

경영데이터분석기초

- SPSS, Excel을 활용한 통계분석 -

유 진 호

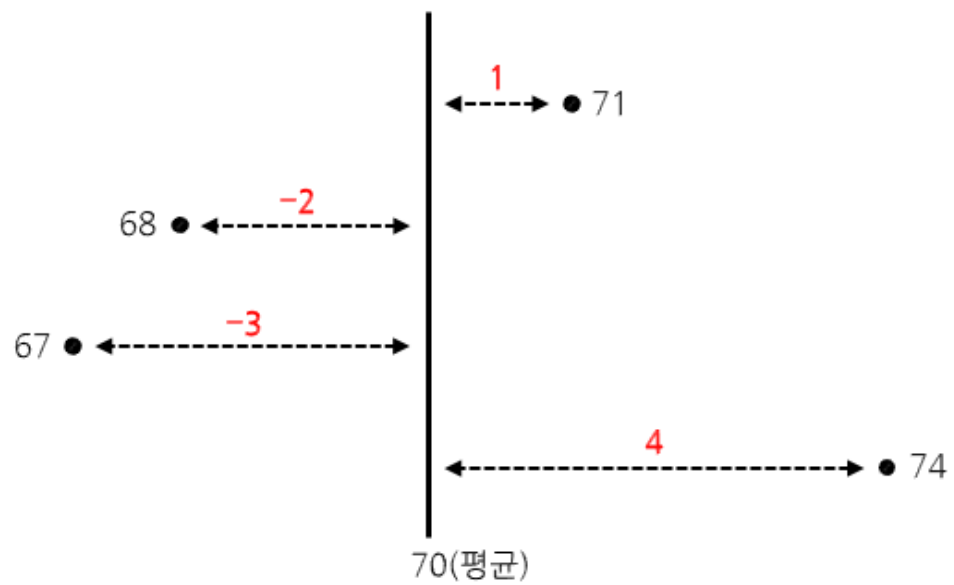
jhyoo@smu.ac.kr

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
$\sum(x_i - \bar{x})^2$	←	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

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

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

교차곱

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

교차곱

제곱합

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

Karl Pearson

🌐 48 languages ▾

Article Talk

Read Edit View history Tools ▾

From Wikipedia, the free encyclopedia

For the English cricketer, see [Karl Pearson \(cricketer\)](#).

Karl Pearson FRS FRSE^[1] (/ˈpiərsən/; 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.

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). Pearson also studied German Law taught by [Bruno](#) and [Meynert](#), medieval and 16th century German

