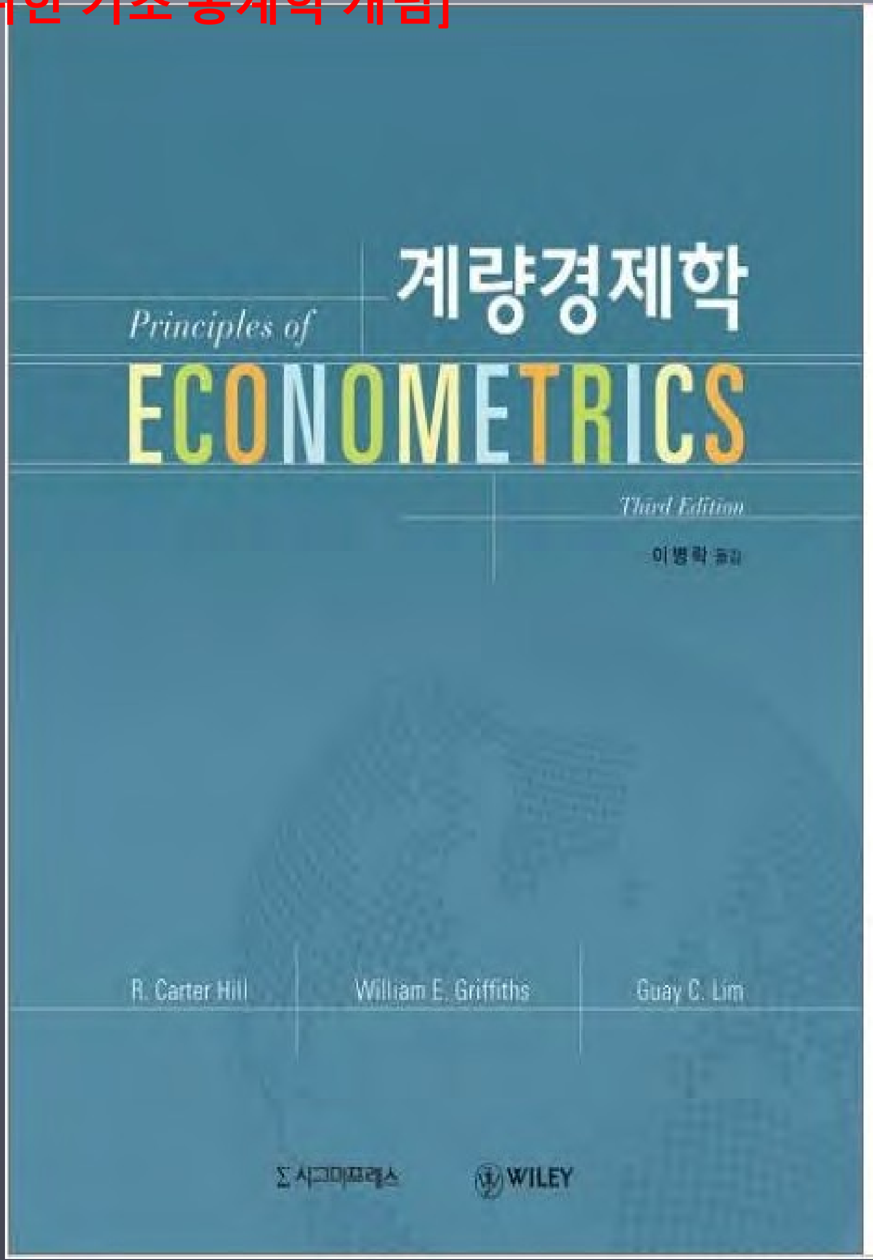
[계량경제학을 위한 기초 통계학 개념]



제0장
확률개념의 소개

1. 확률 (probability)

- Probability(확률) is the likelihood or chance that something is the case or will happen.
- random experiment (임의실험)
  : 발생 가능한 결과들 중 하나가 임의적으로 결정되는 과정
- event (사건, 사상)
  : 임의실험으로 얻게 되는 특정 결과들의 모임

$$P(A) = \lim_{N \to \infty} \frac{N_A}{N}$$

N_A : 특정 사건 A가 발생할 횟수

N : 임의실험 횟수

2. 확률변수 (random variable)

- A **random** <u>variable</u> is a variable whose value is **unknown** *until* it is observed. (대문자)
- The *value* of a **random** variable results from an experiment; it is not perfectly predictable. (소문자)

(예) 공장에서 생산한 두 개의 제품의 불량 여부 임의실험

표본공간: $\Omega = \{(불, 불),\ (불, 정),\ (정, 불),\ (정, 정)\}$ 확률

변수 X : 불량품의 수

확률변수의 값: $x = 0,\ 1,\ 2$

<이산적 확률변수>

- A **discrete** random variable can take only a finite number of values, that can be counted by using the positive integers.

Example: Prize money from the following lottery is a discrete random variable:

first prize: \$1,000
second prize: \$50
third prize: \$5.75

since it has only four (a finite number) (count: 1,2,3,4) of possible outcomes:

\$0.00; \$5.75; \$50.00; \$1,000.00

<연속적 확률변수>

- A **continuous** random variable can take any real value (not just whole numbers) in at least one interval on the real line.

- Examples:

  Gross national product (GNP)
  money supply
  interest rates
  price of eggs
  household income
  expenditure on clothing

<더미변수>

- A discrete random variable that is restricted to two possible values (usually 0 and 1) is called a **dummy variable** (also, binary or indicator variable).

✓ Dummy variables account for qualitative differences:

(예) gender (0=male, 1=female)
race (0=white, 1=nonwhite)
citizenship (0=U.S., 1=not U.S.)
income class (0=poor, 1=rich)

3. 확률분포, 확률(밀도)함수

- **확률분포 (probability distribution)**

: 어떤 확률변수가 취할 수 있는 모든 가능한 값들에 대응하는 확률을 나타낸 것

- **이산적** 확률변수 경우와 **연속적** 확률변수 경우는 확률분포를 나타내는 방식이 조금 다름

- <표현방법>

- 그래프
- 도표
- **확률밀도함수** (probability denisty function; pdf)

<**이산적** 확률변수의 확률밀도함수>

- When the values of a discrete random variable are listed with their chances of occurring,
  the resulting table of outcomes is called
  a *probability function* or a *probability density function*.

(예) (X=동전 한 번 던져 나올 앞면의 수)의 확률분포

동전면	x	$f(x)$
앞면	1	0.5
뒷면	0	0.5

표본공간 $X = \{0, 1\}$
확률밀도함수 $f(x) = 0.5$

그래프로도 표현 가능

<**이산적** 확률변수의 확률밀도함수>

- For a discrete random variable *X* the value of the probability density function *f(x) is* the probability that the random variable *X* takes the value *x* , $f(x)=P(X=x)$.

(예) (*X*=주사위 던져 나올 윗면의 숫자)의 확률분포

die	x	f(x)
one dot	1	1/6
two dots	2	1/6
three dots	3	1/6
four dots	4	1/6
five dots	5	1/6
six dots	6	1/6

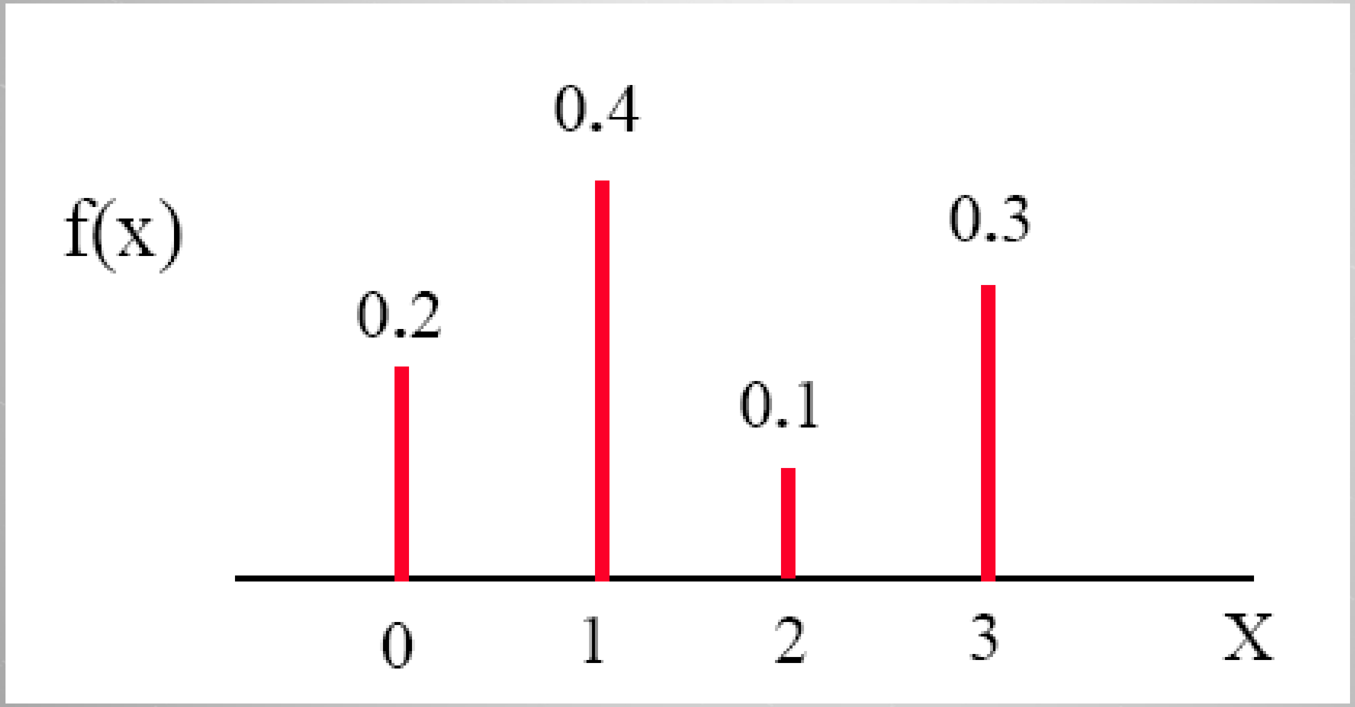
$$f(x)=P(X=x)$$

$$0 \le f(x) \le 1$$

$$\sum_{i=1}^{n} f(x_i) = 1$$

<**이산적** 확률변수의 확률밀도함수>

- Probability, f(x), for a discrete random variable, X, can be represented by **height.**



