경영데이터분석기초

- SPSS, Excel을 활용한 통계분석 -

유 진 호

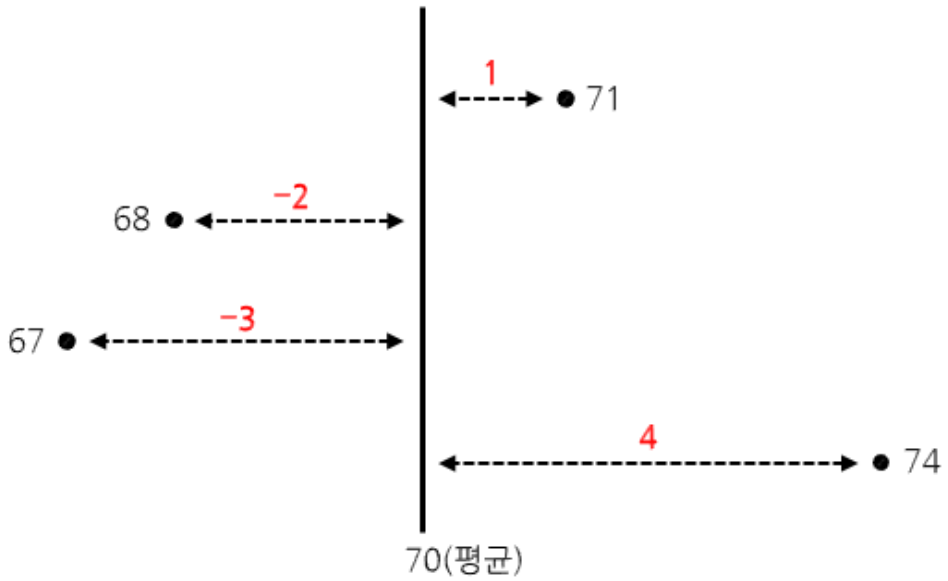
jhyoo@smu.ac.kr

평균, 분산, 표준편차

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

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

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



$(x_i - \bar{x})$ ← 편차

$(x_i - \bar{x})^2$ ← 편차 제곱

제곱합

$\sum(x_i - \bar{x})^2$ ← 편차 제곱합

$$\frac{\sum(x_i - \bar{x})^2}{(n-1)}$$ ← 편차 제곱의 평균(분산)

Covariance, Correlation

공분산, 상관계수

(키, 몸무게)

(age, balance)

평균 $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$

분산 $S^2 = \frac{1}{n-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$

표준편차 $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

COV(X,Y) = $\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$ **교차곱**

Corr(X,Y) = $\frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$ $\frac{\text{교차곱}}{\text{제곱합}}$

상관계수 r는 항상 부등식 $-1 \le r \le 1$을 만족시키며, 양의 상관관계가 있을 때는 $r>0$, 음의 상관관계가 있을 때는 $r<0$이다. 또 무상관일 때는 $r=0$이 된다.

Karl Pearson

From Wikipedia, the free encyclopedia

For the English cricketer, see Karl Pearson (cricketer).

Karl Pearson FRS FRSE[1] (/ˈpɪərsən/; born **Carl Pearson**; 27 March 1857 – 27 April 1936[2]) was an English mathematician and biostatistician. He has been credited with establishing the discipline of mathematical statistics.[3][4] He founded the world's first university statistics department at University College London in 1911, and contributed significantly to the field of biometrics and meteorology. Pearson was also a proponent of social Darwinism and eugenics, and his thought is an example of what is today described as scientific racism. Pearson was a protégé and biographer of Sir Francis Galton. He edited and completed both William Kingdon Clifford's *Common Sense of the Exact Sciences* (1885) and Isaac Todhunter's *History of the Theory of Elasticity*, Vol. 1 (1886–1893) and Vol. 2 (1893), following their deaths.

