SALUTON DENOVE!

확률 및 통계학

동국대학교

2025년 2학기

문 동 휘

월 12:00 ~ 13:30

수 11:00 ~ 12:30

뭔 차이인지
단 하나도
모르겠음

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|$$

$m(X)$ = average value of the data set

n = number of data values

x_i = data values in the set

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

Definition 5.6.1 Definition and p.d.f. A random variable X has the *normal distribution with mean μ and variance σ^2* ($-\infty < \mu < \infty$ and $\sigma > 0$) if X has a continuous distribution with the following p.d.f.:

$$f(x|\mu,\sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \text{for } -\infty < x < \infty. \tag{5.6.1}$$

Definition 5.6.2 Standard Normal Distribution. The normal distribution with mean 0 and variance 1 is called the *standard normal distribution.* The p.d.f. of the standard normal distribution is usually denoted by the symbol ϕ, and the c.d.f. is denoted by the symbol Φ. Thus,

$$\phi(x) = f(x|0, 1) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}x^2\right) \quad \text{for } -\infty < x < \infty \tag{5.6.6}$$

Theorem 8.2.3 Let X have the standard normal distribution. Then the random variable $Y = X^2$ has the χ^2 distribution with one degree of freedom.

Corollary 8.2.1 If the random variables $X_1, \ldots, X_m$ are i.i.d. with the standard normal distribution, then the sum of squares $X_1^2 + \cdots + X_m^2$ has the χ^2 distribution with m degrees of freedom. ∎

Distribution of X_i	Sample size n	Mean μ	Statistic	$1-\alpha$ confidence interval
$X_i \sim \mathcal{N}(\mu, \sigma)$	any	known or unknown	$\frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$	$\left[\frac{s^2(n-1)}{\chi^2_2}, \frac{s^2(n-1)}{\chi^2_1}\right]$

Formula for the Confidence Interval for a Variance

$$\frac{(n-1)s^2}{\chi^2_{\text{right}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{left}}}$$

$$\text{d.f.} = n - 1$$

Formula for the Confidence Interval for a Standard Deviation

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\text{right}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\text{left}}}}$$

$$\text{d.f.} = n - 1$$

책 2장 연습문제 (pg 23 ~pg 25)

- 질문 2) 2번에 측정 단위: 측정단위 = 1 = 정수들, 측정단위 = 0.1 = 소수점 첫째 자리까지 측정하였다, 측정단위 = 0.01 = 소수점 둘째 자리까지 측정하였다
- 즉, 도수분포표 급하한, 급상한도 측정단위에 맞게 표기하라 (급하한이 23 이고 측정단위가 0.01이면 급하한을 23이 아니라 23.00으로 표기)
- 질문 3, 4)도수분포표는 정답이 없어요. 계급의 수와 계급간격은 수업시간에 배운 공식에 따라도 되나 그렇지 않아도 됩니다. 둘 다 맞아요. 자료를 제일 잘 표현하고 분석가의 목적에 맞게 계급의 수와 계급간격을 정하면 됩니다. 같은 자료도 다른 도수분포표 그리기 가능함.
- 수학이 아닌 통계학인거 명심하기 (통계학은 수학이 아닙니다.) 공식은 절대적이지 않습니다. 무시는 하면 안되지만 결국 판단은 분석가가 하는겁니다.

이변량 자료

Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

A **scatter plot** is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y.

