

확률및통계

제4주 두 범주형 변수자료간의 관계표현

hylee@silla.ac.kr

두 범주형 변수의 요약

◎ 분할표 (Contingency Table)

② \ ①	찬성	미결정	반대	합계
남	112	36	28	176
여	84	68	72	224
합	196	104	100	400

칸 (cell)

: 2 * 3 분할표

◎ 상대돏수 분할표

② \ ①	찬성	미결정	반대	합계
남	0.28	0.09	0.07	0.44
여	0.21	0.17	0.18	0.56
합	0.49	0.26	0.25	1

분석하고자 하는
목적에 따라 분할표를
다양하게 정리할 수 있다.

- 남자이면서 찬성한 사람의 비율은 전체 중의 28%이다.
- 남, 여의 찬성 비율 비교 → 남과 여에서 따로 따로 찬성 비율을 계산.
- 남자 중 찬성 비율이 여자 중 찬성 비율보다 훨씬 높다.

두 연속형 변수의 요약 - 1

◎ 산점도 (Scatter plot)

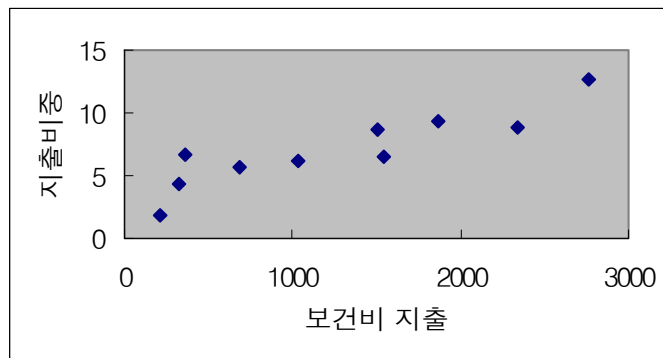


그림 : GDP대비 보건의료비 지출비중과
1인당 보건비 지출에 대한 산점도

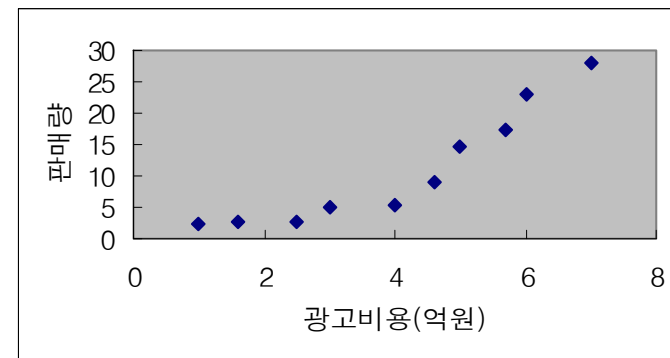


그림 : 광고비용과 판매량의 산점도

두 연속형 변수의 요약 - 2

◎ 상관계수 (Correlation coefficient) : (Karl Pearson)

: 두 변수간의 직선의 관계의 정도를 측정.

■ 자료 : $(x_1, y_1), \dots, (x_n, y_n)$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

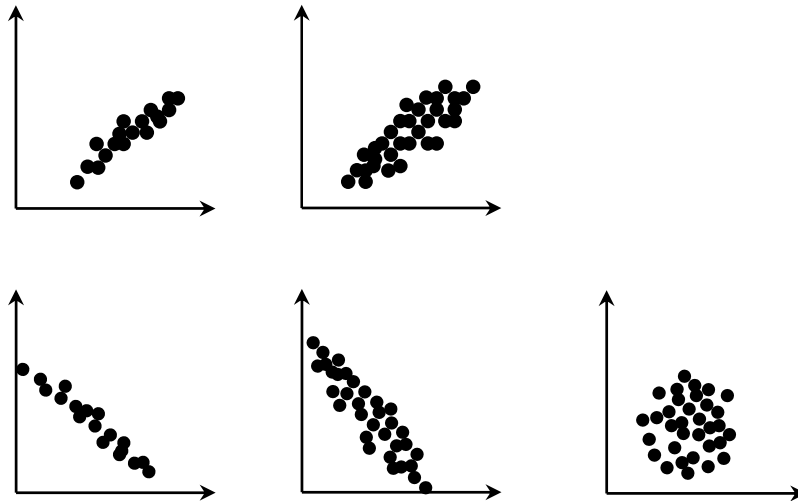
$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - n \bar{x}^2 = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - n \bar{y}^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

상관계수의 특징 - 1

- $-1 \leq r \leq 1$
1. r
 2. $r > 0$: 양의 상관관계 (기울기 : 양수)
 3. $r < 0$: 음의 상관관계 (기울기 : 음수)
 4. r 의 절대값이 클수록 직선의 관계가 강하다.



