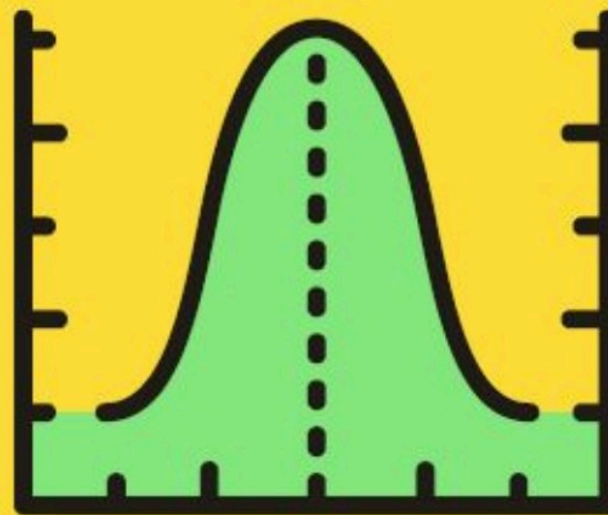


누구나 파이썬 통계분석



CHAPTER 03

2차원 데이터 정리

CONTENTS	3.1 두 데이터 사이의 관계를 나타내는 지표
	3.2 2차원 데이터의 시각화
	3.3 앤스컴의 예

3.0 준비

In [1]:

```
import numpy as np
import pandas as pd

%precision 3
pd.set_option( ' display.precision ' , 3)
```

In [2]:

```
df = pd.read_csv( ' ../data/ch2_scores_em.csv ' ,
                  index_col= ' student number ' )
```

In [3]:

```
en_scores = np.array(df[ ' english ' ])[:10]
ma_scores = np.array(df[ ' mathematics ' ])[:10]

scores_df = pd.DataFrame({ ' english ' :en_scores,
                           ' mathematics ' :ma_scores},
                           index=pd.Index([ ' A ' , ' B ' , ' C ' , ' D ' , ' E ' ,
                                             ' F ' , ' G ' , ' H ' , ' I ' , ' J ' ],
                                             name= ' student ' ))

scores_df
```