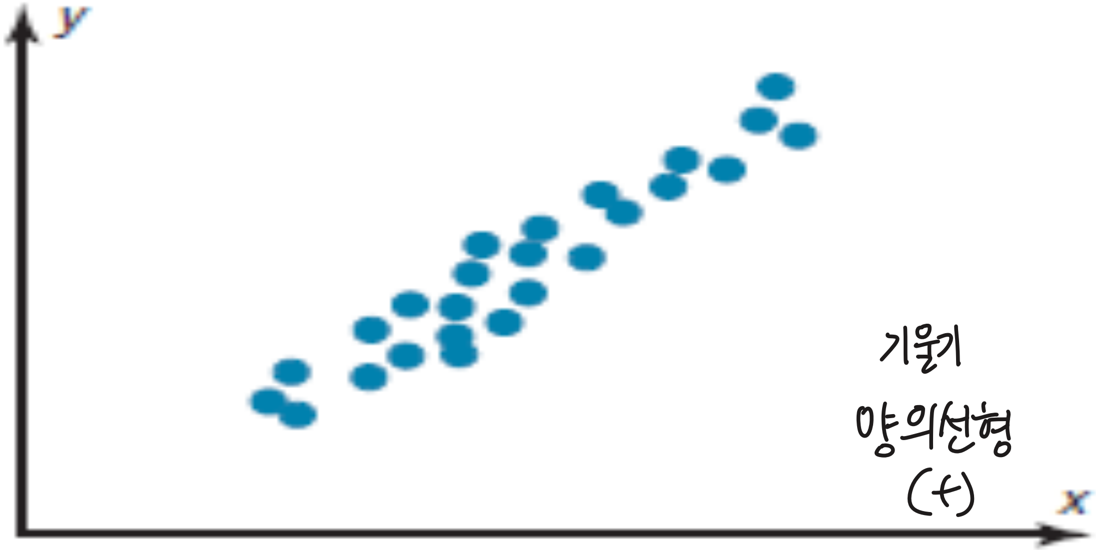
Procedure Table

Drawing a Scatter Plot

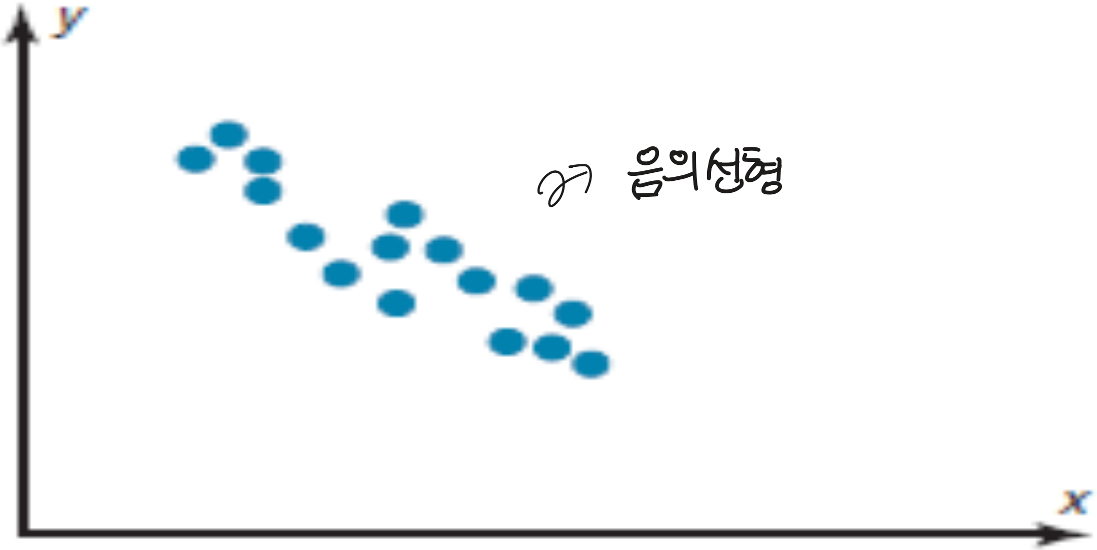
Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph.

Step 3 Determine the type of relationship (if any) that exists for the variables.