$f(0)=0.2$

$f(1)=0.4$

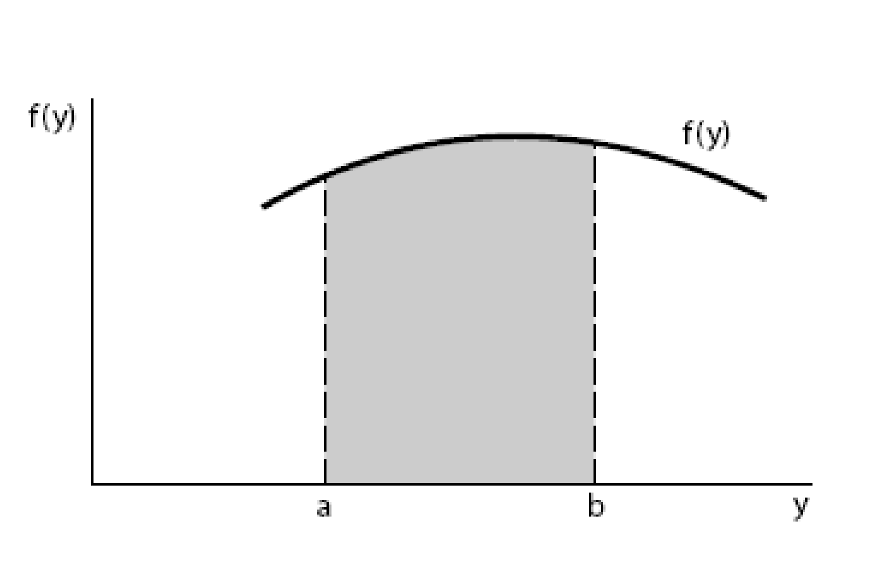
$f(2)=0.1$

$f(3)=0.3$

<**연속적** 확률변수의 확률밀도함수>

- For the continuous random variable Y the probability density function $f(y)$ can be represented by an *equation*, which can be described graphically by a curve.
- For continuous random variables the ***area*** under the probability density function corresponds to probability.

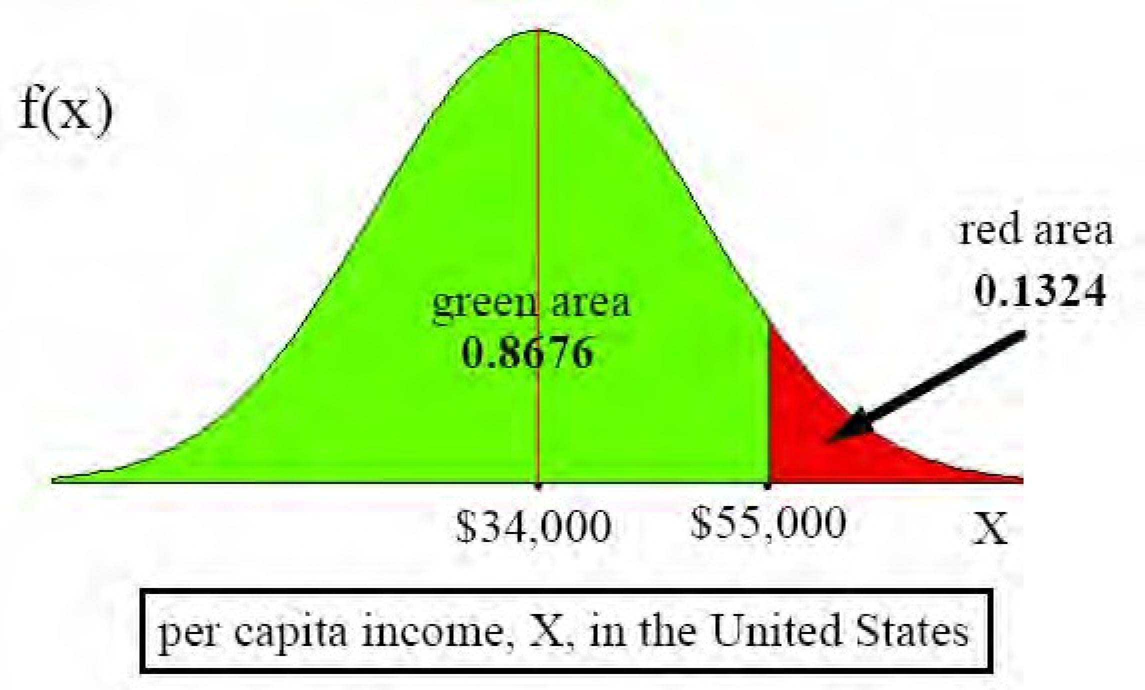
$$P(a \le Y \le b) = \int_a^b f(y)\ dy$$



<**연속적** 확률변수의 확률밀도함수>

- Probability is represented by **area.**
- Height alone has no **area.**
- An interval for X is needed to get an **area under the curve.**

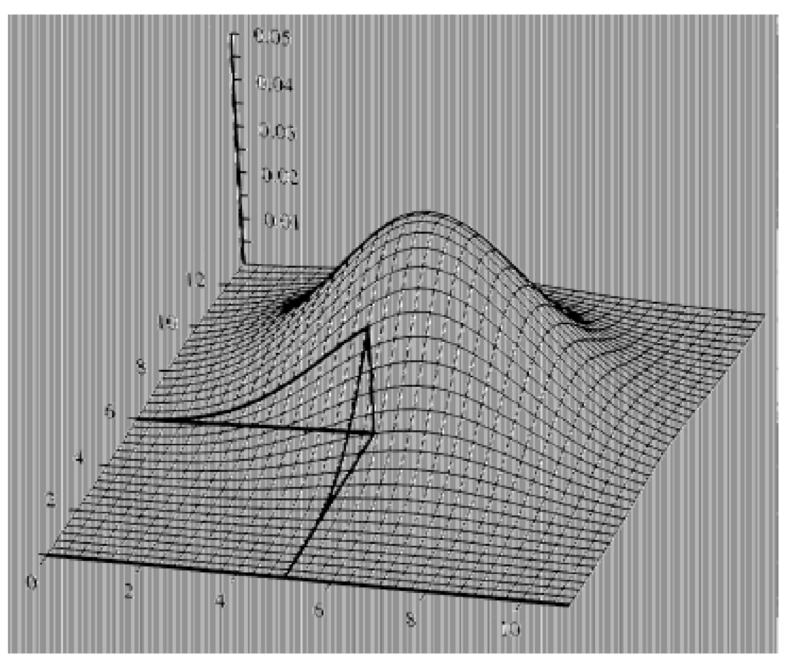
$$P(X \geq 55{,}000) = 0.1324$$



4. 결합확률분포 $f(x, y) = P(X = x, Y = y)$

- Given two random variables X and Y, the joint distribution of X and Y is the distribution of X and Y together.

	Y = 1	Y = 2
X = 0	f(0,1) .45	f(0,2) .15
X = 1	.05 f(1,1)	.35 f(1,2)



- 한계확률밀도함수(marginal probability density function)

$$f(x_i) = \sum_j f(x_i, y_j) \qquad f(y_j) = \sum_i f(x_i, y_j)$$

	Y = 1	Y = 2	marginal pdf for X:
X = 0	.45	.15	.60 f(X = 0)
X = 1	.05	.35	.40 f(X = 1)
marginal pdf for Y:	.50 f(Y = 1)	.50 f(Y = 2)	

- **통계적으로 독립 (Independence)?**
- random variables are **independent** if their joint pdf is the product of their respective marginal pdfs.

$$f(x_i, y_j) = f(x_i) f(y_j)$$

not independent

	Y = 1	Y = 2	marginal pdf for X:
X = 0	.50x.60=.30 .45	.50x.60=.30 .15	.60 f(X = 0)
X = 1	.05 .50x.40=.20	.35 .50x.40=.20	.40 f(X = 1)
marginal pdf for Y:	.50 f(Y = 1)	.50 f(Y = 2)	

The calculations in the boxes show the numbers required to have independence.

5. 조건부확률분포

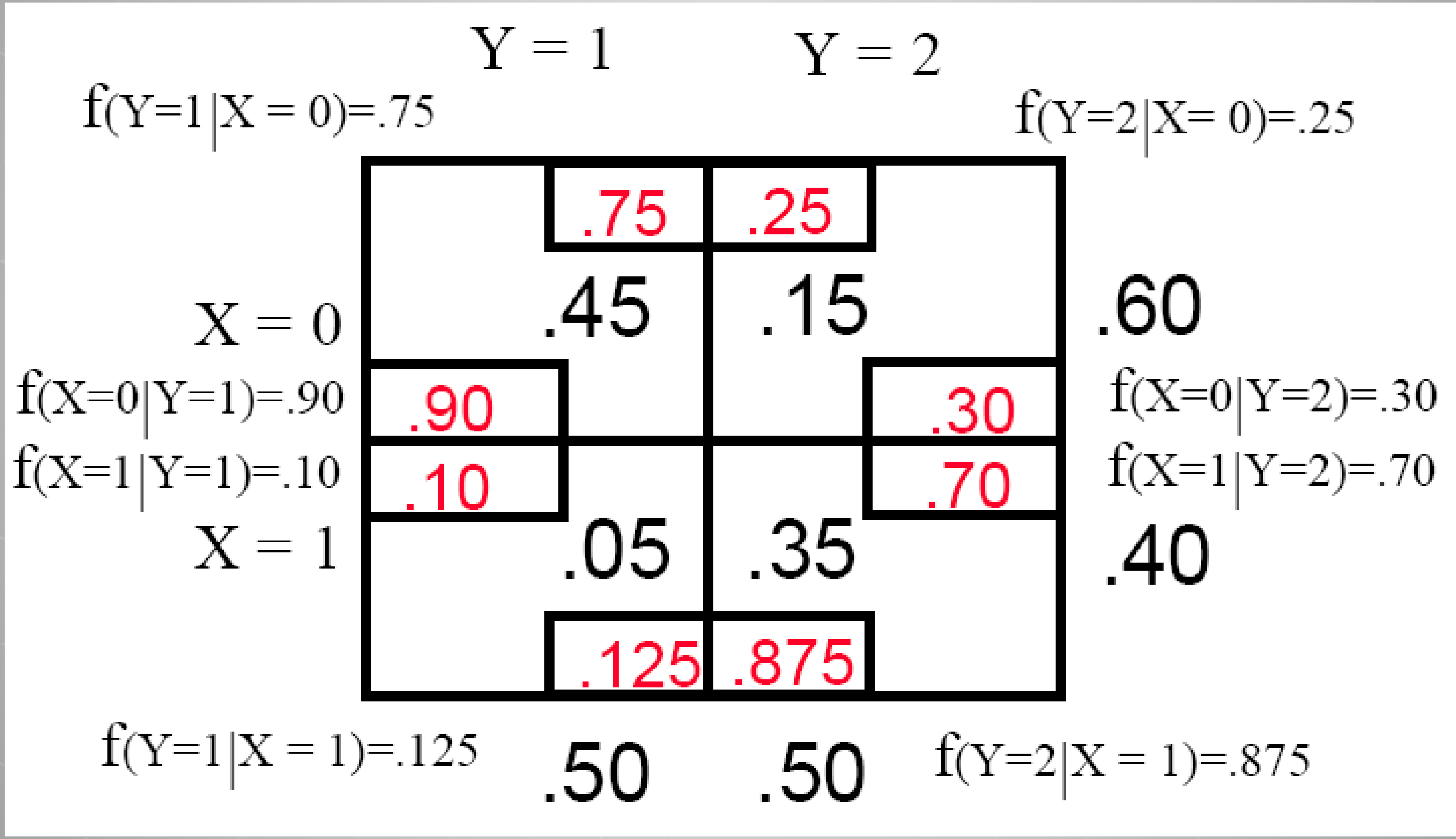
- **조건부**확률밀도함수(conditional PDF)