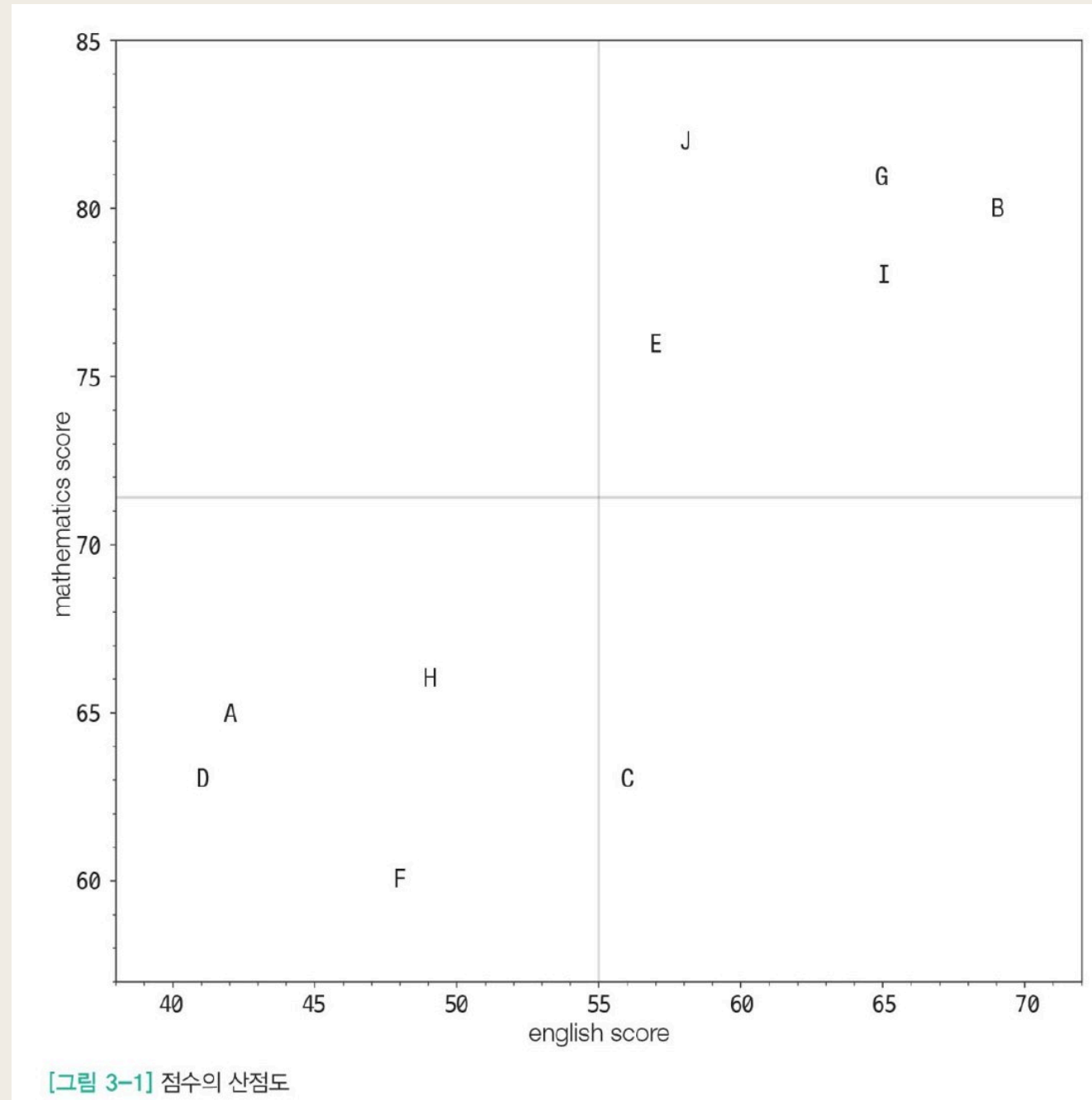
3.13.1 두 데이터 사이의 관계를 나타내는 지표두 데이터 사이의 관계를 나타내는 지표