(a) Positive linear relationship



(b) Negative linear relationship

(c) Curvilinear relationship

(d) No relationship

측정하는지 상관X

Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

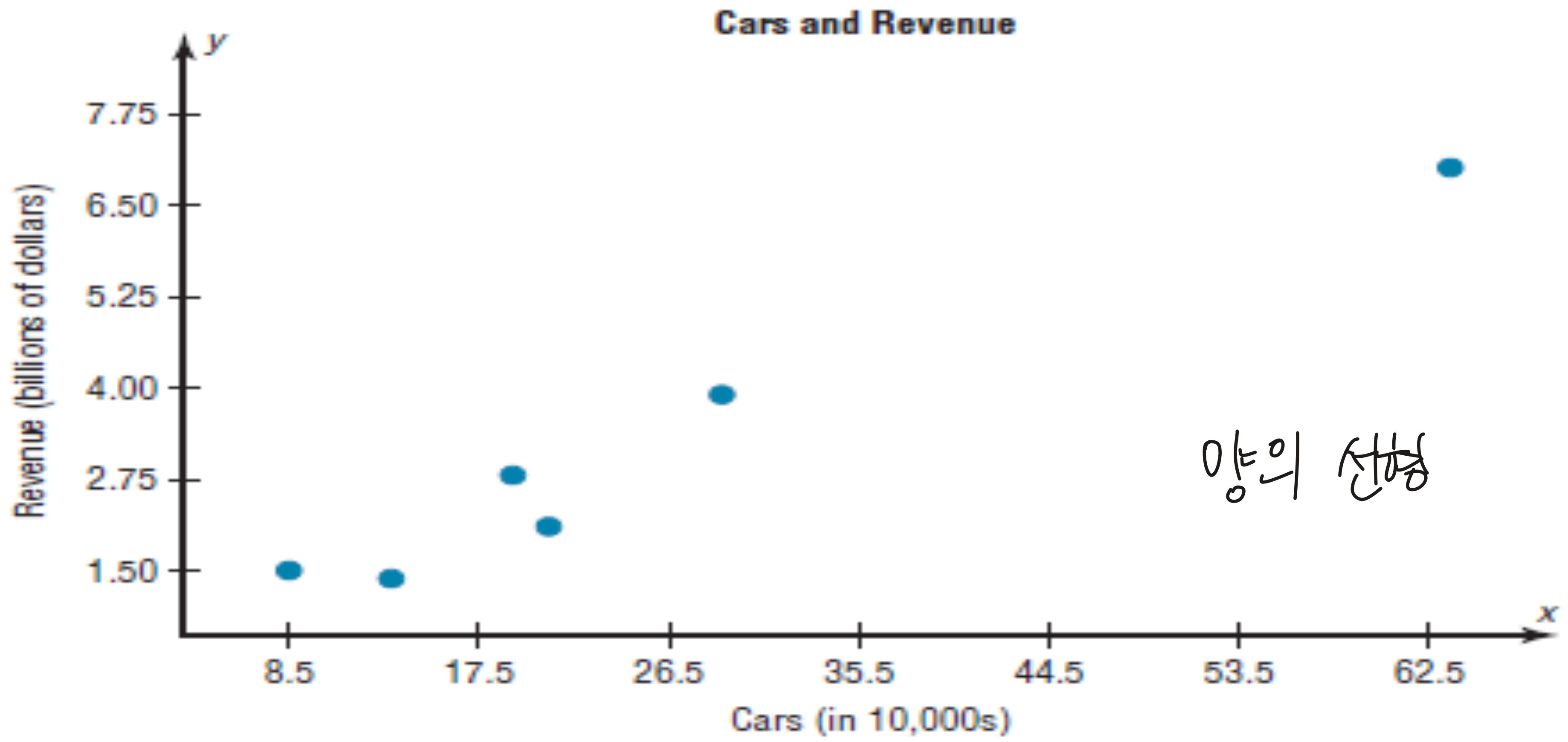
Company	Cars (in ten thousands) x	Revenue (in billions) y
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

SOLUTION

Step 1 Draw and label the *x* and *y* axes.

Step 2 Plot each point on the graph, as shown in Figure 10–2.

FIGURE 10–2 Scatter Plot for Example 10–1



Step 3 Determine the type of relationship (if any) that exists.

In this example, it looks as if a positive linear relationship exists between the number of cars that an agency owns and the total revenue that is made by the company.

Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. The data are shown here.

