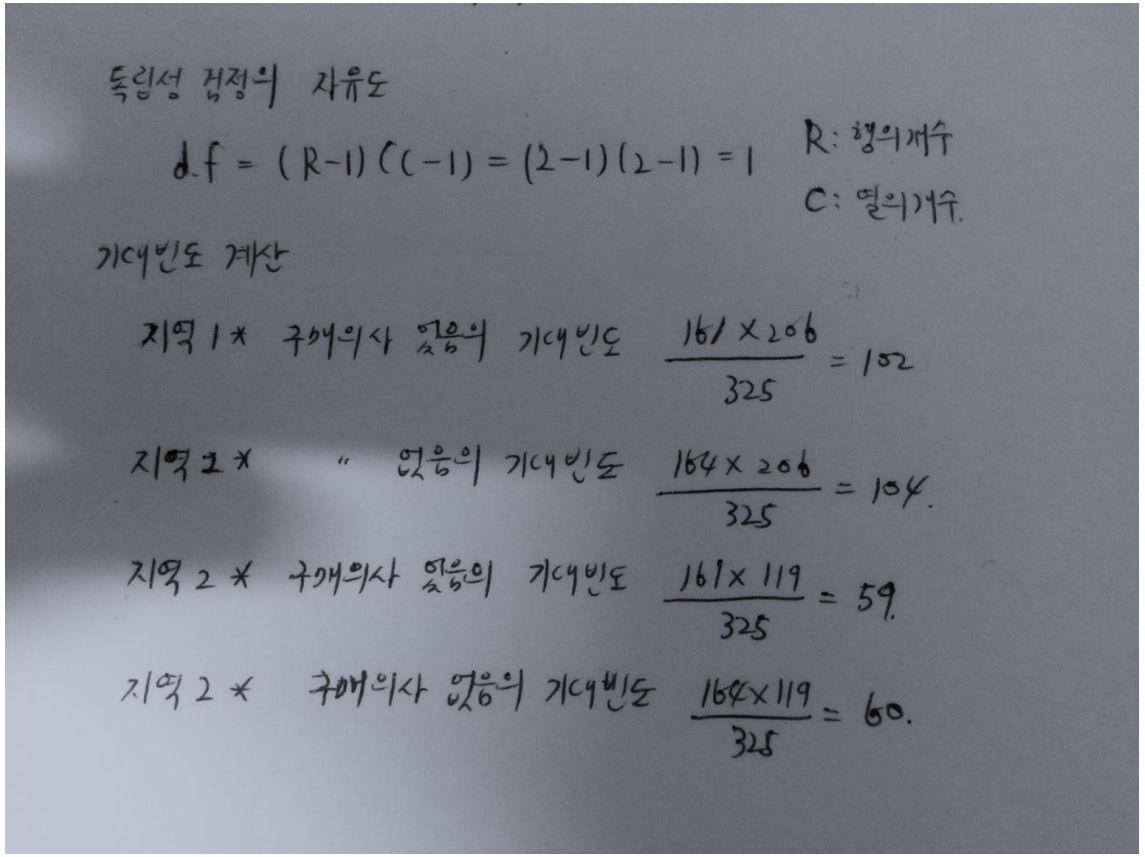
H_1 : 두 **요소가 독립적이지 않다.**

1-2 교차분석(카이제곱 검정)-독립성 검정

H_0 : 지역과 구매의사는 독립적이다.

H_1 : 지역과 구매의사는 독립적이지 않다.

$$E_{ij}(\text{기대빈도}) = \frac{n_{i.} * n_{.j}}{n}$$



	열	구매의사		행의 합계
행		있음(1)	없음(2)	
지역	1	154	52	206
	2	7	112	119
열의 합계		161	164	325

	열	구매의사		행의 합계
행		있음(1)	없음(2)	
지역	1	154	52	206
	기대빈도	102	104	
	2	7	112	119
	기대빈도	59	60	
열의 합계		161	164	325