영어 점수가 높은 학생일수록 수학 점수가 높은 경향이 있다면 영어 점수와 수학 점수는 양의 상관 관계

영어 점수가 높은 학생일수록 수학 점수가 낮은 경향이 있다면 영어 점수와 수학 점수는 음의 상관 관계

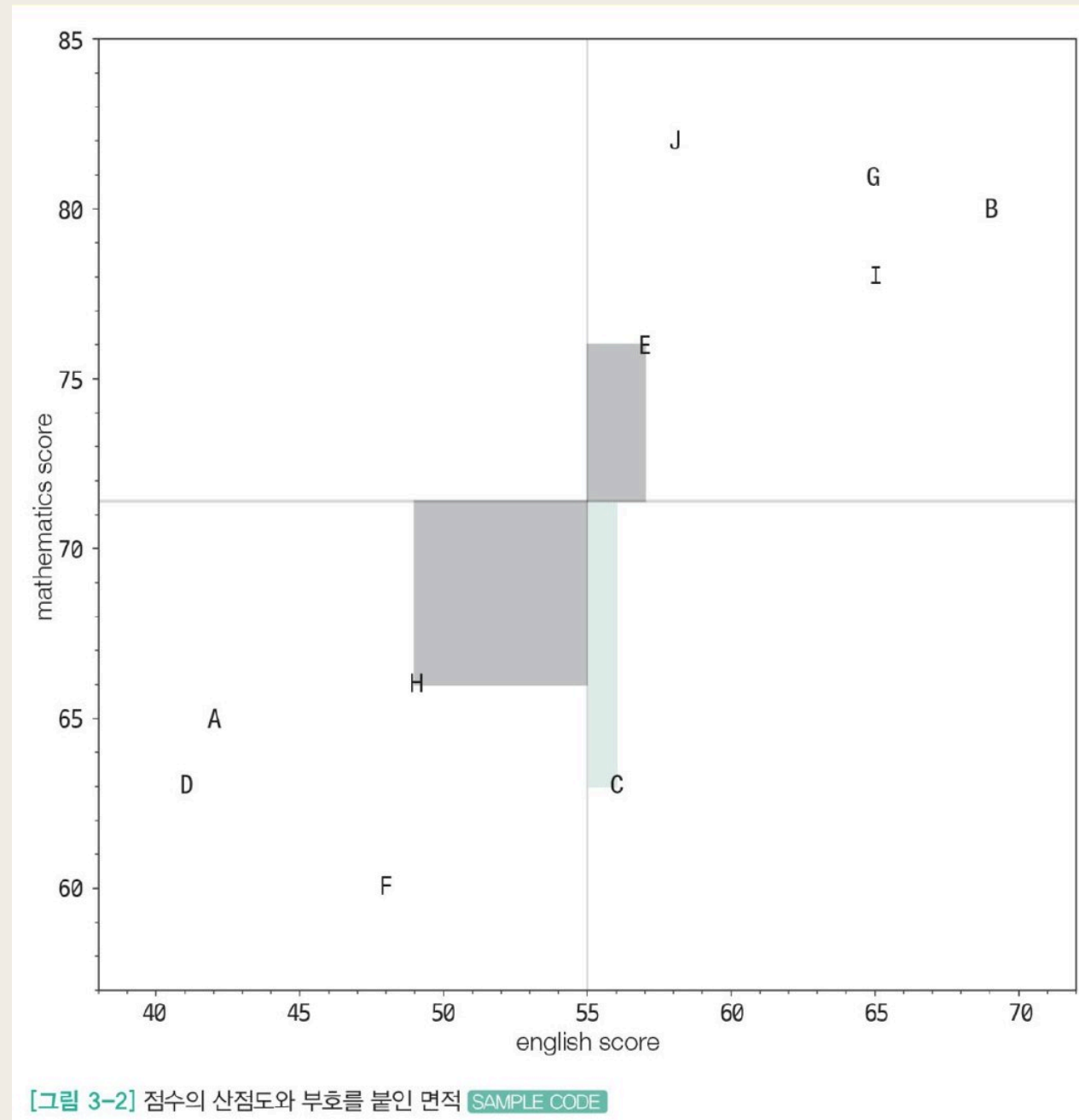
영어 점수가 수학 점수에 직접적으로 영향을 미치지 않을 때, 영어 점수와 수학 점수는 무상관

3.1.1 공분산



- 중간값의 가로선과 세로선은 수학과 영어 평균 점수
- 영어 점수와 수학 점수는 양의 상관 관계

3.1.1 공분산



- 직사각형의 가로길이는 영어 점수의 편차, 세로는 수학 점수의 편차
- 공분산은 면적, 음의 면적도 가능(음의 상관)

3.1.1 공분산

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

$$= \frac{1}{n} \{ (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}) \}$$

In [4] :

```
summary_df = scores_df.copy()
summary_df[ 'english_deviation' ] = \
    summary_df[ 'english' ] - summary_df[ 'english' ].mean()
summary_df[ 'mathematics_deviation' ] = \
    summary_df[ 'mathematics' ] - summary_df[ 'mathematics' ].mean()
summary_df[ 'product of deviations' ] = \
    summary_df[ 'english_deviation' ] * summary_df[ 'mathematics_deviation' ]
summary_df
```

In [5] :

```
summary_df[ 'product of deviations' ].mean()
```

3.1.1 공분산

- NumPy의 cov 함수 반환값은 공분산 행렬(분산공분산 행렬)