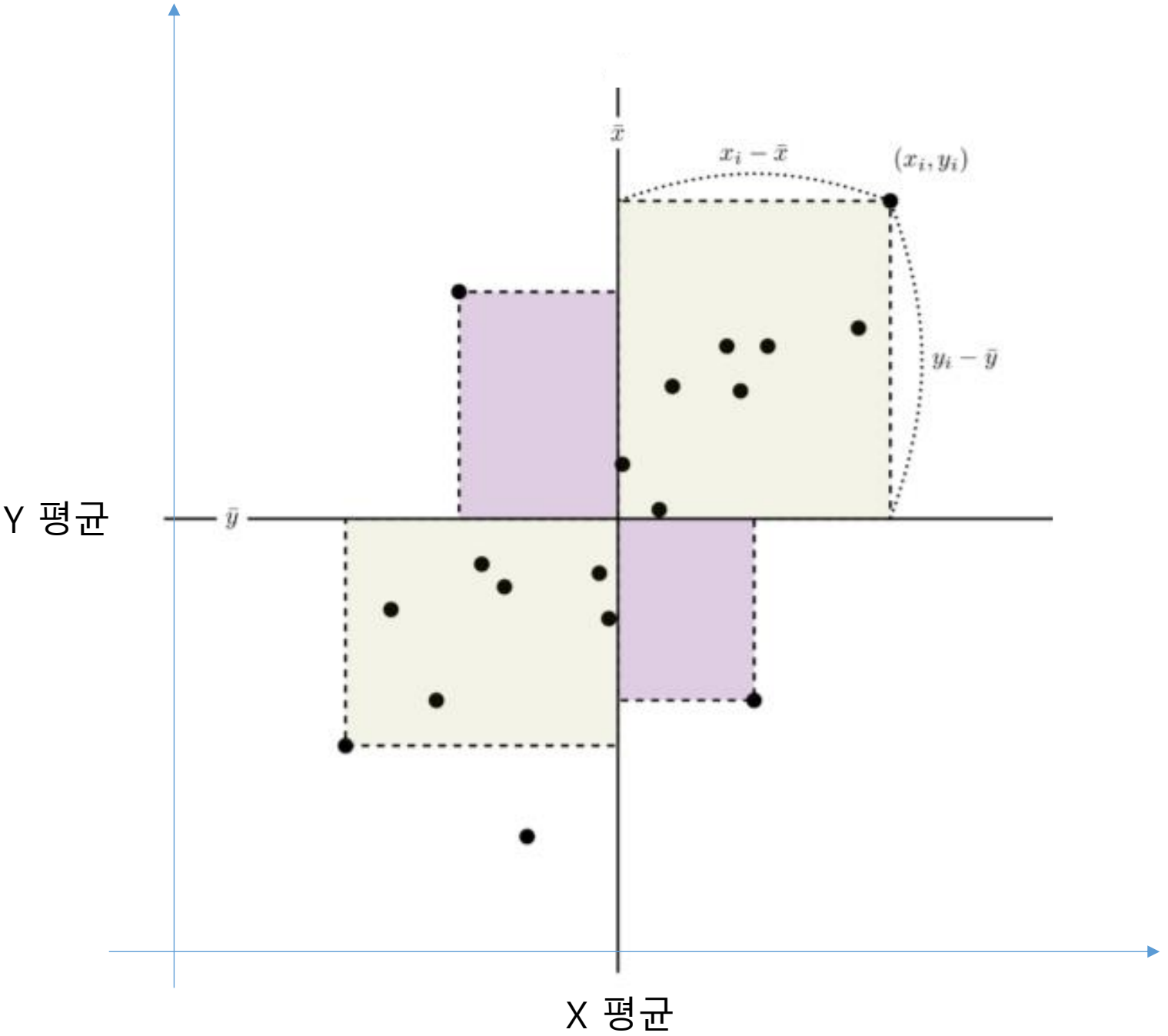
Karl Pearson
FRS

Pearson in 1912

Born	Carl Pearson 27 March 1857 Islington, London, England

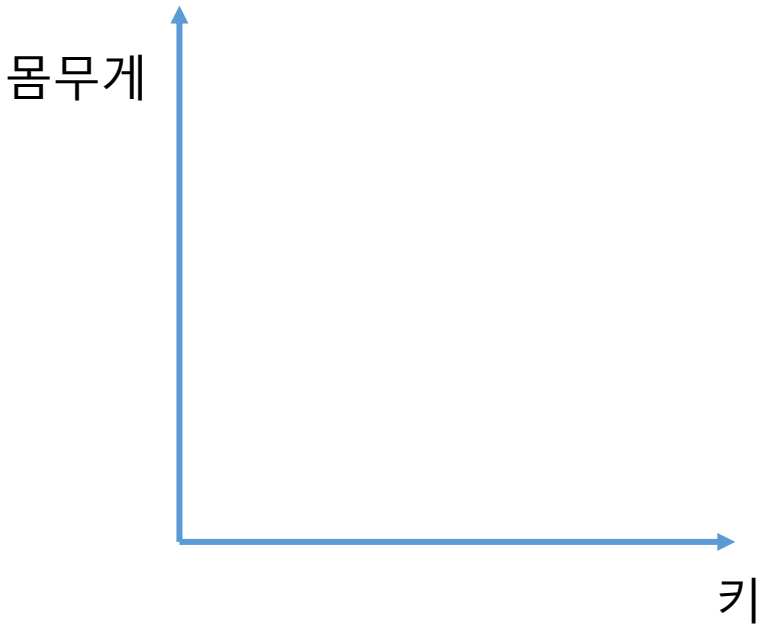
Early life and education [edit]

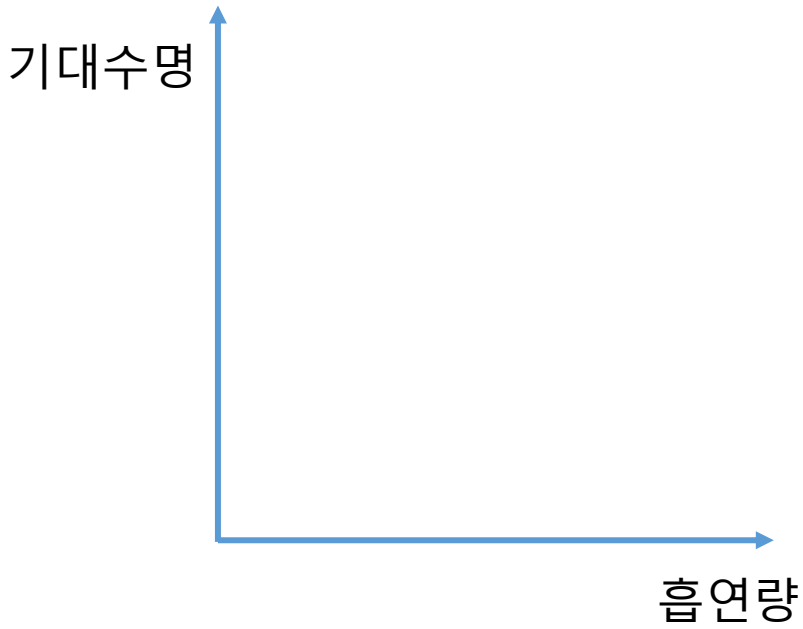
Pearson was born in Islington, London, into a Quaker family. His father was William Pearson QC of the Inner Temple, and his mother Fanny (née Smith), and he had two siblings, Arthur and Amy. Pearson attended University College School, followed by King's College, Cambridge, in 1876 to study mathematics,[5] graduating in 1879 as Third Wrangler in the Mathematical Tripos. He then travelled to Germany to study physics at the University of Heidelberg under G. H. Quincke and metaphysics under Kuno Fischer. He next visited the University of Berlin, where he attended the lectures of the physiologist Emil du Bois-Reymond on Darwinism (Emil was a brother of Paul du Bois-Reymond, the mathematician).



