3. 이변량 자료의 설명

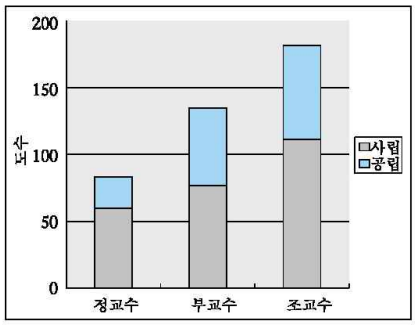
3-1. 이변량 자료

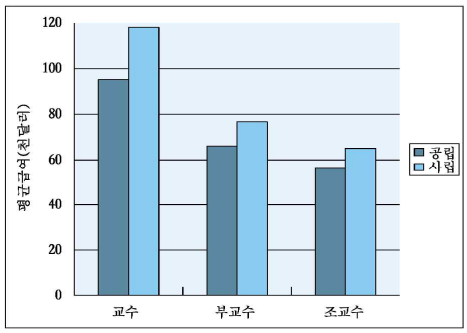
① 한 실험 또는 조사에서 두 변수가 동시에 측정

〖예〗

3-2. 도표/그래프에 의한 설명

3-2-1. 막대도표와 원도표





3-2-2. 산점도(scatter plot)

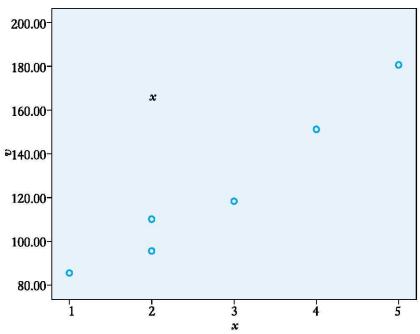
① 한 양적변수를 수평축(x축)으로 잡고 다른 양적변수를 수직축(y축)으로 잡고 각 관측값을 (x, y) 평면에 점으로 나타낸 그림

② 두 변수 x, y 사이의 관계를 설명

- 어떤 경향을 볼 수 있는가?
- 직선경향을 따라 일정하게 증가 또는 감소하는가?
- 곡선경향이 있는가?
- 어떤 경향도 볼 수 없는가?
- 경향은 어느 정도로 강한가? 모든 점들이 정확하게 경향을 따르는가? 혹은 관계가 약하게 보이는가?
- 이상점이 있는가? 점들이 그룹을 형성하는가?

〖예〗 A 지역에 거주하는 6가구, 가구원수 x와 주당 식료품비 지출액 y(단위: 천원) 측정

x	y
2	95.75
2	110.19
3	118.33
4	150.92
1	85.86
5	180.62



가구원수가 2이고 식료품비 지출액이 165인 가구가 있다고 가정한다면 그림 x가 다른 6점의 직선적인 경향과 적합하지 않은 이상점으로 간주된다.

3-3. 수치적 방법에 의한 설명

3-3-1. 공분산(covariance)

① 두 변수 X와 Y가 동시에 변하는 정도

② 두 변수들의 평균점인 $\overline{X}$, $\overline{Y}$를 중심으로 변하는 정도를 계산

$$\sigma_{xy} = Cov(X,\ Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}X_iY_i - \frac{\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n}\right]$$

3-3-2. 상관계수(correlation coefficient)

① 상관(correlation) : 두 변수가 변하는 정도

② 상관계수 : 두 변수가 관계되어 있는 정도를 나타내는 지수, 하나의 변수가 변해감에 따라 다른 변수가 변하는 정도를 나타내는 지수, 즉, 두 변수가 동시에 함께 변하는 정도를 나타내는 지수

③ 모수치에 의한 상관계수는 ρ(로우)라 표기하고, 통계치에 의한 표현은 r로 표기

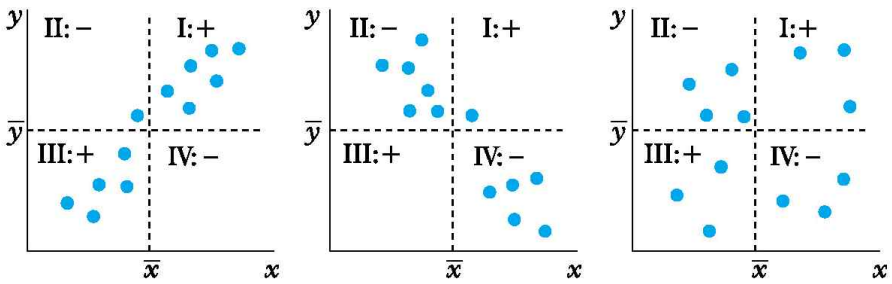
④ $\rho_{XY} = \dfrac{Cov(X,\ Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$

㉠ σ_x는 X의 표준편차

㉡ σ_y는 Y의 표준편차

㉢ σ_{xy}는 X와 Y의 공분산(covariance)

⑤ <양의 경향> <음의 경향> <경향이 없다>



㉠ Ⅰ과 Ⅲ에 있을 때 σ_{xy}와 r의 부호는 양수

㉡ Ⅱ과 Ⅳ에 있을 때 σ_{xy}와 r의 부호는 음수

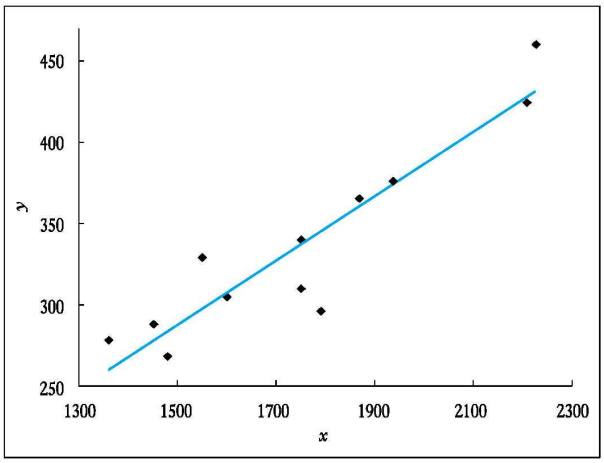
㉢ 점들이 네 영역에 흩어져있을 때는 s_{xy}와 r은 0 가까움.