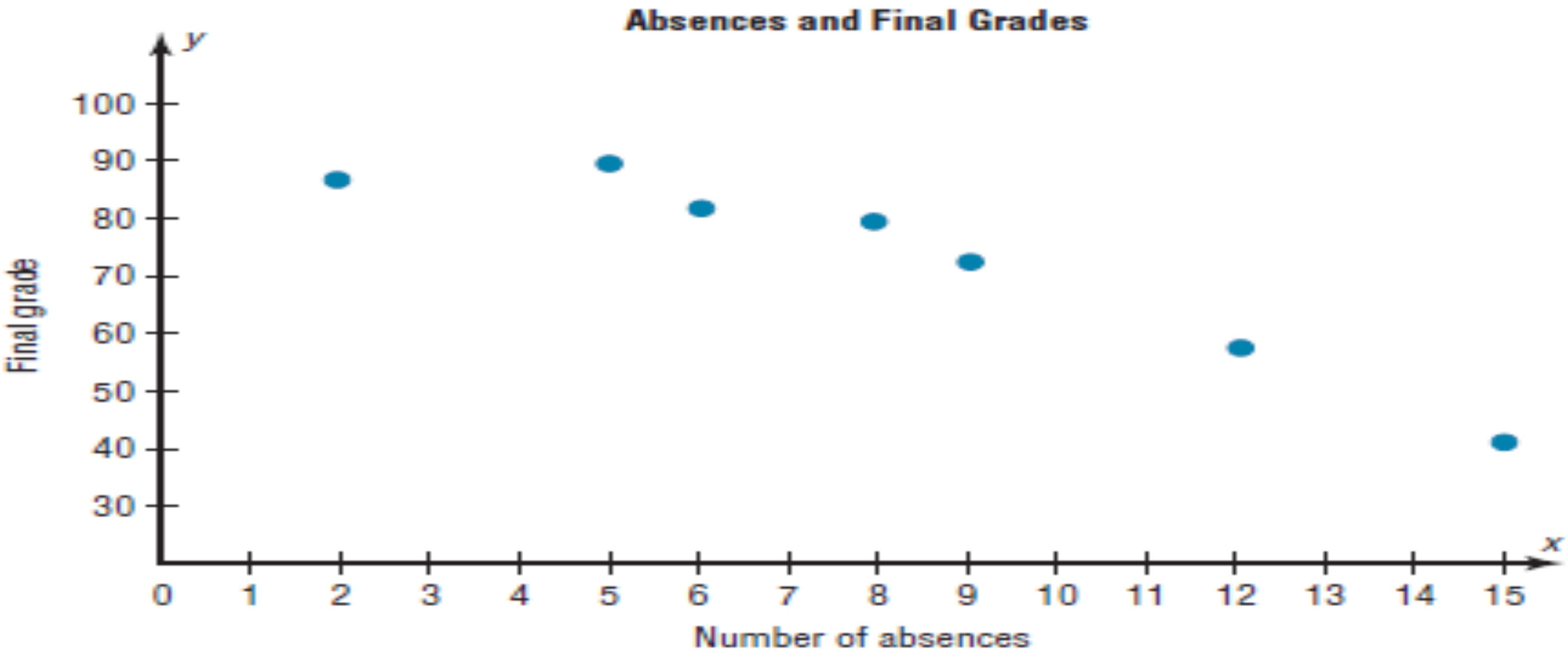
Student	Number of absences *x*	Final grade *y* (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

SOLUTION

Step 1 Draw and label the *x* and *y* axes.

Step 2 Plot each point on the graph, as shown in Figure 10–3.