  : 확률변수 Y 가 어떤 특정한 값 y 를 취한 것이 전제가 된 상태에서 확률변수 X 가 어떤 특정한 값 x 를 취할 조건부확률

$$f(x \mid y) = P(X = x \mid Y = y)$$

$$f(x \mid y) = \frac{f(x, y)}{f(y)}$$

conditional PDF



6. 모집단과 표본

- **모집단 (population)**

  : 연구대상의 전체집단

  - 유한모집단 (finite population)

  - 무한모집단 (infinite population)

- **표본 (sample)**

  : 모집단의 일부

  : 모집단과 가장 유사한 모습(특성)을 가질수록 좋음

  * 임의표본(random sample)

✓ 계량경제학이 분석하는 경제통계자료는 거의 대부분이 임의표본임

7. 모수와 통계량

- **모수 (parameter)**

  : 모집단의 어떤 특성을 수치로 나타낸 것(통계치)

  : 전수조사를 하지 않는 이상 알아낼 수 없음, 미지수

  (예) 모평균, 모분산, 모표준편차, 모비율 등

- **통계량 (statistic)**

  : 표본의 어떤 특성을 수치로 나타낸 것, 표본의 통계치

  : 표본에 어떻게 뽑히는가에 따라 변동하는 확률변수

  (예) 표본평균, 표본분산, 표본표준편차, 표본비율 등

cf. **통계학 (statistics)**

  : 통계량을 이용하여 미지의 모수를 추정하는 학문

8. 평균 (mean)

- mean or arithmetic average of a random variable

  = mathematical expectation or expected value

- 유한 모집단에서의 평균: $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$

- 무한 모집단에서의 평균: $\mu_x = E(X) = \sum_{i=1}^{N} x_i P(X = x_i)$

- 표본의 평균: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

[강의 보충 과제물]

<기초통계 관련 유용한 동영상 4가지> 복습

통계 공식과 개념의 기초 정리
https://www.youtube.com/watch?v=CQA7cdxozHY ☜ 확률분포 / 평균/ 분산 / 표준편차?
정규분포 / 표준화공식?
이항분포 to 정규분포?
모집단과 표본 / 임의추출 n개 하면?
통계적추정 기본식?

표본통계량의 분포1
https://www.youtube.com/watch?v=3lajUt_Alc8 ☜ 표본통계량의 정규분포 to t분포? ,
자유도? 등

표본통계량의 분포2
https://www.youtube.com/watch?v=jIrZHqAZkBs&t=32s ☜ 책 뒤의 통계분포 표 보는 법?

t분포 및 F분포
https://www.youtube.com/watch?v=KyxeBT9XUUU ☜ 계량경제학에서 가장 많이 쓰이는
두 가지 중요한 통계분포에 대한
자세한 소개?

9. 분산 (variance), 표준편차 (standard deviation)

- **모집단**에서의 **분산**:

$$\sigma_x^2 = Var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\sigma_x^2 = \sum_{i=1}^{N}(x_i - E(X))^2 P(X = x_i)$$

- **표본**에서의 **분산**:

$$s_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

- **표준편차**: $\sigma_x = \sqrt{\sigma_x^2}$ (**모집단**)

$s_x = \sqrt{s_x^2}$ (**표본**)

10. 기대치

- **기대치 (expected value)**는 같은 일이 무한히 반복될 때, 해당 **확률변수의 평균**
- 산술평균을 나타냄
- **Empirical** sample mean:

$$E(X) = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

N: 관측치의 수

- **Analytical** sample mean:

$$E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

n: 가능한 수치 그룹의 수

<기대치 관련 공식>

The expected value of X:

$$E\,X = \sum_{i=1}^{n} x_i\, f(x_i)$$

The expected value of X-squared:

$$E\,X^2 = \sum_{i=1}^{n} x_i^2\, f(x_i)$$

It is important to notice that $f(x_i)$ does not change!

The expected value of X-cubed:

$$E\,X^3 = \sum_{i=1}^{n} x_i^3\, f(x_i)$$

<기대치 관련 공식>

① $E(a) = a$ (a는 상수)

② $E(aX) = aE(X)$

③ $E(a + bX) = a + bE(X)$

④ $E(X + Y) = E(X) + E(Y)$ $E(X - Y) = E(X) - E(Y)$

11. 분산

- **분산 (variance)**

  : 확률변수의 값들이 **중심으로부터 얼마나 퍼져 있는가**를 나타냄

- 계산식

$$Var(X) = E[(X-\mu)^2]$$
$$= E[X^2 - 2\mu X + \mu^2]$$
$$= E(X^2) - 2\mu E(X) + \mu^2$$
$$= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$$

여기서. $\mu = E(X)$

<분산 계산 방법: 예시>

x_i	$f(x_i)$	$(x_i - EX)$	$(x_i - EX)^2 f(x_i)$
2	.1	2 - 4.3 = -2.3	5.29 (.1) = .529
3	.3	3 - 4.3 = -1.3	1.69 (.3) = .507
4	.1	4 - 4.3 = - .3	.09 (.1) = .009
5	.2	5 - 4.3 = .7	.49 (.2) = .098
6	.3	6 - 4.3 = 1.7	2.89 (.3) = .867

$$\sum_{i=1}^{n} x_i f(x_i) = .2 + .9 + .4 + 1.0 + 1.8 = 4.3$$

$$\sum_{i=1}^{n} (x_i - EX)^2 f(x_i) = .529 + .507 + .009 + .098 + .867 = 2.01$$

<분산 관련 공식>

① $Var(X) \geq 0$

② $Var(a) = 0$ (a는 상수)

③ $Var(X+a) = Var(X)$

④ $Var(aX) = a^2 Var(X)$

⑤ $Var(a+bX) = b^2 Var(X)$

⑥ $Var(X \pm Y) = Var(X) + Var(Y) \pm 2\operatorname{cov}(X,Y)$

12. 정규분포

- 정규분포 (normal distribution)

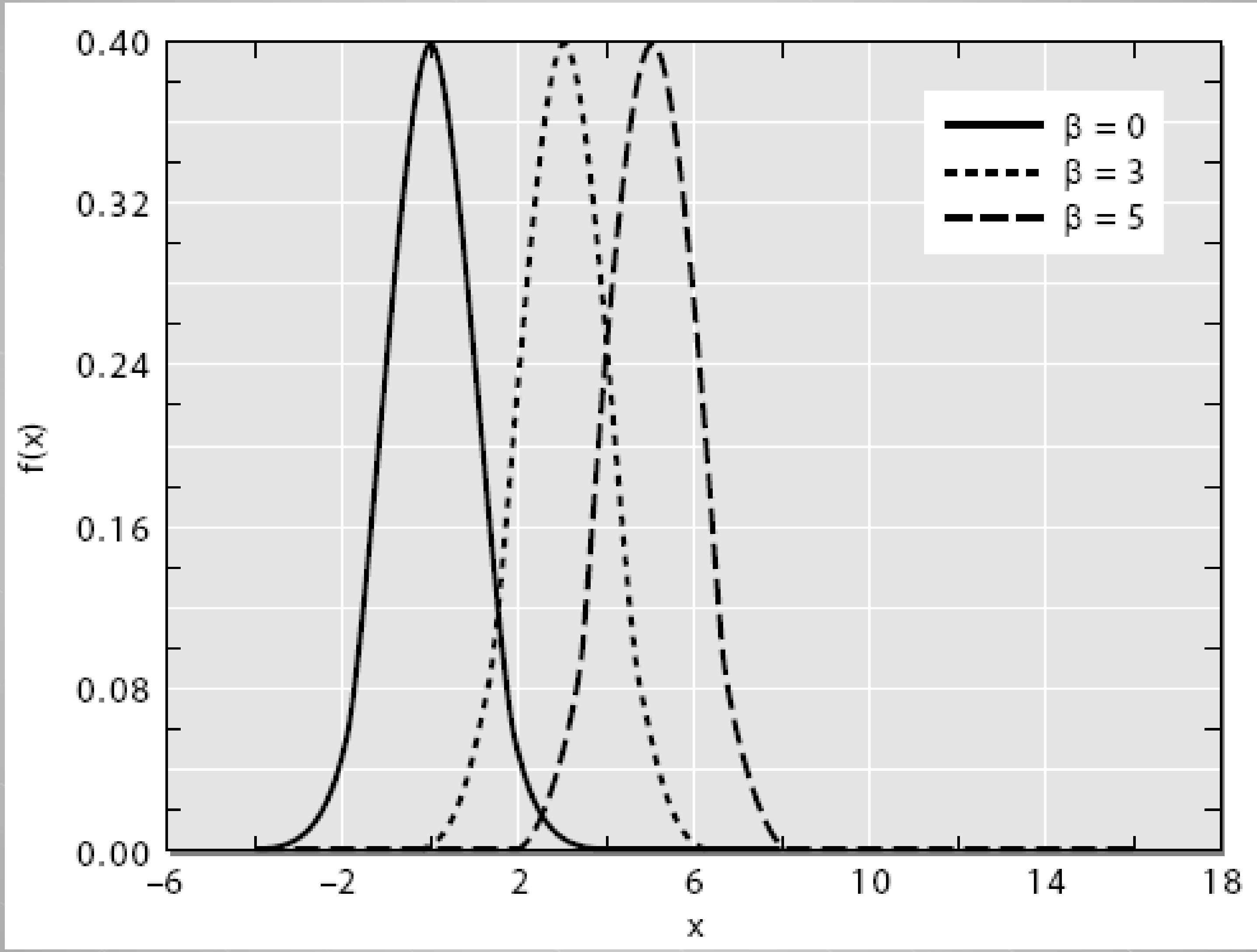
$$X \sim N(\beta,\ \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\beta)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty$$

