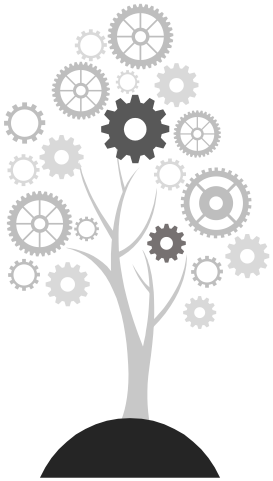
Contents

Chapter 4　두 변수 자료의 요약



Chapter 4

두 변수 자료의 요약

01 두 범주형 변수의 요약 : 분할표

두 변수가 모두 범주형이 측정될 경우

- 분할표 (contingency table)
  : 한 변수에 대한 범주는 왼쪽에, 또 다른 변수에 대한 범주는 위쪽에 표시하고, 두 변수의 범주들이 교차하는 칸마다 각 변수의 범주를 동시에 갖는 관측값들의 수를 그 칸의 도수(상대도수)로 기록

예제 1. 성별과 정책 지지여부 (표본의 개수 :n=400)

남와 여에서 찬성한 관측값들의 개수.

		열 속성			행 합계
		찬성	미결정	반대	
행 속성	남자	112	36	28	176
	여자	84	68	72	224
열 합계		196	104	100	400

두 변수 자료의 요약

- 분할표 (contingency table) 각 도수를 전체 표본의 개수로 나누어 상대도수 분할표를 만들 수 있다
- 어떤 특정 조합에서 상대적 비율이 높은지를 쉽게 확인할 수 있음

		열 속성			행 합계
		찬성	미결정	반대	
행 속성	남자	0.28	0.09	0.07	0.44
	여자	0.21	0.17	0.18	0.56
열 합계		0.49	0.26	0.25	1

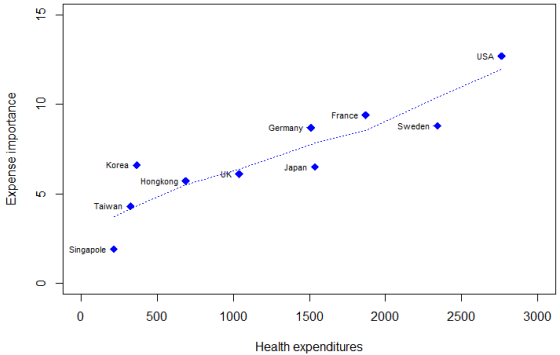
- 각 도수를 행 합계 혹은 열 합계로 나누어 상대도수를 구할 수도 있음
- (남자의 경우 찬성의 비율이 점점 높아지며 여자의 경우는 반대)

		열 속성			행 합계
		찬성	미결정	반대	
행 속성	남자	0.636	0.205	0.159	1
	여자	0.375	0.304	0.321	1

02 그림을 통한 두 연속형 변수의 요약 : 산점도

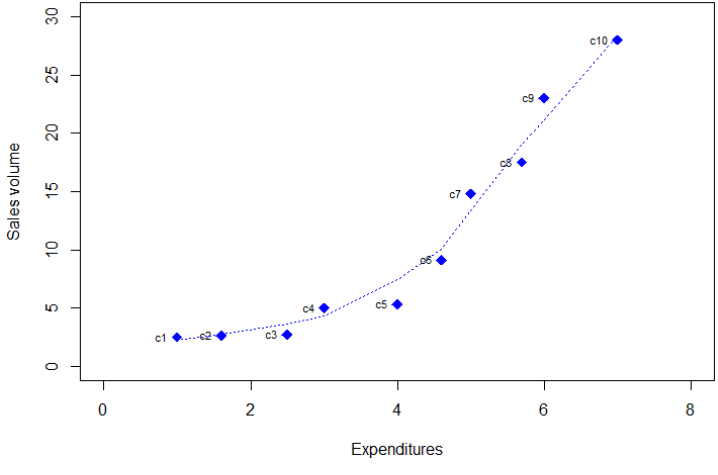
- 산점도(scatter diagram)
① 변수 x를 수평축에 놓고 변수 y를 수직축에 놓은 후에 각 관측값의 짝 (x, y)를 좌표 위에 표시
② 두 연속형 변수의 연관 관계를 알고자 할 때, 그림을 통하여 시각적으로 대략 파악
③ 직선만이 아니라 곡선 등 여러 가지 형태가 나타날 수 있음