Karl Pearson

FRS



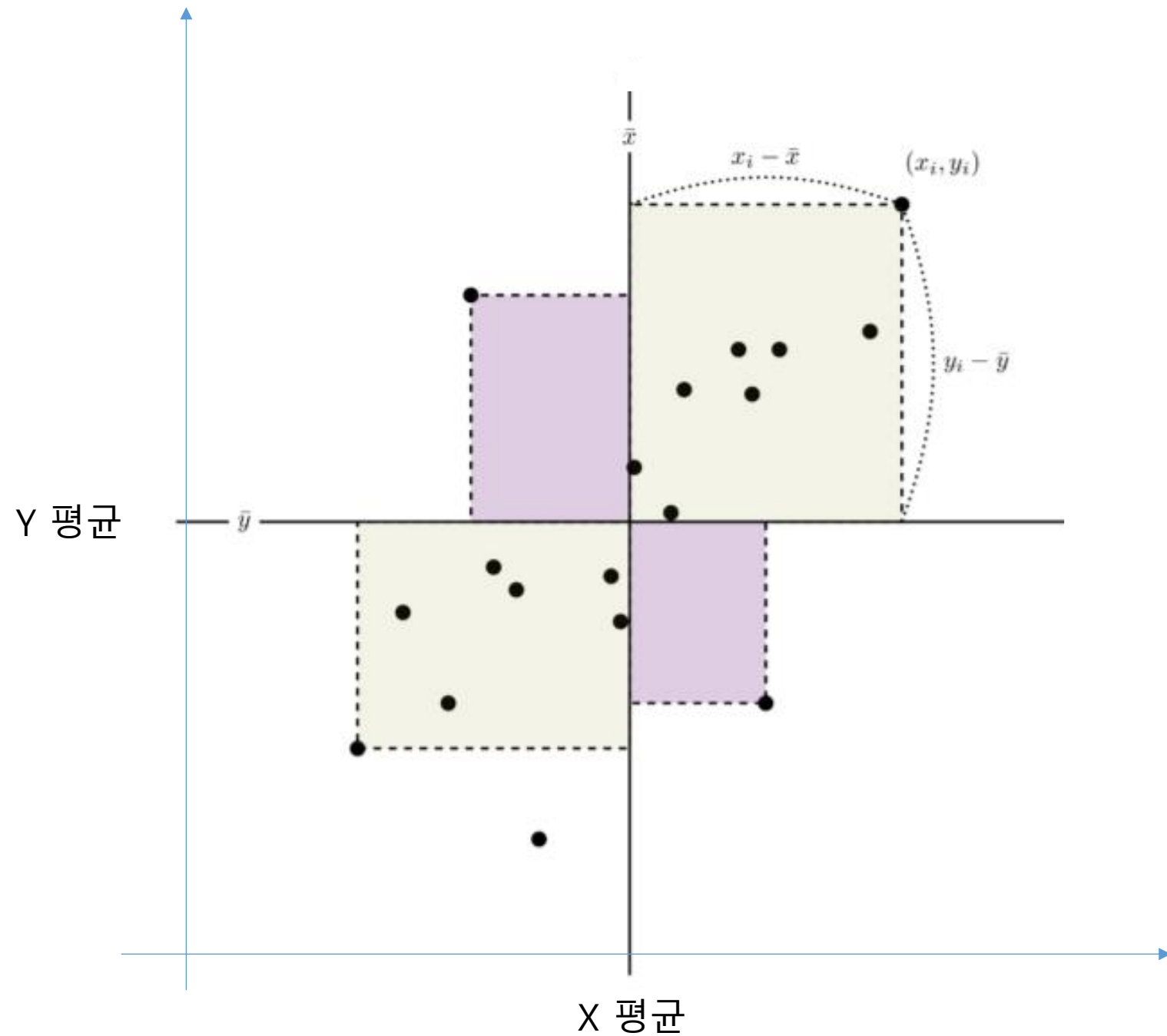
Pearson in 1912

Born

Carl Pearson

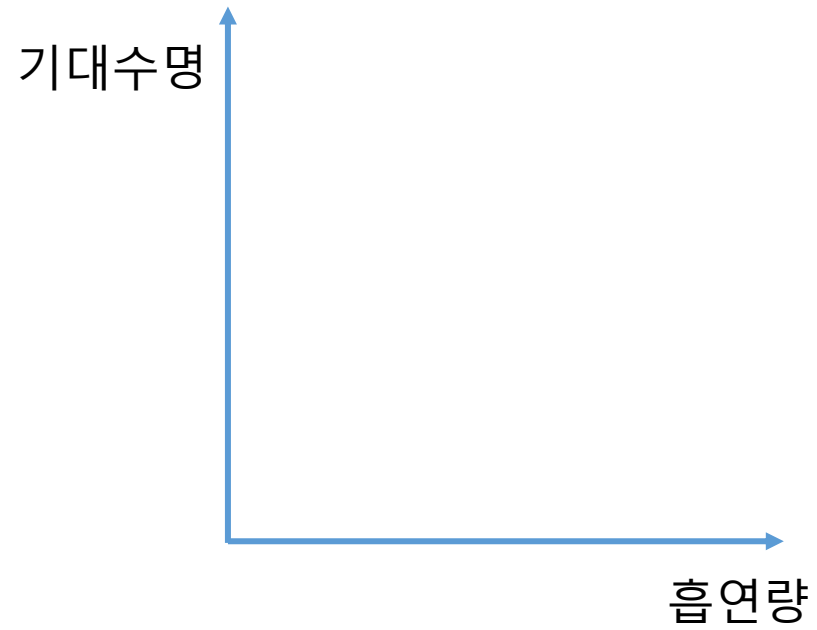
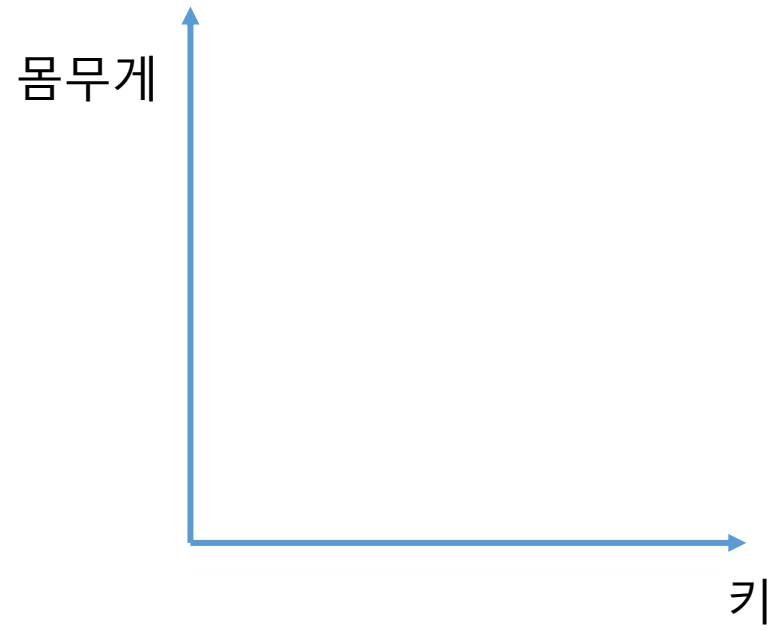
27 March 1857