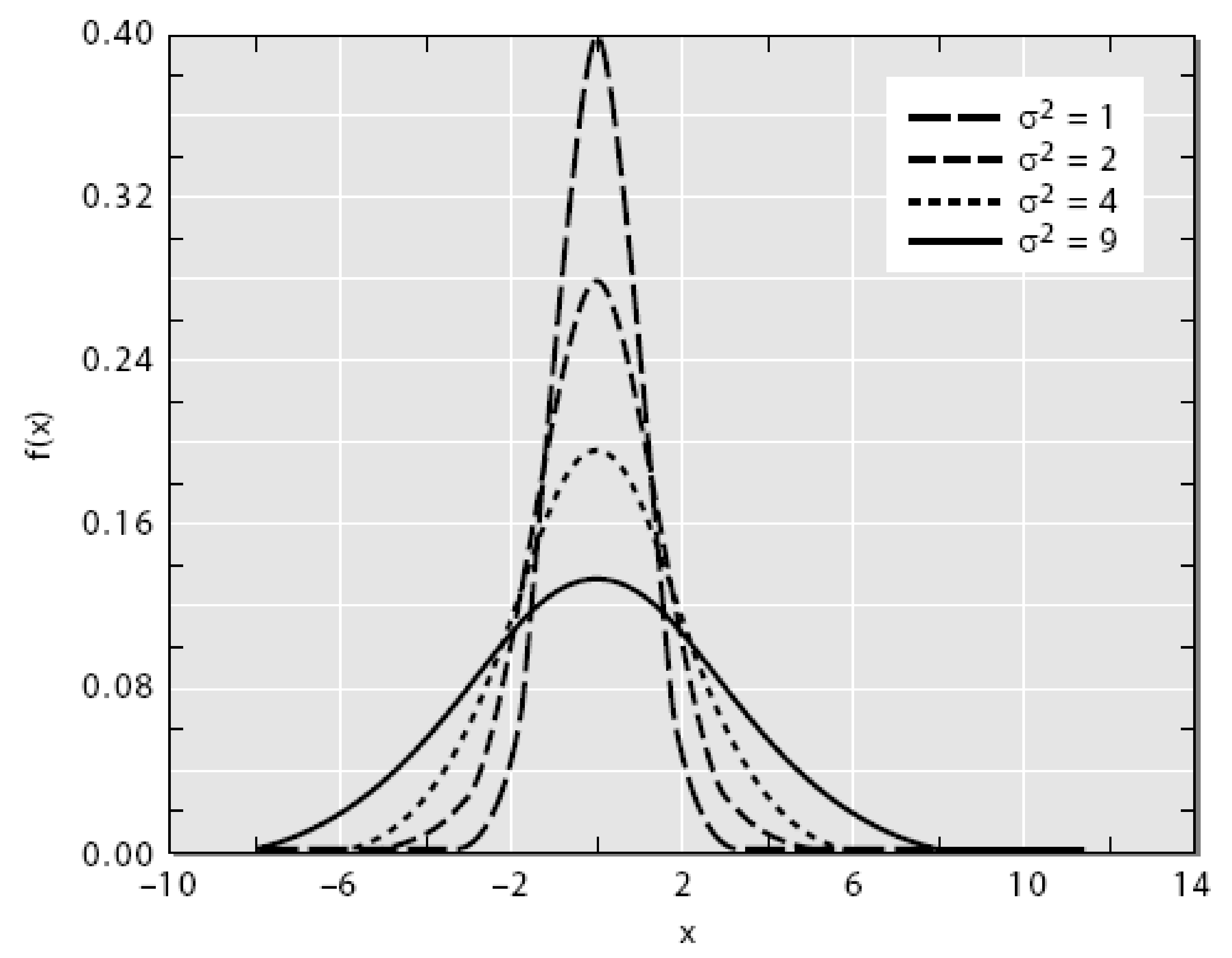
- 표준정규분포 (standard normal distribution)