예제 2. 보건의료비 지출비중 (x) 대 보건비 지출 (y) : 선형의 패턴이 관측됨



두 변수 자료의 요약

예제 3. 광고비용 (x) 대 판매량 (y) : 비선형의 패턴이 관측됨



산점도와 관련한 것임!!!

03 수치를 통한 두 연속형 변수의 요약 : 상관계수

- 표본상관계수(sample correlation coefficient)
  : 산점도에서 점들이 얼마나 직선에 가까운가의 정도를 나타내는 데 쓰이는 측도
  (두 변수 x 와 y 사이의 선형관계를 나타내는 측도)

For $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

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

$$S_{xx} = \sum (x_i - \bar{x})^2, \quad S_{yy} = \sum (y_i - \bar{y})^2, \quad S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$S_{xx} = \sum (x_i - \bar{x})^2$

$S_{yy} = \sum (y_i - \bar{y})^2$

$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

※ p.101의 간편식을 참고

Chapter 4 두 변수 자료의 요약

역슬래시형 직선은 두 변수가 반비례 적인 상관관계를 갖고 있다는 것을 뜻함.

- 특징

① $-1 \leq r \leq 1$

② 표본상관계수의 절대값의 크기는 직선관계에 가까운 정도를 나타내고 표본상관계수의 부호는 직선관계의 방향을 나타낸다.

③ 상관계수의 절대값이 1에 가까울 수록 두 변수는 강한 선형관계를 가진다고 할 수 있다.
($|r| = 1$이면 모든 점이 정확히 직선 위에 존재)

핵심!!!

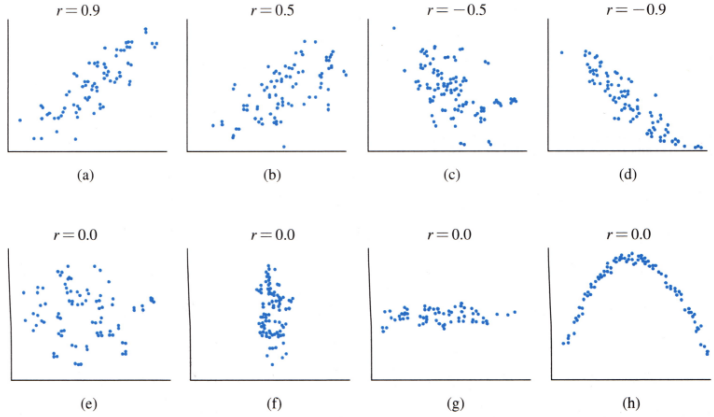


그림 4-3 표본상관계수와 산점도의 대응비교

두 변수 자료의 요약

④ 표본상관계수의 절대값이 0에 가깝다고 해서 x 와 y 가 서로 관계가 없다고 할 수 없다.
⑤ 표본상관계수의 절대값이 1에 가깝다고 해서 항상 직선의 관계가 강하다고 할 수 없다.
⑥ 단위가 없다