[Islington](#), London, England

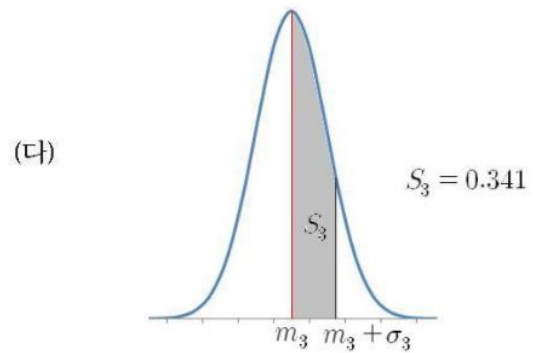
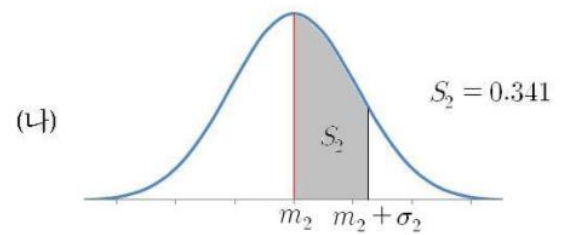
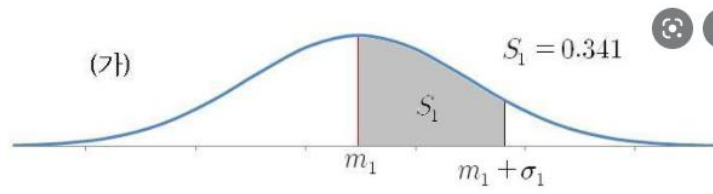


$$\text{COV}(X,Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

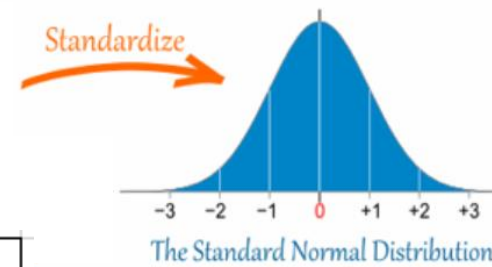
$$\text{Corr}(X,Y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$



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.

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

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 $-1 \leq r \leq 1$ 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

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W) 도움말(H)									
<div> <div> </div> <div> <div> <div>보고서(P)</div> <div>기술통계량(E)</div> <div>표</div> <div>평균 비교(M)</div> <div>일반선형모형(G)</div> <div>일반화 선형 모형(Z)</div> <div>혼합 모형(X)</div> <div>상관분석(C)</div> <div>회귀분석(R)</div> <div>로그선형분석(O)</div> <div>신경망(W)</div> <div>분류분석(Y)</div> <div>차원 감소(D)</div> <div>척도(A)</div> <div>비모수 검정(N)</div> <div>예측(T)</div> <div>생존확률(S)</div> <div>다중응답(U)</div> <div>결측값 분석(V)</div> <div>다중 대입(T)</div> <div>복합 표본(L)</div> <div>시뮬레이션...</div> <div>품질 관리(Q)</div> <div>ROC 곡선(V)</div> </div> <div> <div>single</div> <div>married</div> <div>married</div> </div> <div> <div>secondary</div> <div>secondary</div> <div>tertiary</div> </div> </div> </div>									
	age	job	default	balance	housing	loan	contact	day	mon
1	30	unemployed	no	1787	no	no	cellular	19	oct
2	33	services	no	4789	yes	yes	cellular	11	may
3	35	management	no	1350	yes	no	cellular	16	apr
4	30	management	no	1476	yes	yes	unknown	3	jun
5	59	blue-collar	no	0	yes	no	unknown	5	may
6	35	management	no	747	no	no	cellular	23	feb
7	36	self-employed	no	307	yes	no	cellular	14	may
8	39	technician	no	147	yes	no	cellular	6	may
9	41	entrepreneur	no	221	yes	no	unknown	14	may
10	43	services	no	-88	yes	yes	cellular	17	apr
11	39	services	no	9374	yes	no	unknown	20	may
12	43	admin.	no	264	yes	no	cellular	17	apr
13	36	technician	no	1109	no	no	cellular	13	aug
14	20	student	no	502	no	no	cellular	30	apr
15	31	blue-collar	no	360	yes	yes	cellular	29	jan
16	40	management	no	194	no	yes	cellular	29	aug
17	56	technician	no	4073	no	no	cellular	27	aug
18	37	admin.	no	2317	yes	no	cellular	20	apr
19	25	blue-collar	no	-221	yes	no	unknown	23	may
20	31	services	no	132	no	no	cellular	7	jul
21	38	management	no	0	yes	no	cellular	18	nov
22	42	management	no	16	no	no	cellular	19	nov
23	44	services	no	106	no	no	unknown	12	jun
24	44	entrepreneur	no	93	no	no	cellular	7	jul
25	26	housemaid	no	543	no	no	cellular	30	ian

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 balance는 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