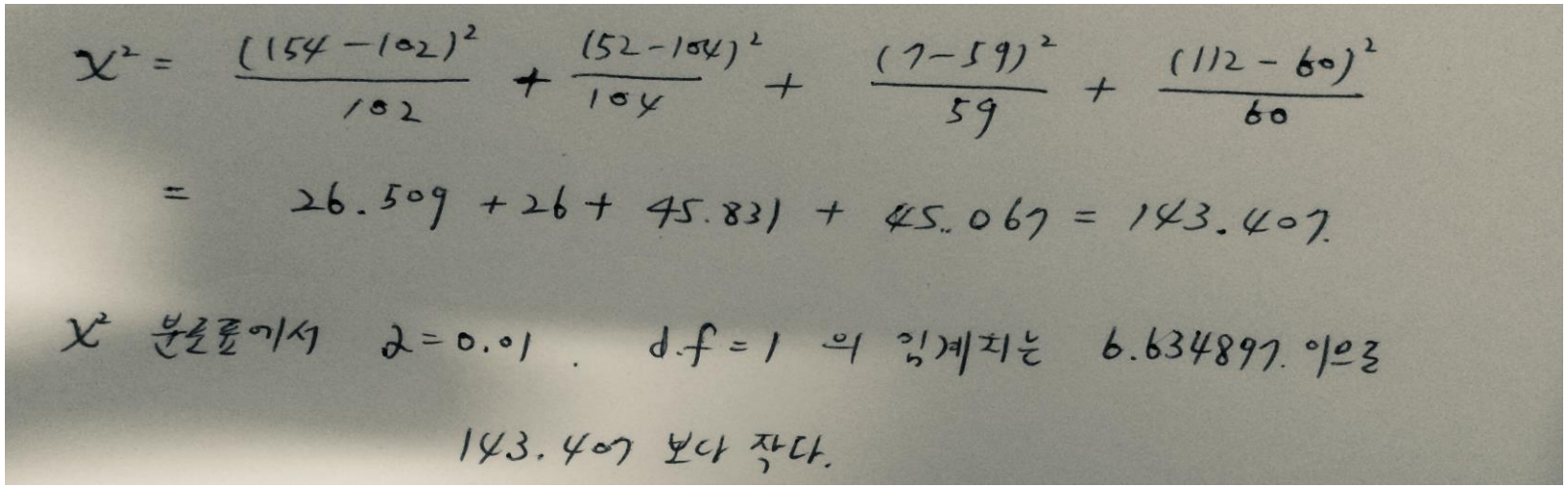
1-2 교차분석(카이제곱 검정)

▶ 독립성 검정을 위한 카이제곱 검정

	열	구매의사		행의 합계
행		있음(1)	없음(2)	
지역	1	154	52	206
	기대빈도	102	104	
	2	7	112	119
	기대빈도	59	60	
열의 합계		161	164	325

카이제곱 통계량

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



1-2 교차분석(카이제곱 검정)

▶ 독립성 검정을 위한 카이제곱 검정 정리

		Factor 2 Levels					
		1	· · ·	*j*	· · ·	*J*	Row Total
Factor 1 Levels	1	*O*	· · ·	*O*	· · ·	*O*	*R*
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	*O*	· · ·	*O*	· · ·	*O*	*R*
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	I	*O*	· · ·	*O*	· · ·	*O*	*R*
Column Total		*C*	· · ·	*C*	· · ·	*C*	*n*

기대 빈도

$$E = \frac{R \times C}{n}$$

$$E_{ij}(\text{기대빈도}) = \frac{n_{i.} * n_{.j}}{n}$$

1-2 교차분석(카이제곱 검정)

▶ 독립성 검정을 위한 카이제곱 검정 정리

		Factor 2 Levels					
		1	· · ·	*j*	· · ·	*J*	Row Total
Factor 1 Levels	1	O E	· · ·	O E	· · ·	O E	*R*
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	O E	· · ·	O E	· · ·	O E	*R*
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	I	O E	· · ·	O E	· · ·	O E	*R*
Column Total		*C*	· · ·	*C*	· · ·	*C*	*n*

카이제곱 통계량

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

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

1-2 교차분석(카이제곱 검정)

▶ 도전 문제

	열	심박수		행의 합계
행		Low	High	
성별	Girl	11	7	18
	Boy	17	5	22
열의 합계		28	12	40

가. 기대빈도를 구해보자.

나. 카이 제곱 통계량을 구해보자.