$$Z = \frac{X-\beta}{\sigma} \sim N(0,1)$$

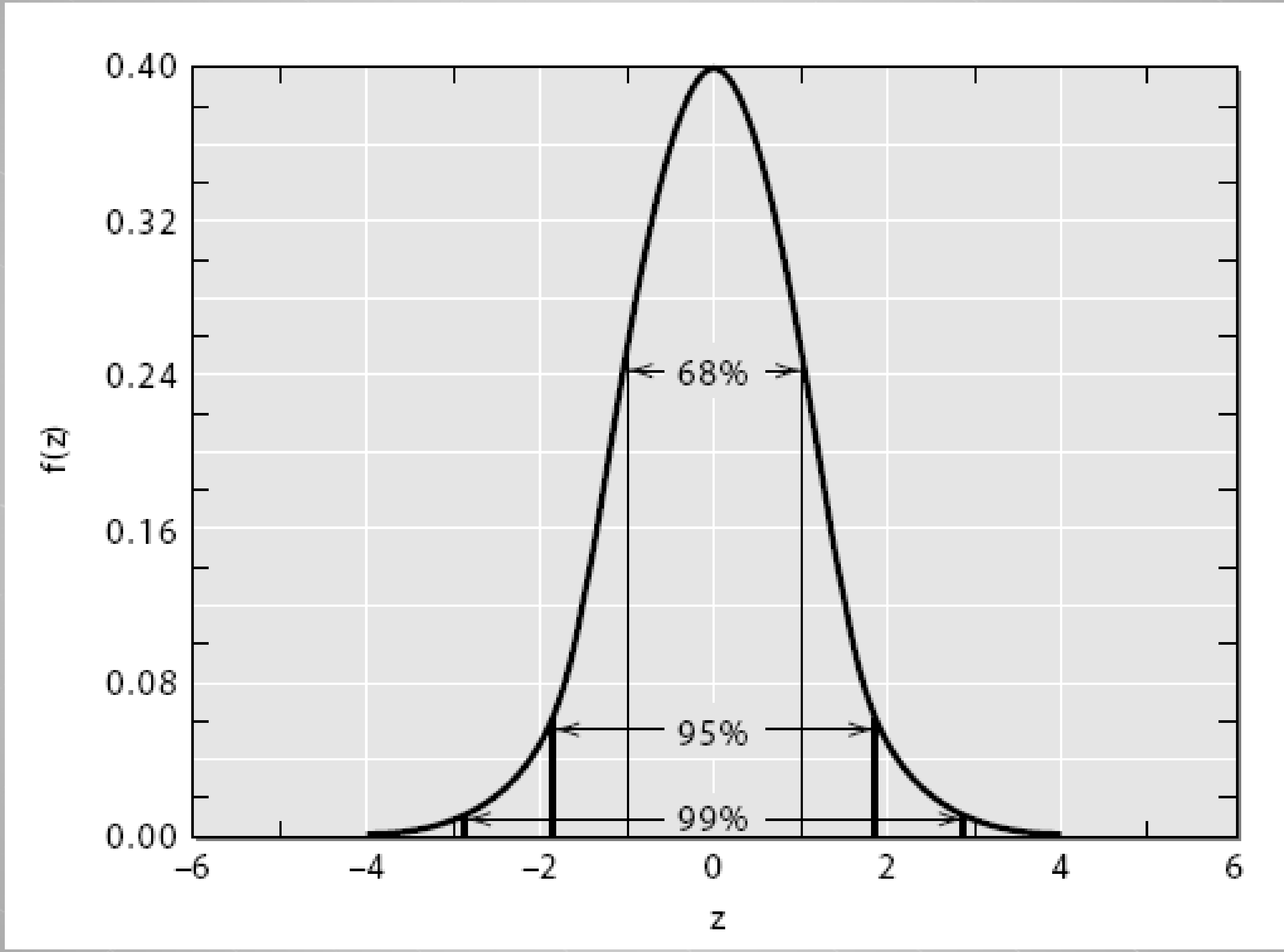
$$X \sim N(\beta, 1^2)$$



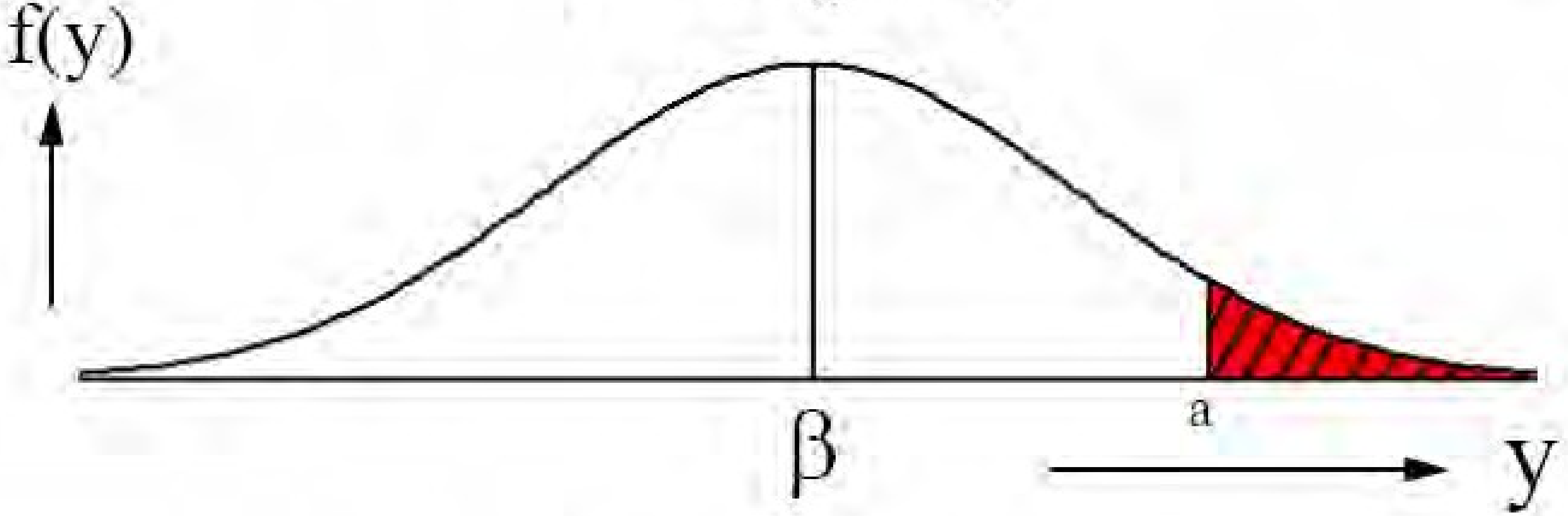
$$X \sim N(0,\ \sigma^2)$$



$$X \sim N(0, 1^2)$$

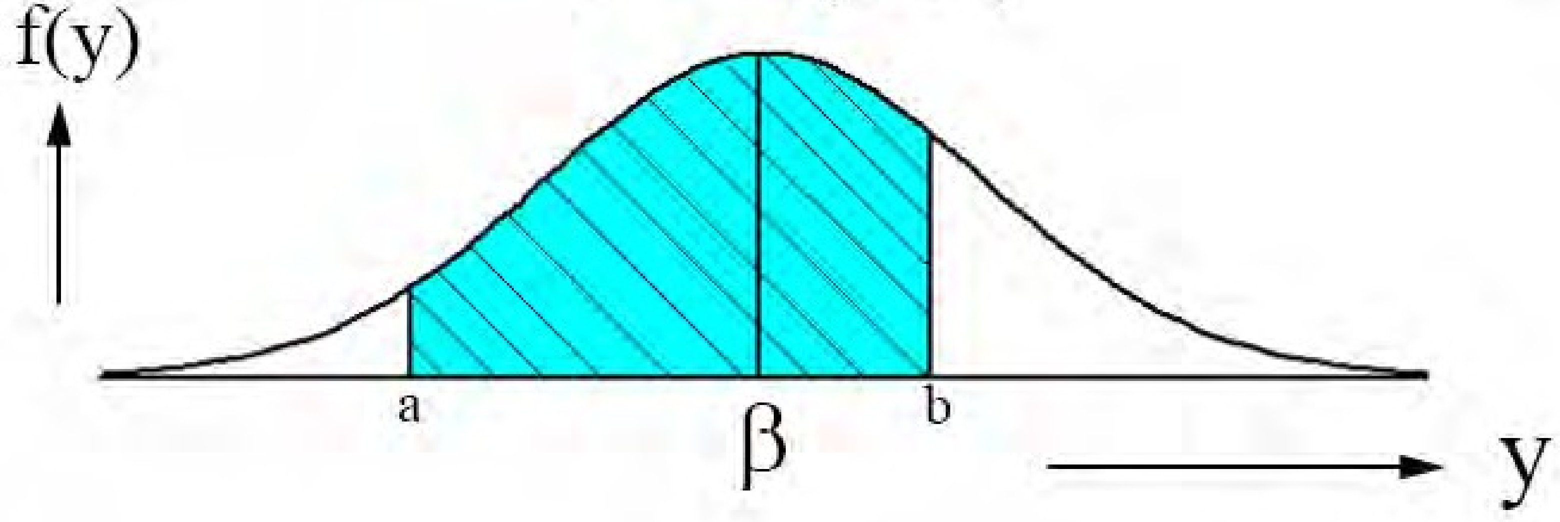


$$Y \sim N(\beta, \sigma^2)$$



$$P[Y \geq a] = P\left[\frac{Y-\beta}{\sigma} \geq \frac{a-\beta}{\sigma}\right] = P\left[Z \geq \frac{a-\beta}{\sigma}\right]$$

$$Y \sim N(\beta, \sigma^2)$$



$$P\left[a \le Y \le b\right] = P\left[\frac{a-\beta}{\sigma} \le \frac{Y-\beta}{\sigma} \le \frac{b-\beta}{\sigma}\right]$$