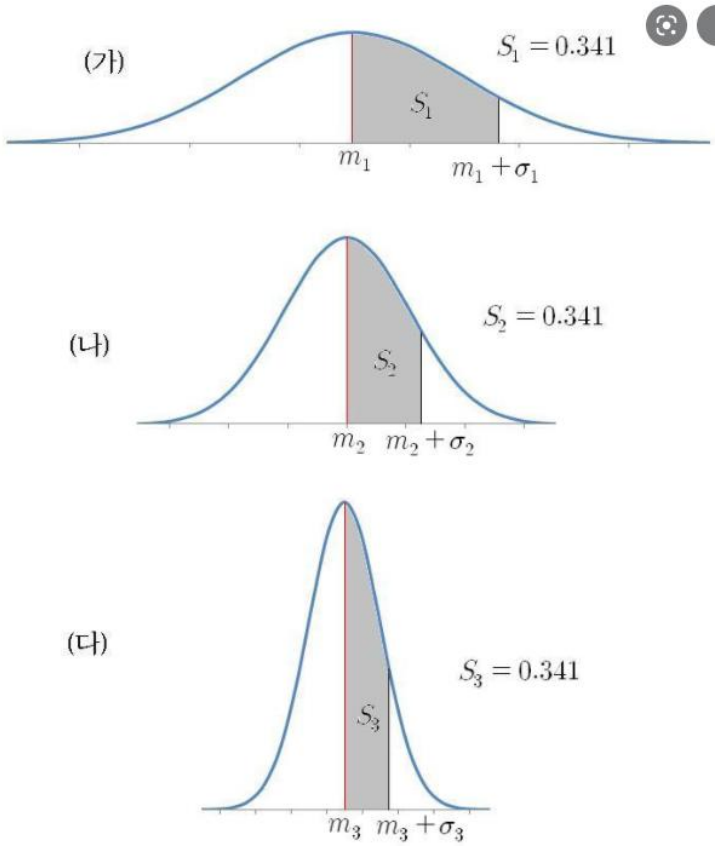
$$\text{COV}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)}$$

$$\textbf{Corr}(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

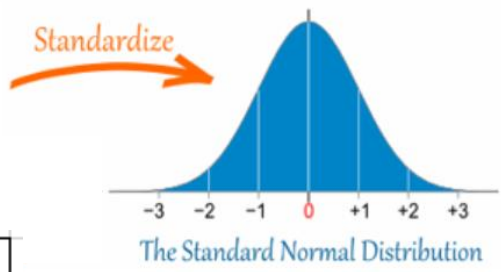

몸무게
키


기대수명
흡연량

표준화



$$Z = \frac{X - \mu}{\sigma}$$



	국어	수학
점수	90	70
평균	70	50
표준편차	10	5
표준점수	2.0	4.0

◆ 피어슨 상관계수에 대한 이해[coefficient of correlation, 相關係數]

상관계수는 다음과 같이 계산된다.

$$r_{xy} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 (y-\bar{y})^2}}$$

두 변량 X, Y 사이의 상관관계의 정도를 나타내는 수치(계수)이다.

즉, x, y 두 변량의 교차곱을 각각의 표준편차로 나눈 값이다.

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

공분산을 표준화한 값

상관계수 r는 항상 부등식 $-1 \le r \le 1$을 만족시키며, 양의 상관관계가 있을 때는 $r>0$, 음의 상관관계가 있을 때는 $r<0$이다. 또 무상관일 때는 $r=0$이 된다.

공분산[covariance, 共分散] : 두 변수(變數)의 관계를 나타내는 양(量)을 말한다.

x, y의 공분산(共分散)은 다음과 같이 계산된다.