(그림 참고)

$$\tilde{a}, \tilde{b} \in \mathbb{R}^n$$

$$\tilde{a} \cdot \tilde{b} = |\tilde{a}| |\tilde{b}| \cos \theta \leq |\tilde{a}| |\tilde{b}|$$

$$\sum a_i b_i \leq \sqrt{\sum a_i^2} \sqrt{\sum b_i^2}$$

$$a_i = (x_i - \bar{x})$$

$$b_i = (y_i - \bar{y})$$

상관계수의 특징 - 2

◎ 상관관계와 인과관계

- 상관계수 값이 크다고 해서 항상 높은 인과관계를 갖는 것은 아니다.

(eg) 종교집회건수, 살인사건 발생건수

◎ 표본 상관계수의 성질

$$x \rightarrow ax + b = x'$$

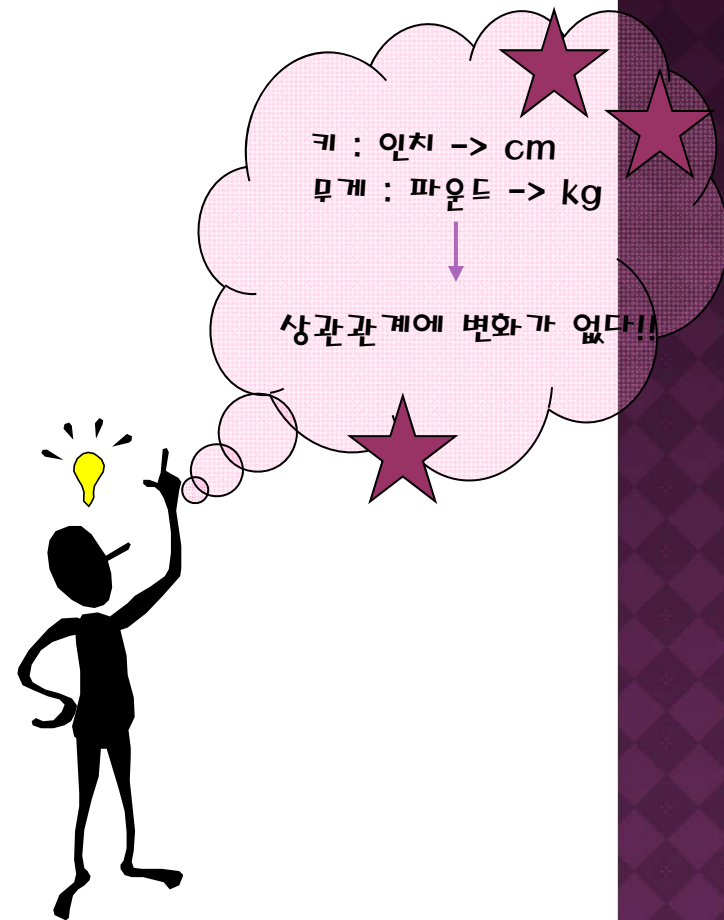
$$y \rightarrow cy + d = y'$$

$$r' = \frac{S_{xy}'}{\sqrt{S_{xx}'} \sqrt{S_{yy}'}} = \frac{a \cdot c}{|a||c|} r$$

$$S_{xy}' = \sum_{i=1}^n (x_i' - \bar{x}') \sum_{i=1}^n (y_i' - \bar{y}') = acS_{xy}$$

$$S_{xx}' = a^2 S_{xx}$$

$$S_{yy}' = c^2 S_{yy}$$



Thank You!