FIGURE 10–3 Scatter Plot for Example 10–2

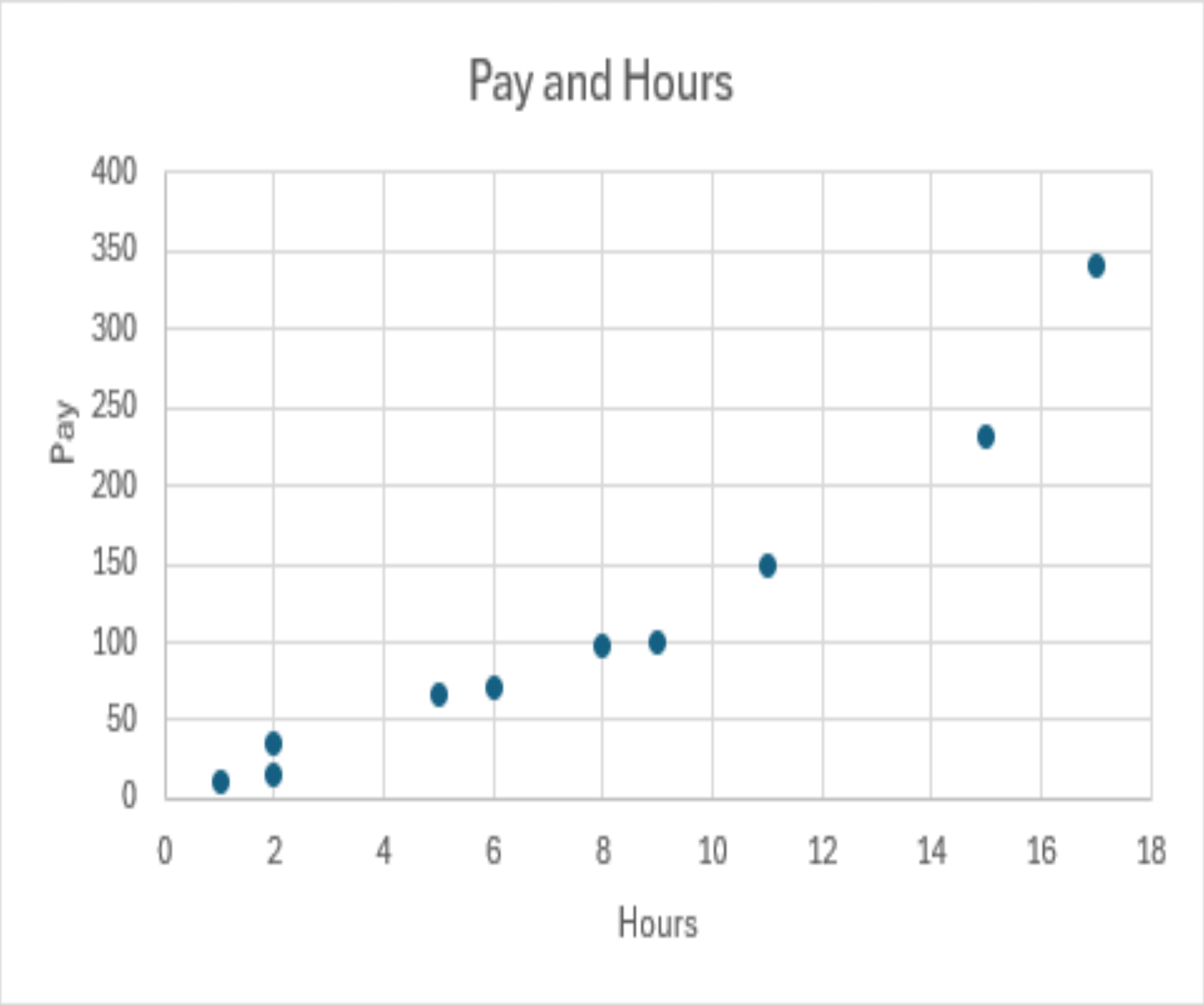


음의선형

Step 3 Determine the type of relationship (if any) that exists.

In this example, it looks as if a negative linear relationship exists between the number of student absences and the final grade of the students.

hours	pay
1	10
2	35
6	70
8	98
2	15
5	66
9	100
11	150
15	231
17	341



A researcher wishes to see if there is a relationship between the ages of the wealthiest people in the world and their net worth. A random sample of 10 persons was selected from the *Forbes* list of the 400 richest people for a recent year. The data are shown. Draw a scatter plot for the data.