In [6] :

```
cov_mat = np.cov(en_scores, ma_scores, ddof=0)
cov_mat
```

Out [6] :

```
array([[86.  , 62.8 ],
       [62.8 , 68.44]])
```

- 1행 2열, 2행 1열 성분이 영어 수학의 공분산

In [7] :

```
cov_mat[0, 1], cov_mat[1, 0]
```

3.1.2 상관계수

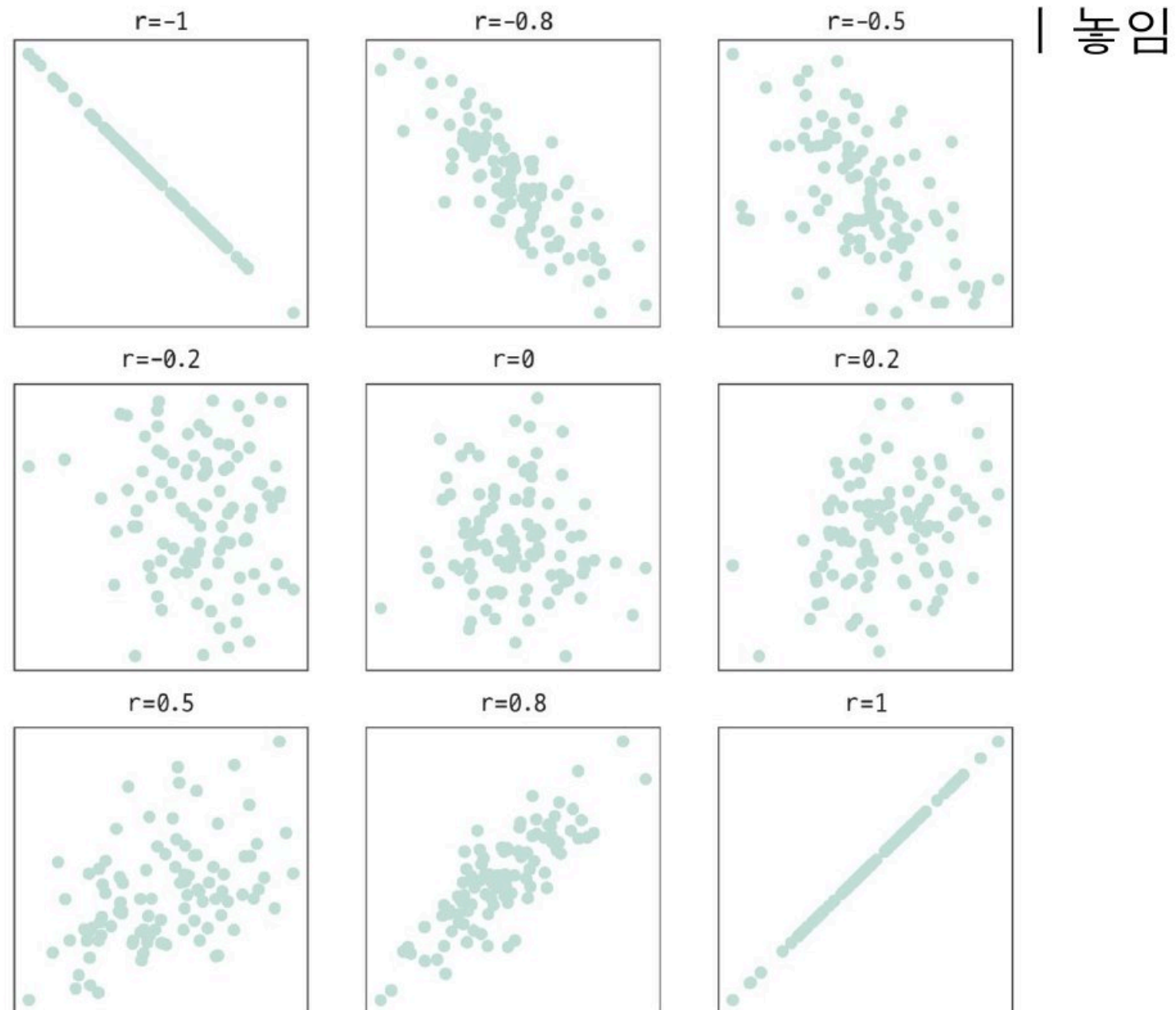
- 공분산의 단위는 직감적으로 이해하기 어려우므로, 단위에 의존하지 않는 상관을 나타내는 지표
 - 시험 점수간의 공분산 (점수 $\times$ 점수), 키와 점수 (cm $\times$ 점수)
- 상관계수는 공분산을 각 데이터의 표준편차로 나누어 단위에 의존하지 않음