$$= P\left[\frac{a-\beta}{\sigma} \le Z \le \frac{b-\beta}{\sigma}\right]$$

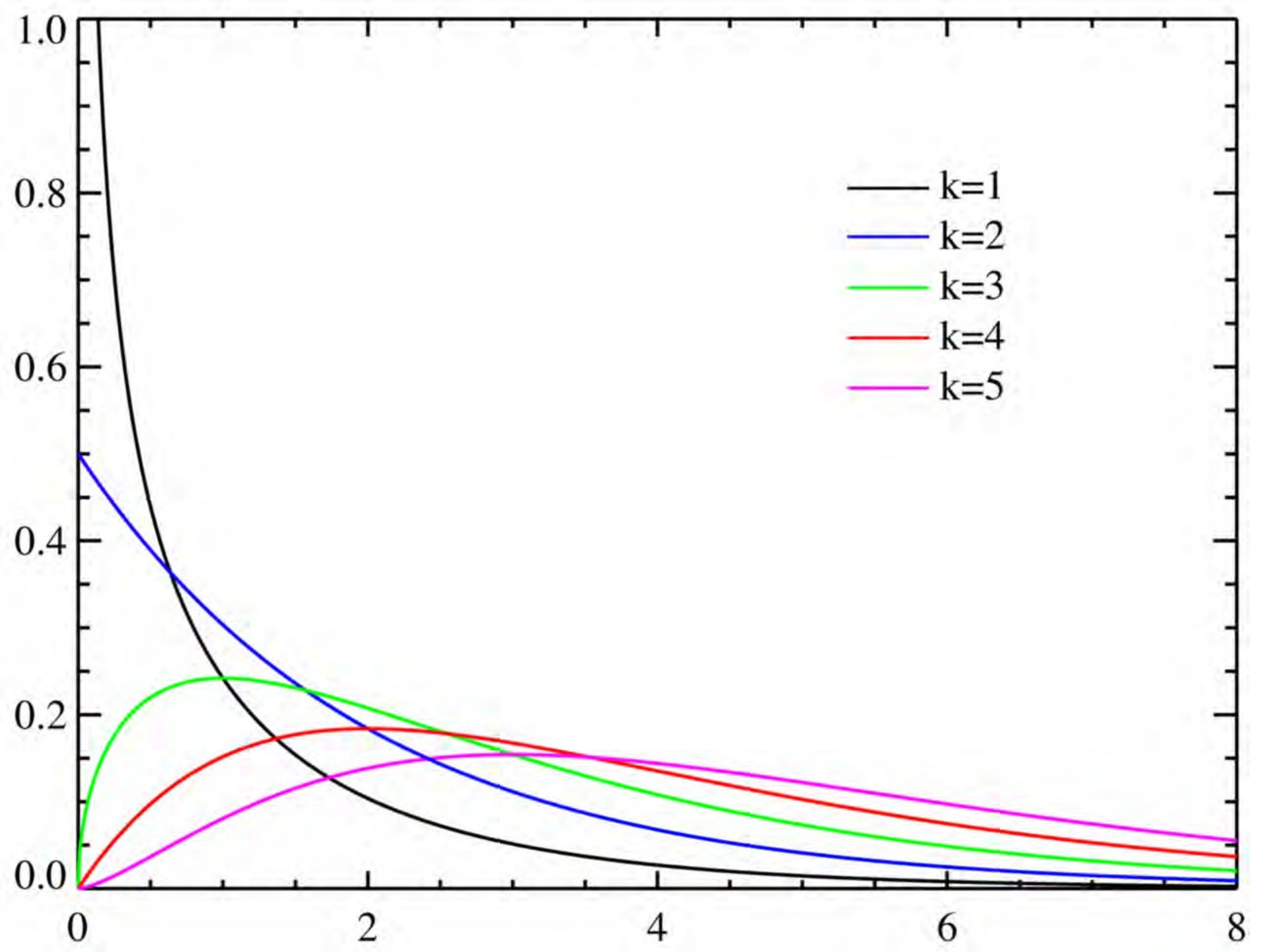
13. 카이제곱 분포

- χ^2 distribution, chi-square distribution

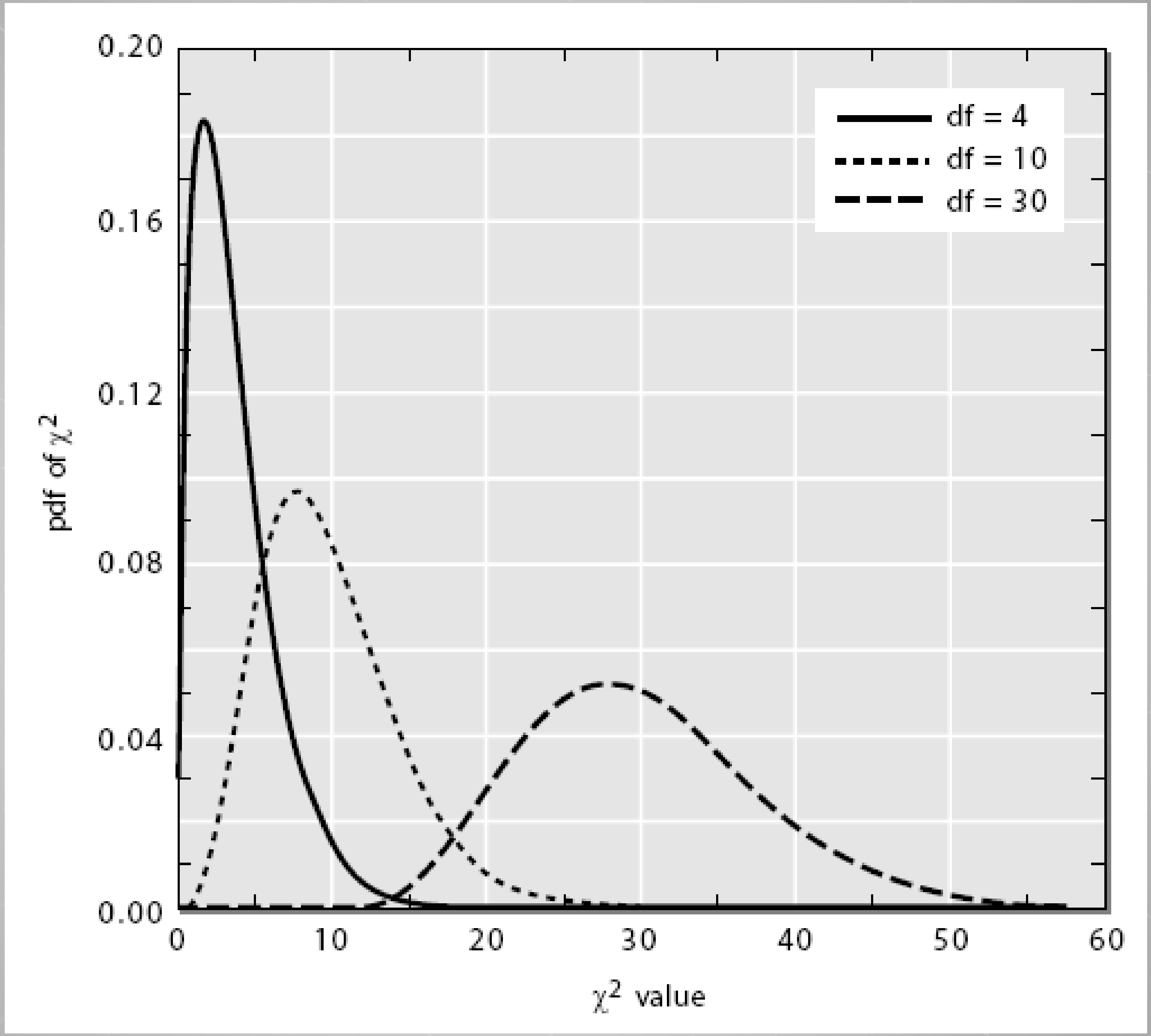
  : 확률변수 Z_i 가 표준정규분포를 따를 때 이 변수들의 제곱의 합 $\sum_{i=1}^{q} Z_i^2$ 은 자유도가 q 인 χ^2–분포를 함

- χ^2 -분포의 모양은 **자유도(degrees of freedom)**에 따라 달라짐

- 자유도가 커질수록 정규분포에 가까운 모양을 가짐


1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
k=1
k=2
k=3
k=4
k=5

$\chi^2(q)$


0.20
0.16
0.12
0.08
0.04
0.00
pdf of χ2
0
10
20
30
40
50
60
χ2 value
df = 4
df = 10
df = 30

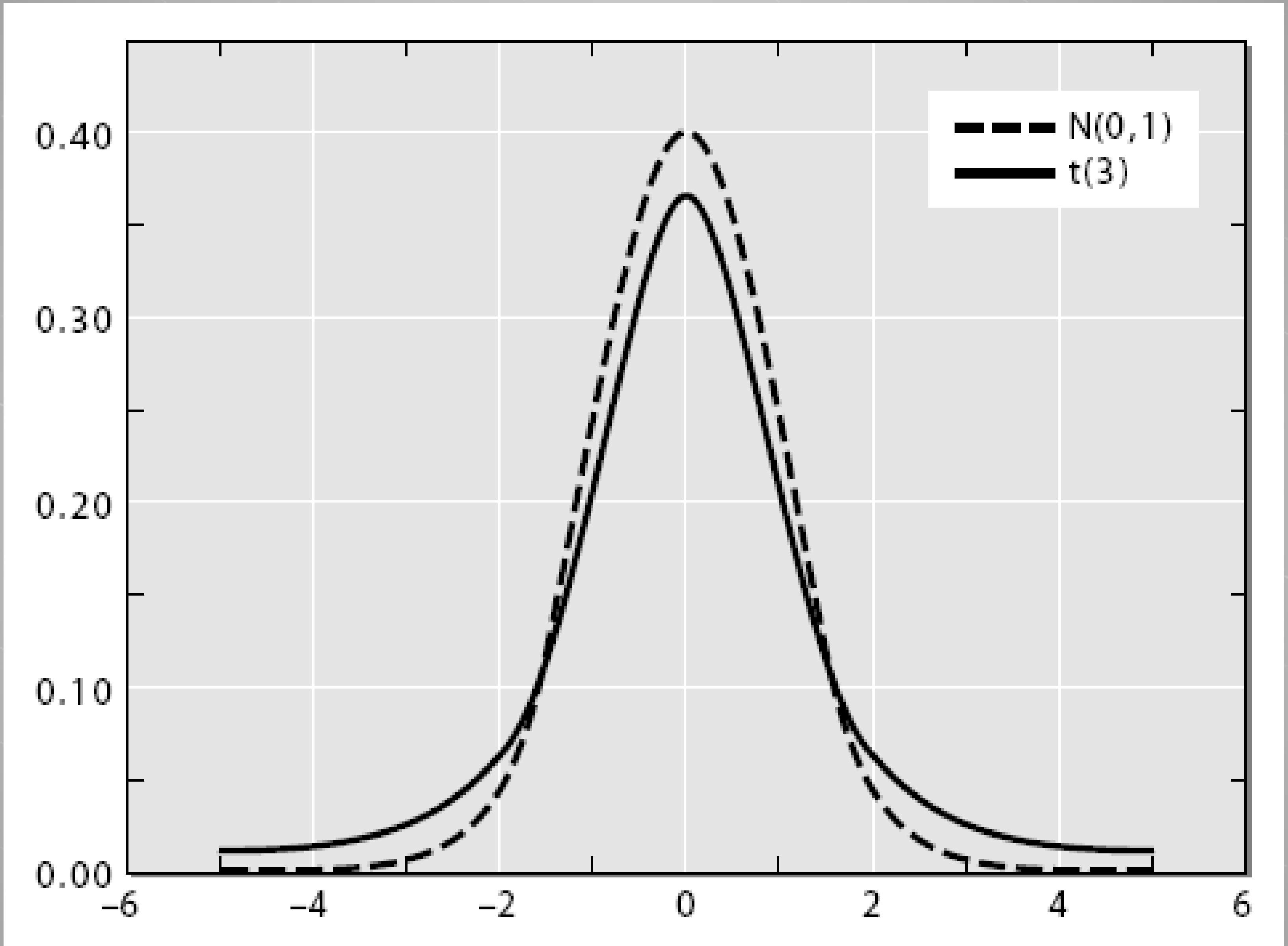
14. t-분포

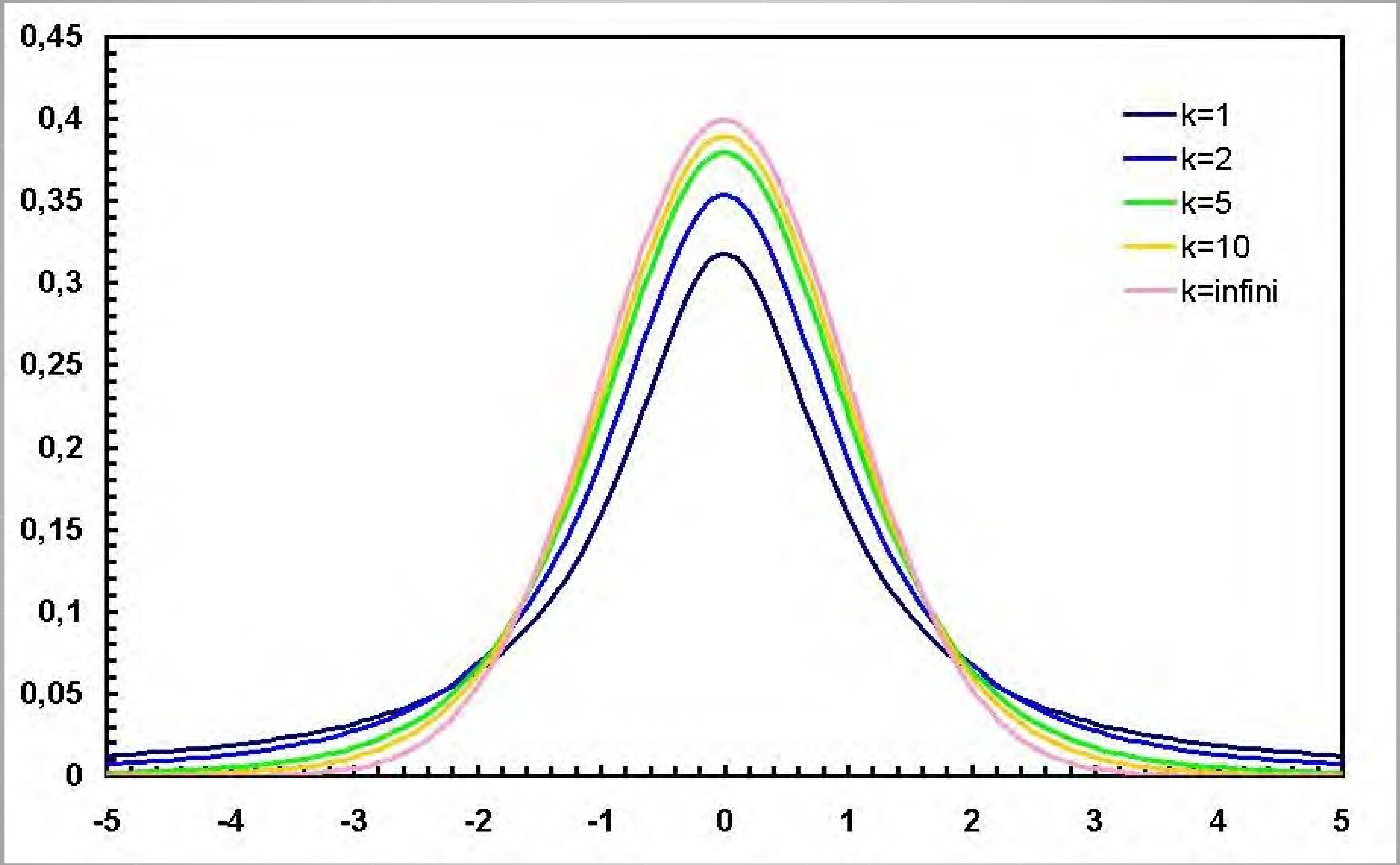
- Student's t distribution

If $Z \sim N(0,1)$ and $V \sim \chi^2_{(m)}$, and if Z and V are independent, then

$$t = \frac{Z}{\sqrt{V/m}} \sim t_{(m)}$$

- t-분포의 모양은 $N(0, 1^2)$ 보다 덜 뾰족하고 꼬리가 두터움 (fat-tail)
- 자유도 $m \rightarrow \infty$ 함에 따라 t-분포는 $N(0, 1^2)$ 에 수렴함