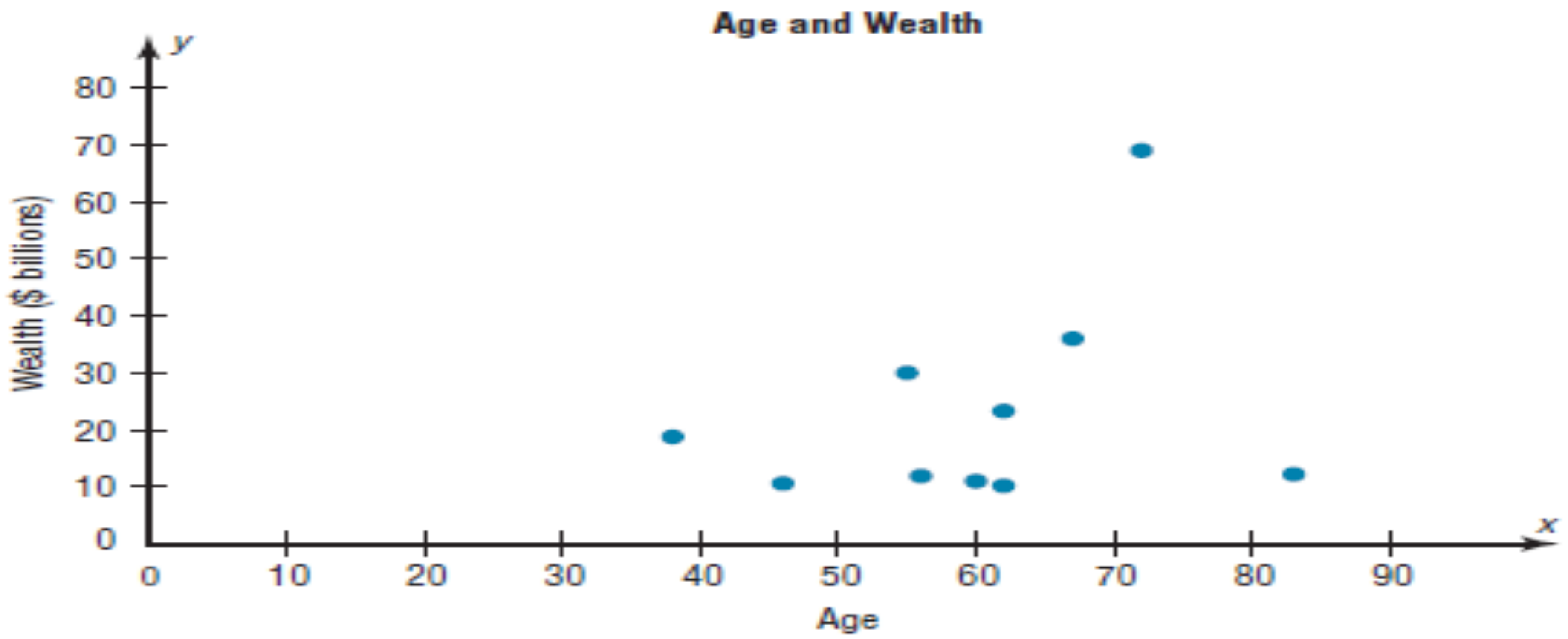
Person	Age x	Net worth y (in billions of dollars)
A	60	11
B	72	69
C	56	11.9
D	55	30
E	83	12.2
F	67	36
G	38	18.7
H	62	10.2
I	62	23.3
J	46	10.6

No relationship

SOLUTION

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure 10–4.



Step 3 Determine the type of relationship (if any) that exists.

In this example, there is no type of a strong linear or curvilinear relationship between a person's age and his or her net worth.

Correlation Coefficient Statisticians use a measure called the *correlation coefficient* to determine the strength of the linear relationship between two variables. There are several types of correlation coefficients.

선형상관계수

: 선형적으로 얼마큼 관계 가지는지

→ 모집단 상관계수

> The **population correlation coefficient** denoted by the Greek letter ρ is the correlation computed by using all possible pairs of data values (x, y) taken from a population.

: 모집단 전체 데이터로 계산한 것

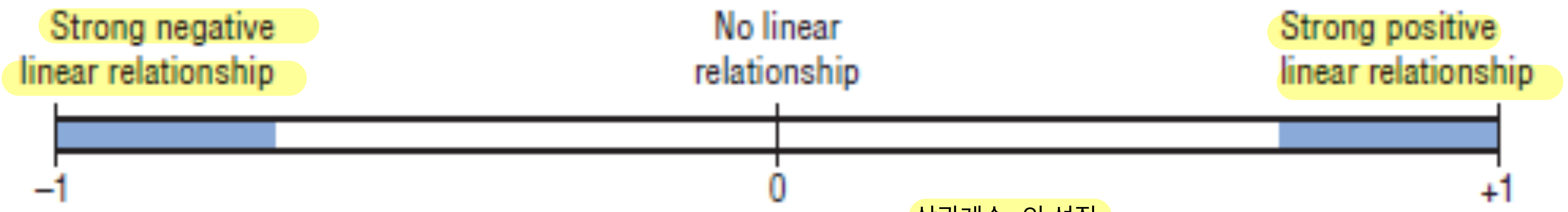
현실에선 사용 X — 모집단 구하지 X

> The **linear correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two quantitative variables. The symbol for the sample correlation coefficient is r. — 표본상관계수 : 실제 많이 사용