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

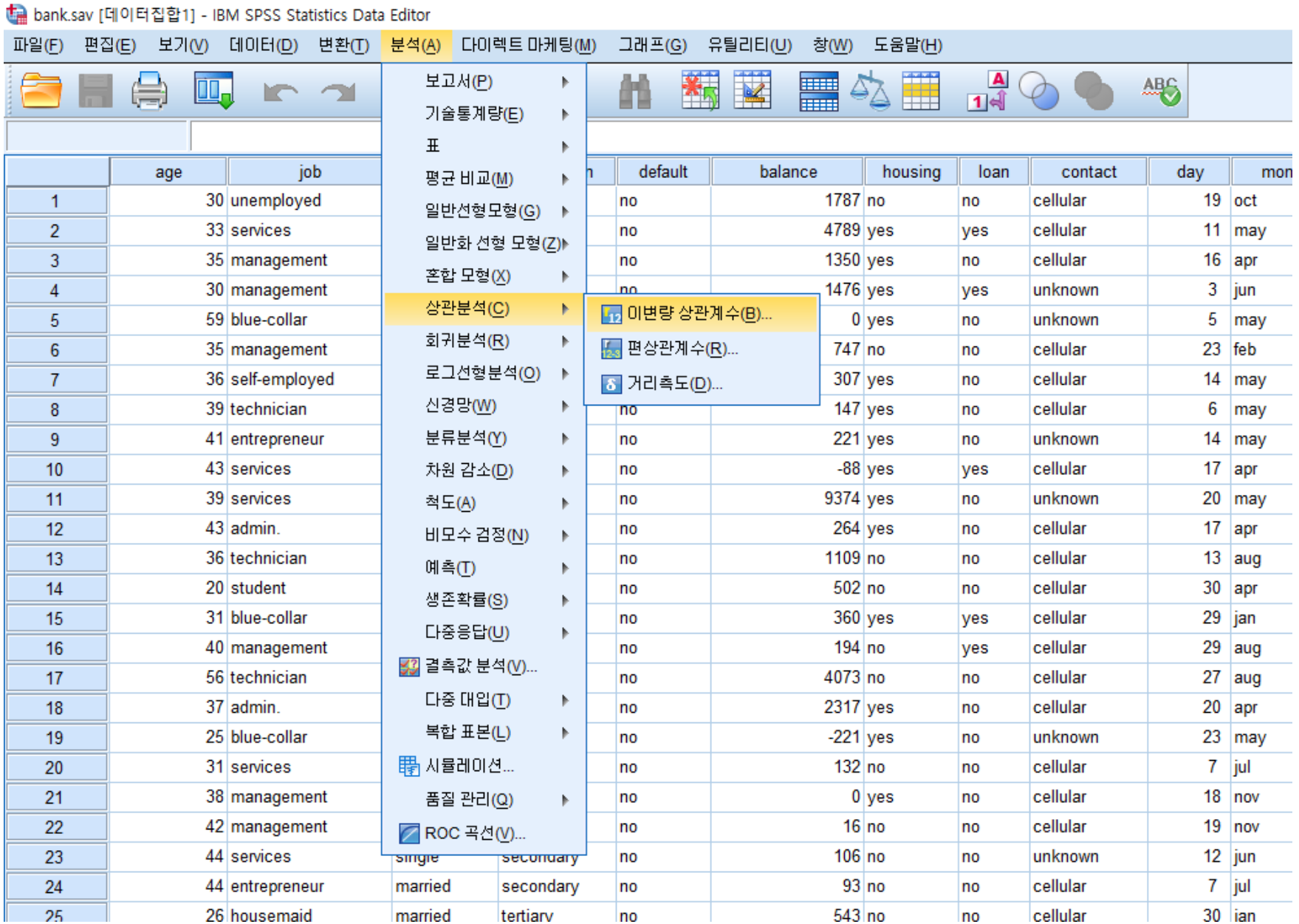
단위에 따라 변함

즉, x, y 두 변량의 교차곱을 (n−1)로 나눈 값이다.

◆ 공분산과 상관계수

과제1 : bank.csv를 엑셀로 읽어 들여 피어슨 상관계수 산출하기

과제2 : bank.csv를 SPSS로 읽어 들여 피어슨 상관분석 실시하기



연속형 변수들간의 관계 파악하기(상관관계분석)

.Correlate > Bivariate

(상관분석) > (이변량상관계수)

결과해석하기

- 상관분석
  - Age와 balance의 상관계수는 ###, 유의확률은 ###이므로 유의수준 0.05에서 상관성이 존재하다. 다만, 상관계수값이 작으므로 아주 작은 상관관계가 존재한다고 할 수 있다.

- 회귀분석
  - 분산분석표에 따라 F 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 회귀모형이 통계적으로 유의하다고 할 수 있다.
  - t 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 회귀계수는 0이 아니다. 즉, [나이]는 [balance]에 영향을 준다(나이에 따라 balace는 달라진다)고 할 수 있다.

- 교차분석(CrossTab)
  - 카이제곱(X^2) 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 [Y 집단]별로 [결혼상태]는 서로 다르다(서로 연관성이 있다)고 할 수 있다.

- 평균차이분석(Means)
  - F통계량 값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 [Y 집단]별로 [duration]은 통계적으로 유의한 차이가 있다.

과제

- 엑셀로 (age, balance) 공분산, 상관계수 구하기
- SPSS.빈도분석/기술통계/데이터탐색/상관계수 실행하기
- 엑셀에 붙여서 + 상관계수 부분 해석 설명달기