N(0,1)
t(3)
0.40
0.30
0.20
0.10
0.00
−6
−4
−2
0
2
4
6

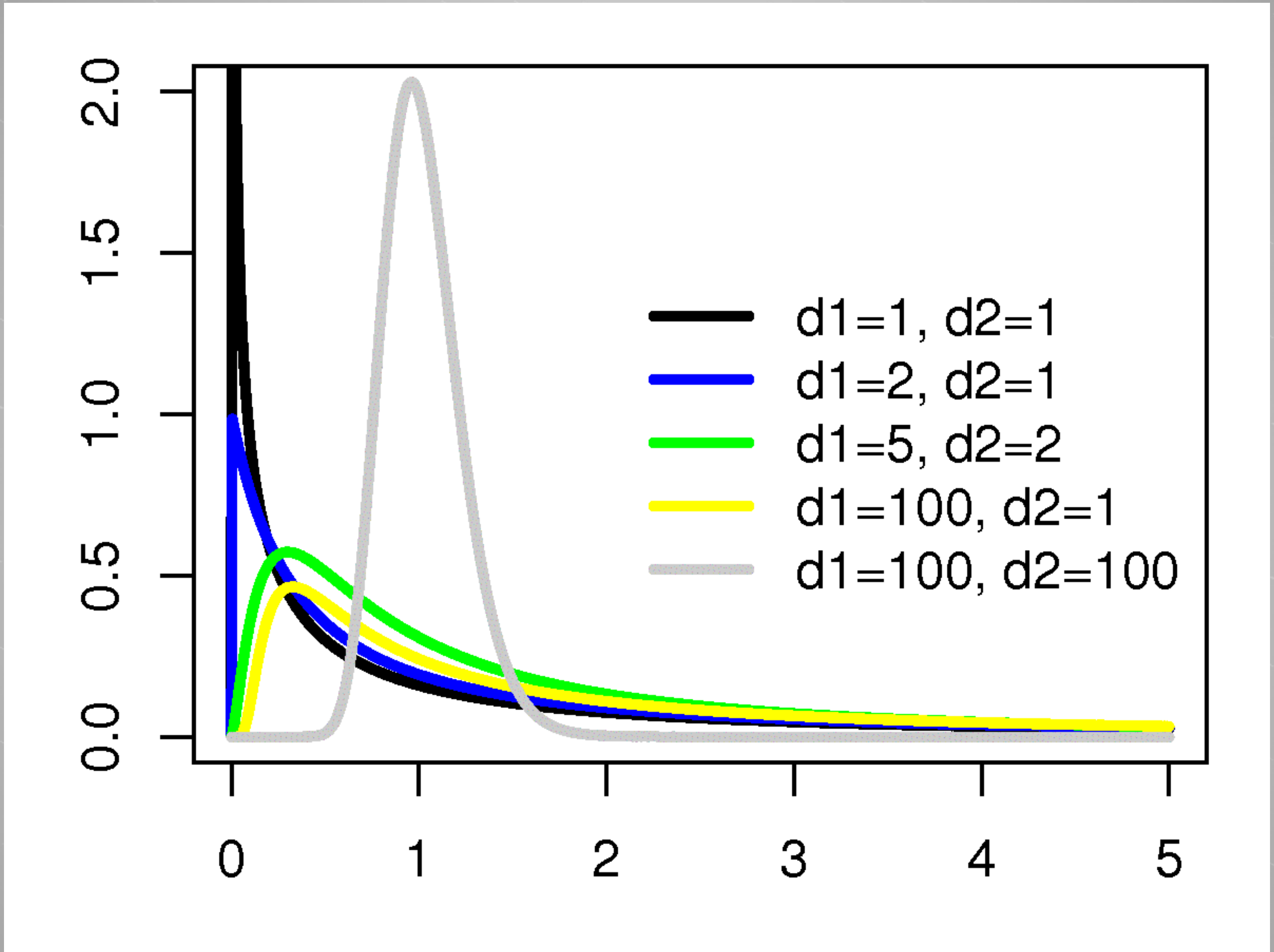

0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
k=1
k=2
k=5
k=10
k=infini

15. F-분포

- F-distribution

: 각각 자유도가 q_1, q_2 인 카이제곱분포를 갖는 두 확률 변수의 비율은 F-분포를 따름

$$F_{q_1,q_2} = \frac{X_1/q_1}{X_2/q_2}$$


2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=100, d2=1
d1=100, d2=100

16. Rules of Summation

Rule 1: $$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Rule 2: $$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$

Rule 3: $$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

16. Rules of Summation (continued)

Rule 4: $$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

Rule 5: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

The definition of $\bar{x}$ as given in Rule 5 implies the following important fact:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

16. Rules of Summation (continued)

Rule 6: $$\sum_{i=1}^{n} f(x_i) = f(x_1) + f(x_2) + \ldots + f(x_n)$$

Notation: $$\sum_{x} f(x_i) = \sum_{i} f(x_i) = \sum_{i=1}^{n} f(x_i)$$

Rule 7: $$\sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i,y_j) = \sum_{i=1}^{n} [\, f(x_i,y_1) + f(x_i,y_2) + \ldots + f(x_i,y_m)]$$

The order of summation does not matter :

$$\sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i,y_j) = \sum_{j=1}^{m} \sum_{i=1}^{n} f(x_i,y_j)$$

17. 공분산, $Cov(X,Y)$

- 유한 모집단의 공분산: $$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

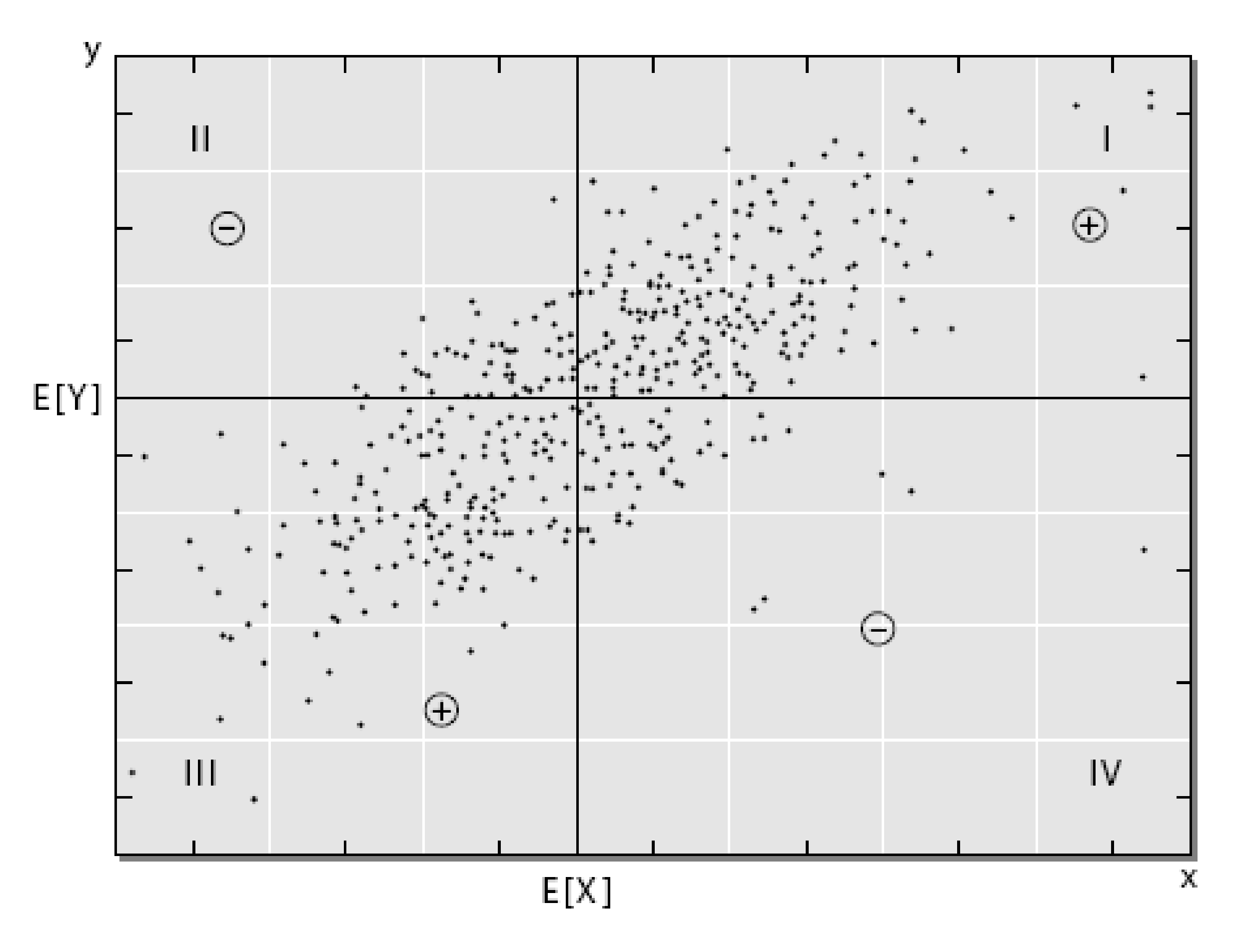
- 무한 모집단의 공분산: $$\sigma_{XY} = E[\{X - E(X)\}\{Y - E(Y)\}]$$