: 표본(샘플) 데이터로 계산한 값

The linear correlation coefficient explained in this section is called the **Pearson product moment correlation coefficient (PPMC)**, named after statistician Karl Pearson, who pioneered the research in this area.

↳ 상관계수



상관계수 r의 성질

상관계수는 단위가 없는 값이다.

값의 범위는 항상 -1에서 +1 사이에 있다.

x와 y를 서로 바꿔도 r 값은 변하지 않는다.

x나 y를 다른 단위(스케일)로 바꿔도 r 값은 변하지 않는다.

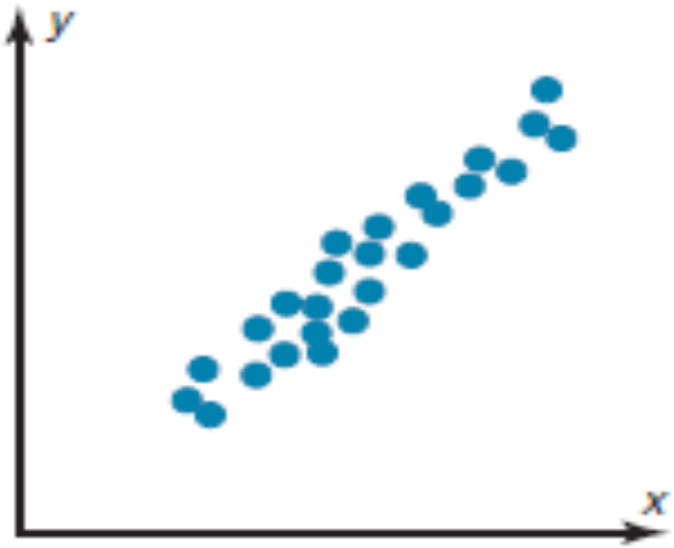
r 값은 **이상치(극단적인 값)**에 민감해서, 데이터에 이상치가 있으면 크게 변할 수 있다.

Properties of the Linear Correlation Coefficient

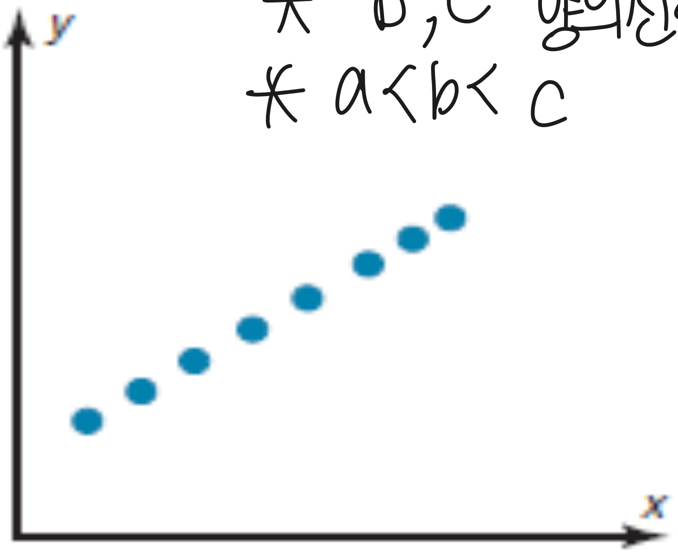
1. The correlation coefficient is a unitless measure.
2. The value of r will always be between -1 and $+1$ inclusively. That is, $-1 \le r \le 1$.
3. If the values of x and y are interchanged, the value of r will be unchanged.
4. If the values of x and/or y are converted to a different scale, the value of r will be unchanged.
5. The value of r is sensitive to outliers and can change dramatically if they are present in the data.