⑥ X와 Y에 대하여 n쌍의 표본자료가 주어졌다면, 표본 공분산(sample covariance) S_{xy}와 표본 상관계수(sample correlation coefficient) r_{xy}를 계산할 수 있다.

$$r_{xy} = \frac{S_{xy}}{S_xS_y} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

〖예〗 12개 주거용 건물의 주거면적 x와 판매가격 y의 상관계수

건물	x (제곱피트)	y (천만원)
1	1360	278.5
2	1940	375.7
3	1750	339.5
4	1550	329.8
5	1790	295.6
6	1750	310.3
7	2230	460.5
8	1600	305.2
9	1450	288.6
10	1870	365.7
11	2210	425.3
12	1480	268.8



[풀이]

㉠ x와 y의 합과 표준편차를 구하면 각각

$$\sum_{i=1}^{n} x = 20{,}980,\ S_x = 281.4842,\ \sum_{i=1}^{n} y = 4043.5,\ S_y = 59.7592$$

㉡ 공분산은

$$S_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n\right]$$

$$= \frac{1}{11}[7{,}240{,}383 - (20{,}980)(4043.5)/12] = 15{,}545.19697$$

㉢ 상관계수 $r = \dfrac{S_{xy}}{S_x S_y} = \dfrac{15{,}545.19697}{(281.4842)(59.7592)} = 0.9241$

3-3-3. 회귀직선

① 모든 점들이 직선 위에 존재하지는 않지만 직선적인 경향을 볼 수 있다면 이 점들을 관통하는 최선의 직선을 적합하여 설명 가능

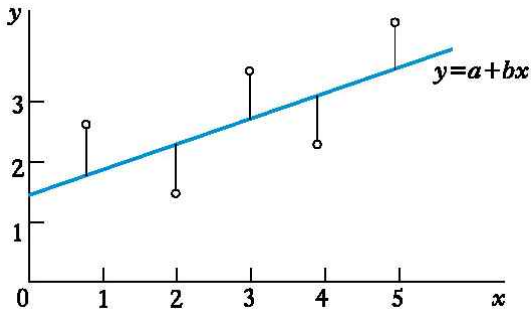
② x에 대한 y의 최적적합직선을 회귀선(regression line) 또는 최소제곱직선이라 함

③ 산점도에 있는 모든 점들이 자료를 대표하는 직선 쪽으로 회귀한다는 이유

④ 회귀선은 $\overline{X}$와 $\overline{Y}$인 점을 반드시 지난다.

⑤ 자료와 직선과의 차의 제곱의 합이 최소가 되도록 하여 a와 b를 구함.

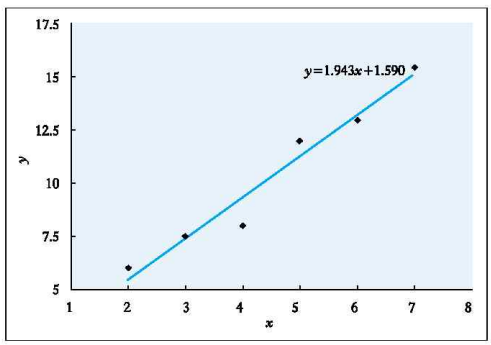
$$\text{회귀등식 } y = a + bx,\ b = r\left(\frac{s_y}{s_x}\right) = \frac{s_{xy}}{{s_x}^2},\ a = \bar{y} - b\bar{x}$$



⑥ X값을 회귀등식 $y=a+bx$에 대입하여 나타난 기대되는 값과 Y값과의 차이를 가장 작게 해야 함
⑦ s_x>0, s_y>0이므로 b와 r은 같은 부호를 갖음
⑧ r이 양수라면 b도 양수 ⇒ 직선은 x에 따라 증가
⑨ r이 음수라면 b도 음수 ⇒ 직선은 x에 따라 감소
⑩ r이 0에 가깝다면 b도 0에 가까움

〖예〗 근무경력 년 수 x와 시간당 초기임금 y(단위: 천 원)에 관한 자료

x	2	3	4	5	6	7
y	6.00	7.50	8.00	12.00	13.00	15.50



[풀이]

㉠ x와 y에 대하여 필요한 계산을 하면

$\bar{x}=4.5$, $\bar{y}=10.333$, $s_x=1.871$,

$s_y=3.710$, $r=0.980$

㉡ 기울기 $b=r\left(\frac{s_y}{s_x}\right)=0.980\left(\frac{3.710}{1.871}\right)=1.9432389\approx 1.943$

㉢ y절편 $a=\bar{y}-b\bar{x}=10.333-1.943(4.5)=1.590$

㉣ 최적회귀직선 $y=1.590+1.943x$

㉤ 회귀직선은 x값을 알고 있는 경우 y값을 예측하는데 사용, 예를 들어 근무 경력이 3년인 사람의 초기임금은 $y=1.590+1.943\times 3=7.419$라 예측

【참고】 상관계수와 회귀직선 사용의 차이점

상관계수	회귀직선
실험단위가 임의로 추출되고 동시에 x와 y에 대하여 측정이 이루어질 때	x의 값을 미리 정하고 난 다음에 y를 측정하는 경우