$$\begin{aligned} r_{xy} &= \frac{S_{xy}}{S_x S_y} \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right) \end{aligned}$$

- 양의 상관은 1에 가까워지고, 음의 상관은 -1에 가까워지고, 무상관은 0

3.1.2 상관계수

- 양의 상관은 1에 가까워지고, 음의 상관은 -1에 가까워지고, 무상관은 0
- 상관계수가



[그림 3-3] 상관계수

3.1.2 상관계수

- 수식대로 계산하는 영어 점수와 수학 점수의 상관계수

In [10]:

```
np.cov(en_scores, ma_scores, ddof=0)[0, 1] /\n    (np.std(en_scores) * np.std(ma_scores))
```

- NumPy의 `corrcoef` 함수(상관행렬의 [0,1] [1,0] 성분)

In [11]:

```
np.corrcoef(en_scores, ma_scores)
```

- DataFrame의 `corr` 메서드

In [12]:

```
scores_df.corr()
```

3.2.1 산점도

In [13]:

```
import matplotlib.pyplot as plt
```

```
%matplotlib inline
```

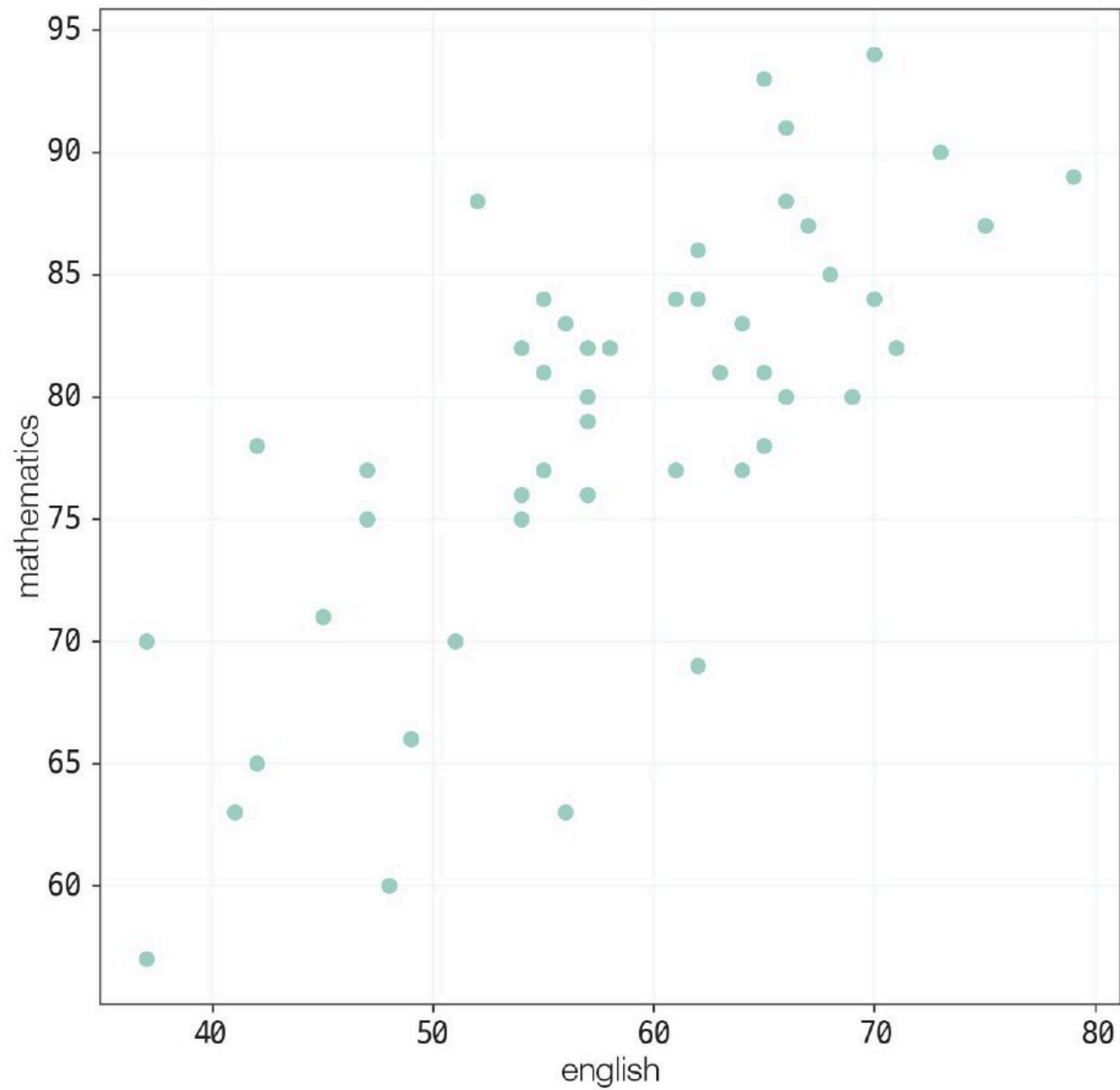
In [14]:

```
english_scores = np.array(df[ ' english ' ])  
math_scores = np.array(df[ ' mathematics ' ])
```

```
fig = plt.figure(figsize=(8, 8))  
ax = fig.add_subplot(111)  
# 산점도  
ax.scatter(english_scores, math_scores)  
ax.set_xlabel( ' english ' )  
ax.set_ylabel( ' mathematics ' )
```

```
plt.show()
```