* b,c 양의선형
* a<b<c

(a) $r = 0.50$ no relationship



(b) $r = 0.90$

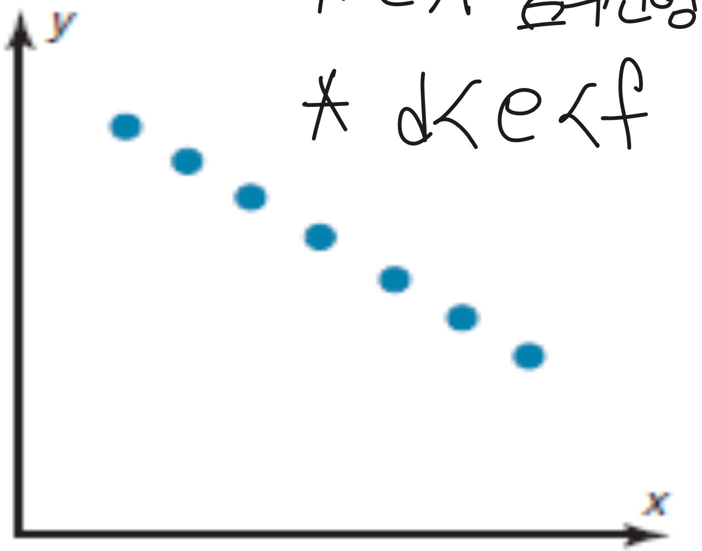


(c) $r = 1.00$

* e,f 음의선형
* d<e<f

(d) $r = -0.50$ no relationship

(e) $r = -0.90$



(f) $r = -1.00$

Assumptions for the Correlation Coefficient

1. The sample is a random sample. : 표본은 랜덤이여야함
2. The data pairs fall approximately on a straight line and are measured at the interval or ratio level. :
3. The variables have a bivariate normal distribution. (This means that given any specific value of *x*, the *y* values are normally distributed; and given any specific value of *y*, the *x* values are normally distributed.)

표본은 무작위 표본(random sample)이어야 한다.
→ 데이터를 임의로 뽑아야지 편향되지 않는다.

데이터 쌍은 직선 관계(straight line)를 따라야 하고, 구간척도(interval)나 비율척도(ratio)로 측정되어야 한다.
→ 즉, 두 변수 사이 관계가 대체로 직선 형태여야 하고, 데이터는 숫자로 의미 있게 측정된 값이어야 한다.

변수들은 이변량 정규분포(bivariate normal distribution)를 따라야 한다.
→ 즉,주어진 x 값에 대해서 y 값들이 정규분포를 하고,
주어진 y 값에 대해서 x 값들이 정규분포를 해야 한다.

* 수업 때 도수분포표로 계산함

책에 있는걸로 풀기

Formula for the Linear Correlation Coefficient r

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

where n is the number of data pairs.

Procedure Table

Finding the Value of the Linear Correlation Coefficient

Step 1 Make a table as shown.

x	y	xy	x^2	y^2

Step 2 Place the values of x in the x column and the values of y in the y column. Multiply each x value by the corresponding y value, and place the products in the xy column.

Square each x value and place the squares in the x^2 column.

Square each y value and place the squares in the y^2 column.

Find the sum of each column.

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

Step 3 Substitute in the formula and find the value for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{S_{xy}}{S_x S_y}$$

where n is the number of data pairs.

EXAMPLE 10–4 Car Rental Companies

Compute the linear correlation coefficient for the data in Example 10–1.

SOLUTION

Step 1 Make a table as shown here.

Company	Cars x (in ten thousands)	Revenue y (in billions)	xy	x^2	y^2
A	63.0	\$7.0			
B	29.0	3.9			
C	20.8	2.1			
D	19.1	2.8			
E	13.4	1.4			
F	8.5	1.5			

Step 2 Find the values of xy, x^2, and y^2, and place these values in the corresponding columns of the table.

Company	Cars x (in 10,000s)	Revenue y (in billions of dollars)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Step 3 Substitute in the formula and solve for r.

선형관계 보여줌.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The linear correlation coefficient suggests a strong positive linear relationship between the number of cars a rental agency has and its annual revenue. That is, the more cars a rental agency has, the more annual revenue the company will have.

$\hat{G}$*is revido!*