- 표본의 공분산: $$s_{XY} = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

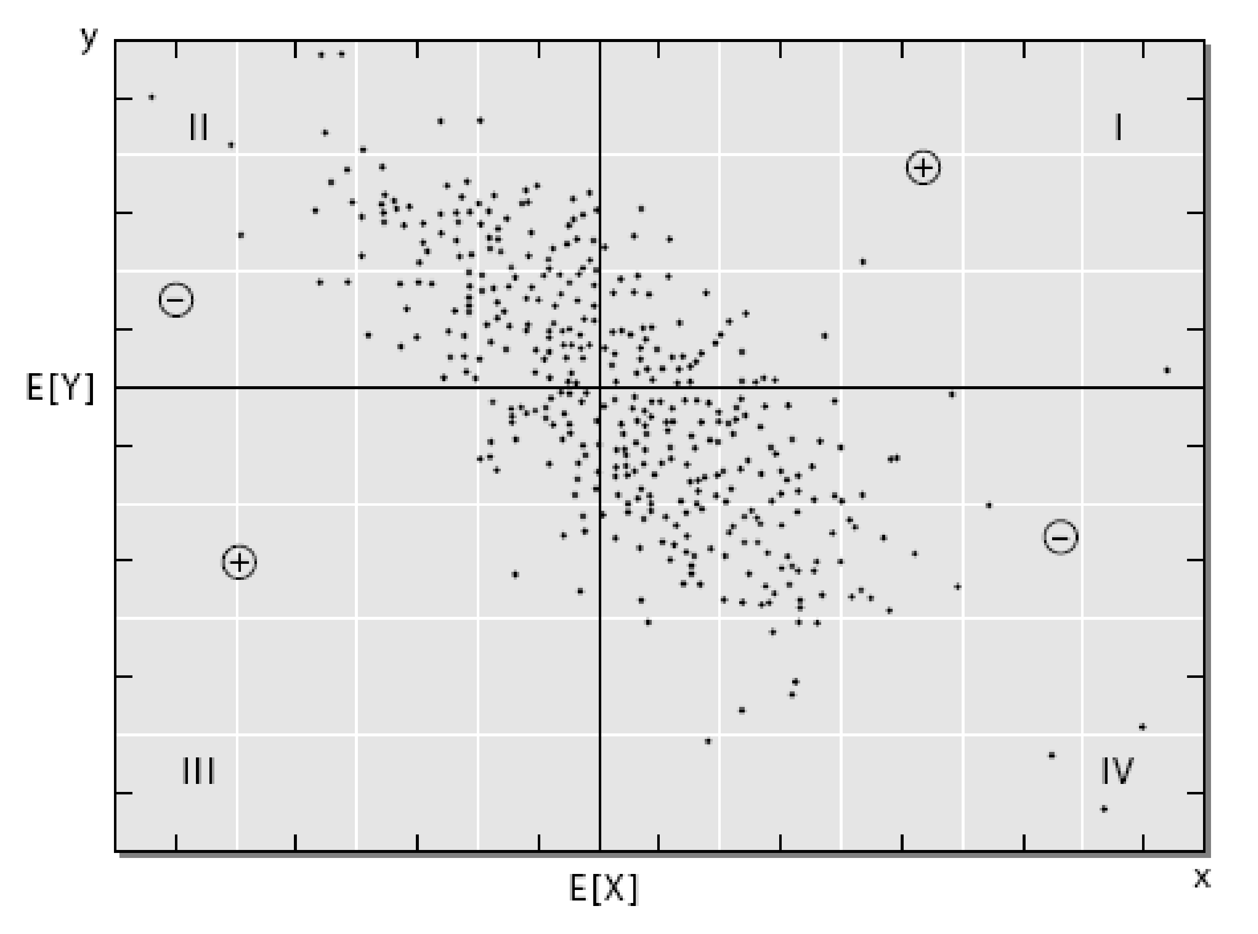
➢ 분산(variance)은 공분산(covariance)의 특수한 경우임

$$Var(X) = Cov(X, X) = E[\{X - E(X)\}\{X - E(X)\}] = E[X - E(X)]^2$$

- 양의 공분산: $\sigma_{XY} > 0,\ E[\{X - E(X)\}\{Y - E(Y)\}] > 0$



- 음의 공분산: $\sigma_{XY} < 0,\ E[\{X - E(X)\}\{Y - E(Y)\}] < 0$



- 공분산 계산식

$$
\begin{aligned}
Cov(X,Y) &= E\,[\{X - E(X)\}\{Y - E(Y)\}] \\
&= E\,[XY - X\,E(Y) - Y\,E(X) + E(X)\,E(Y)] \\
&= E(XY) - E(X)\,E(Y) - E(Y)\,E(X) + E(X)\,E(Y) \\
&= E(XY) - 2\,E(X)\,E(Y) + E(X)\,E(Y) \\
&= E(XY) - EX\,EY
\end{aligned}
$$

$$Cov(X,Y) = E(XY) - E(X)\,E(Y)$$

<공분산 계산 방법: 예시>

	$Y = 1$	$Y = 2$	
$X = 0$	.45	.15	.60
$X = 1$	.05	.35	.40
	.50	.50	

$EX=0(.60)+1(.40)=.40$

$EY=1(.50)+2(.50)=1.50$

$EX\ EY = (.40)(1.50) = .60$

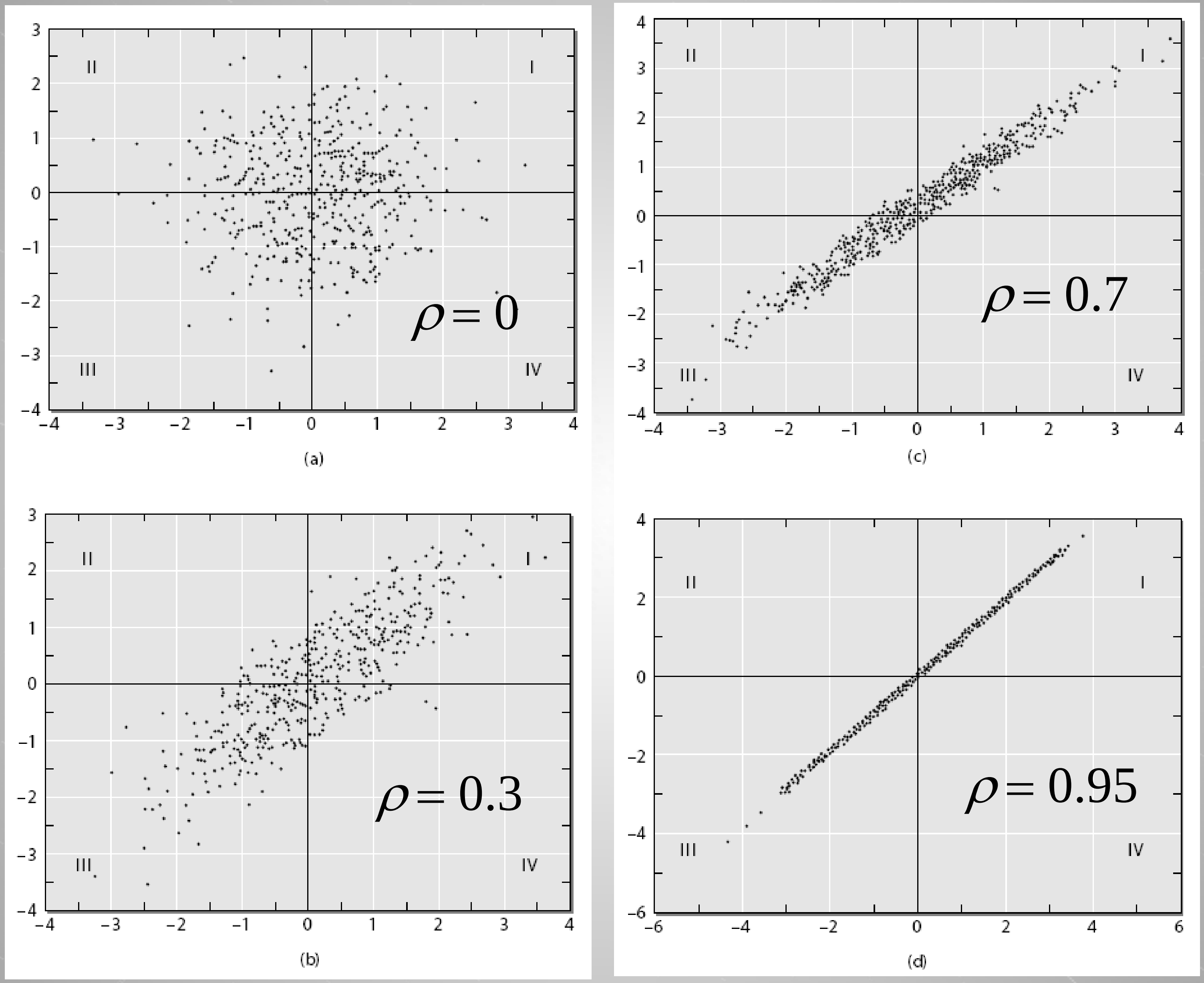
$E(XY) = (0)(1)(.45)+(0)(2)(.15)+(1)(1)(.05)+(1)(2)(.35)=.75$

covariance

$$
\begin{aligned}
cov(X,Y) &= E(XY) - EX\ EY \\
&= .75 - (.40)(1.50) \\
&= .75 - .60 \\
&= .15
\end{aligned}
$$

18. 상관계수

- *Correlation coefficient* $\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$
- $-1 \le \rho \le 1$
- positive correlation if $\rho > 0$
- negative correlation if $\rho < 0$
- no correlation if $\rho = 0$


II
I
III
IV
ρ = 0
(a)
II
I
III
IV
ρ = 0.7
(c)
II
I
III
IV
ρ = 0.3
(b)
II
I
III
IV
ρ = 0.95
(d)

<상관계수 계산 방법: 예시>

2.44

	Y = 1	Y = 2	
X = 0	.45	.15	.60
X = 1	.05	.35	.40
	.50	.50	

$EX = .40$

$EX^2 = 0^2(.60) + 1^2(.40) = .40$

$var(X) = E(X^2) - (EX)^2$
$= .40 - (.40)^2$
$= .24$

$cov(X,Y) = .15$

correlation

$EY = 1.50$

$EY^2 = 1^2(.50) + 2^2(.50)$
$= .50 + 2.0$
$= 2.50$

$var(Y) = E(Y^2) - (EY)^2$
$= 2.50 - (1.50)^2$
$= .25$

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)\ var(Y)}}$$

$\rho(X,Y) = .61$

- **Independence?**

⇒ Zero Covariance & Correlation

- Independent random variables have zero covariance and, therefore, zero correlation.

- The converse is **not true**.

<과제> (교과서 연습문제 풀이)

0.3
0.6
0.7
0.16
0.23

※ 참고: 필요한 data는 WILEY 교과서 홈페이지에 있음

- http://principlesofeconometrics.com/poe3/poe3.htm