3.2.1 산점도



[그림 3-4] 산점도

3.2.2 회귀직선

$$y = \beta_0 + \beta_1 x$$

In [15] :

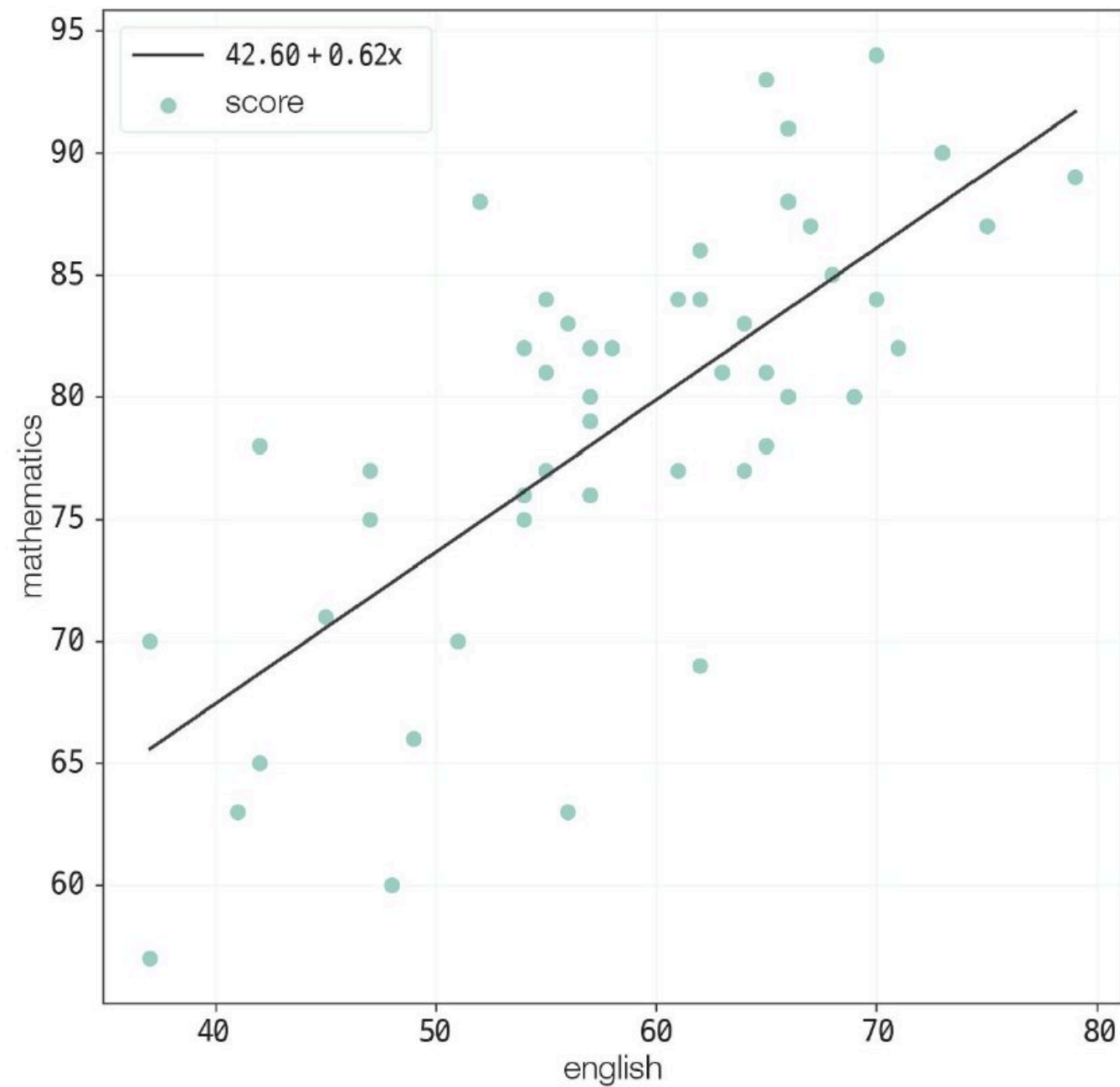
```
# 계수  $\beta_0$ 과  $\beta_1$ 을 구한다
poly_fit = np.polyfit(english_scores, math_scores, 1)
#  $\beta_0 + \beta_1 x$ 를 반환하는 함수를 작성
poly_1d = np.poly1d(poly_fit)
# 직선을 그리기 위해 x좌표를 생성
xs = np.linspace(english_scores.min(), english_scores.max())
# xs에 대응하는 y좌표를 구한다
ys = poly_1d(xs)

fig = plt.figure(figsize=(8, 8))
ax = fig.add_subplot(111)
ax.plot(xs, ys, color='gray',
        label=f'{poly_fit[1]:.2f}+{poly_fit[0]:.2f}x')
ax.scatter(english_scores, math_scores, label='score')
ax.set_xlabel('english')
ax.set_ylabel('mathematics')
# 범례 표시
ax.legend(loc='upper left')

plt.show()
```