- 직선의 관계가 아니지만 표본상관계수는 매우 높은 경우

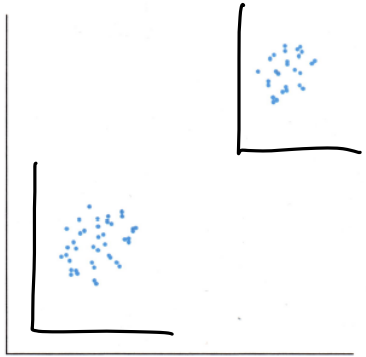


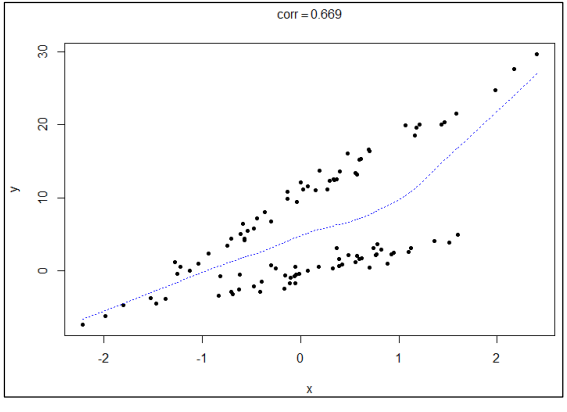
그림 4-4 표본상관계수가 무적절한 경우 – 두 모집단으로부터의 표본

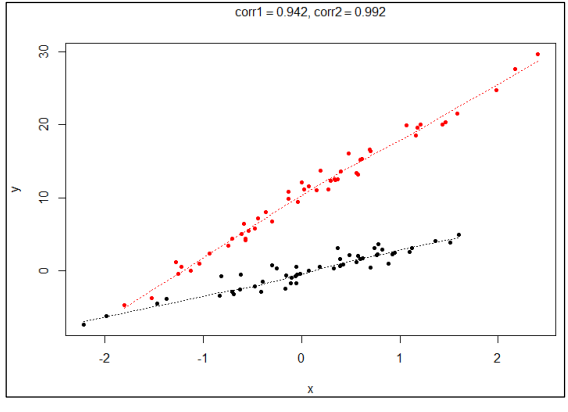
이렇게 각각 나누어보면, 상관관계가 낮다 보일 수 있다.

※ 변동계수와 표본상관계수는 단위가 없는 값이다! (<=>계산 과정중에 단위가 약분된다)

④ 집단 구분없이 그린 산점도에서는 직선의 패턴이 왜곡되어 나타낸다

- 직선의 관계가 아니지만 표본상관계수는 매우 높은 경우





두 변수 자료의 요약

예제 2. 보건의료비 지출비중 (x) 대 보건비 지출 (y)

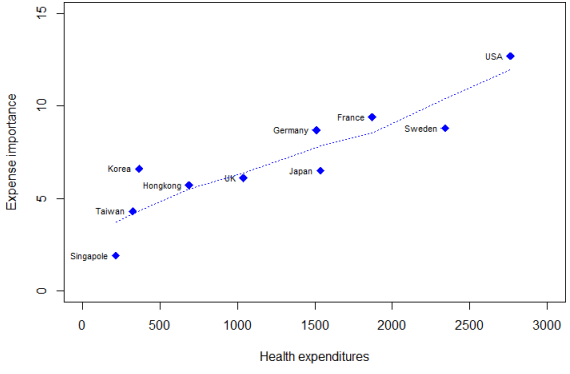
$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 7.07, \qquad \bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = 1265.5$$

$$S_{xx} = \sum(x_i - 7.07)^2 = 80.541$$

$$S_{yy} = \sum(y_i - 1265.5)^2 = 7096966.5$$

$$S_{xy} = \sum(x_i - 7.07)(y_i - 12.65.5) = 21434.55$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}} = \frac{21434.55}{\sqrt{80.541} \cdot \sqrt{7096966.5}} = 0.897$$

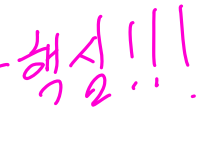


↑ 보건의료비 지출비중

Chapter 4 두 변수 자료의 요약

- 상관관계과 인과관계

상관관계 ≠ 인과관계



종교집회의 수와 살인사건의 수 사이의 표본상관계수가 매우 높다.
-> 종교집회의 수가 원인이고 살인사건의 수가 결과인 인과관계가 성립한다?

No!!!
종교집회가 많으면 인구수가 많을 것이고, 인구가 많으면 살인이 자주 일어 남
(여기서 인구수는 잠재변수의 역할)

표본상관계수는 단순히 두 (연속형) 변수 사이의 선형관계의 강도를 측정한 것

※ 인과관계에 대한 강도를 측정한 것이 아님