3.2.2 회귀직선

$$y = \beta_0 + \beta_1 x$$



[그림 3-5] 산점도와 회귀직선

3.2.3 히트맵

- 히스토그램의 2차원 버전으로 색을 이용해 표현하는 그래프
- 영어 점수 35점부터 80점, 수학 점수 55점부터 95점까지 5점 간격

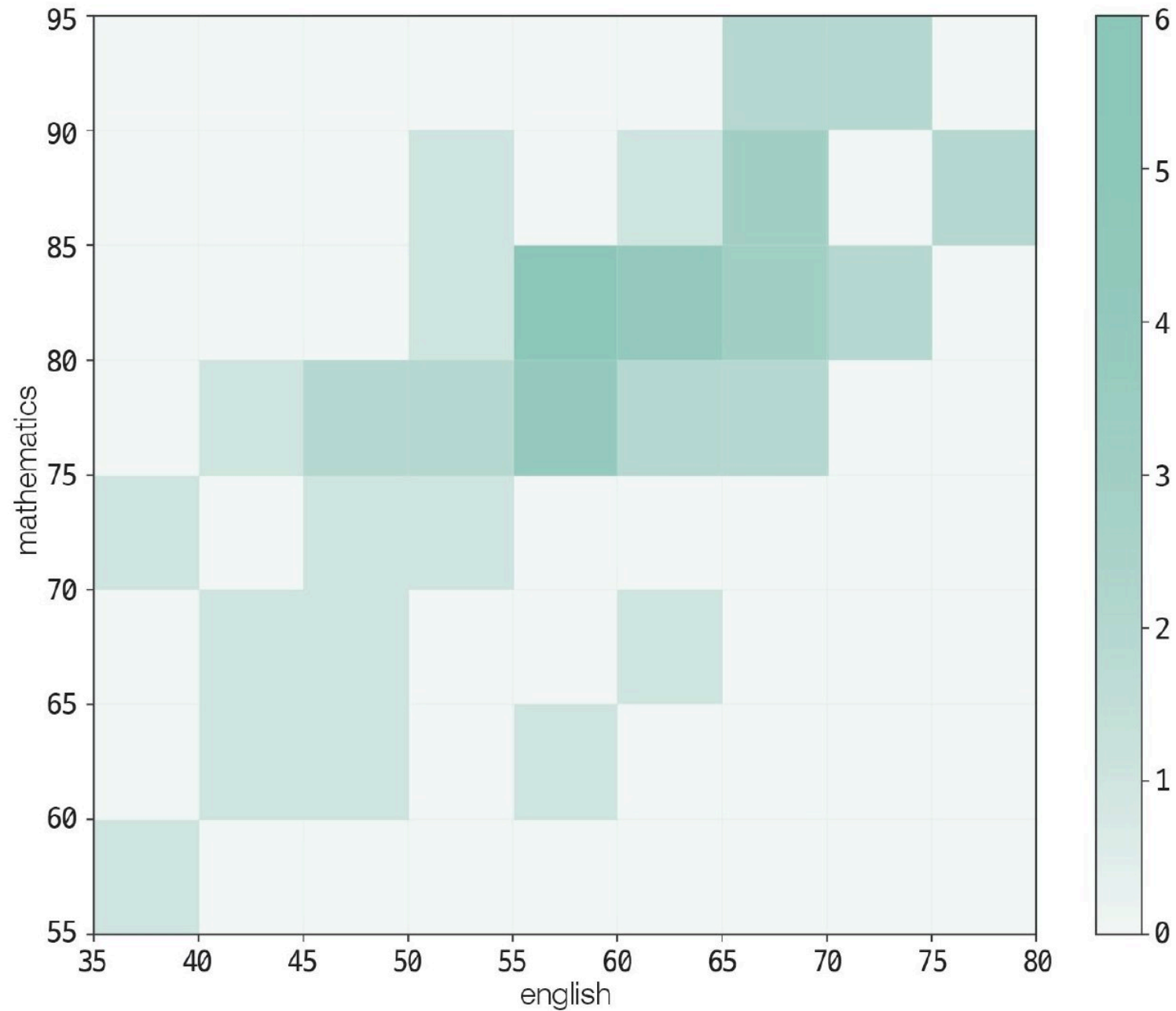
In [16]:

```
fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111)

c = ax.hist2d(english_scores, math_scores,
               bins=[9, 8], range=[(35, 80), (55, 95)])
ax.set_xlabel('english')
ax.set_ylabel('mathematics')
ax.set_xticks(c[1])
ax.set_yticks(c[2])
# 컬러 바 표시
fig.colorbar(c[3], ax=ax)
plt.show()
```

3.2.3 히트맵

- 색이 진한 영역일수록 많은 학생이 분포



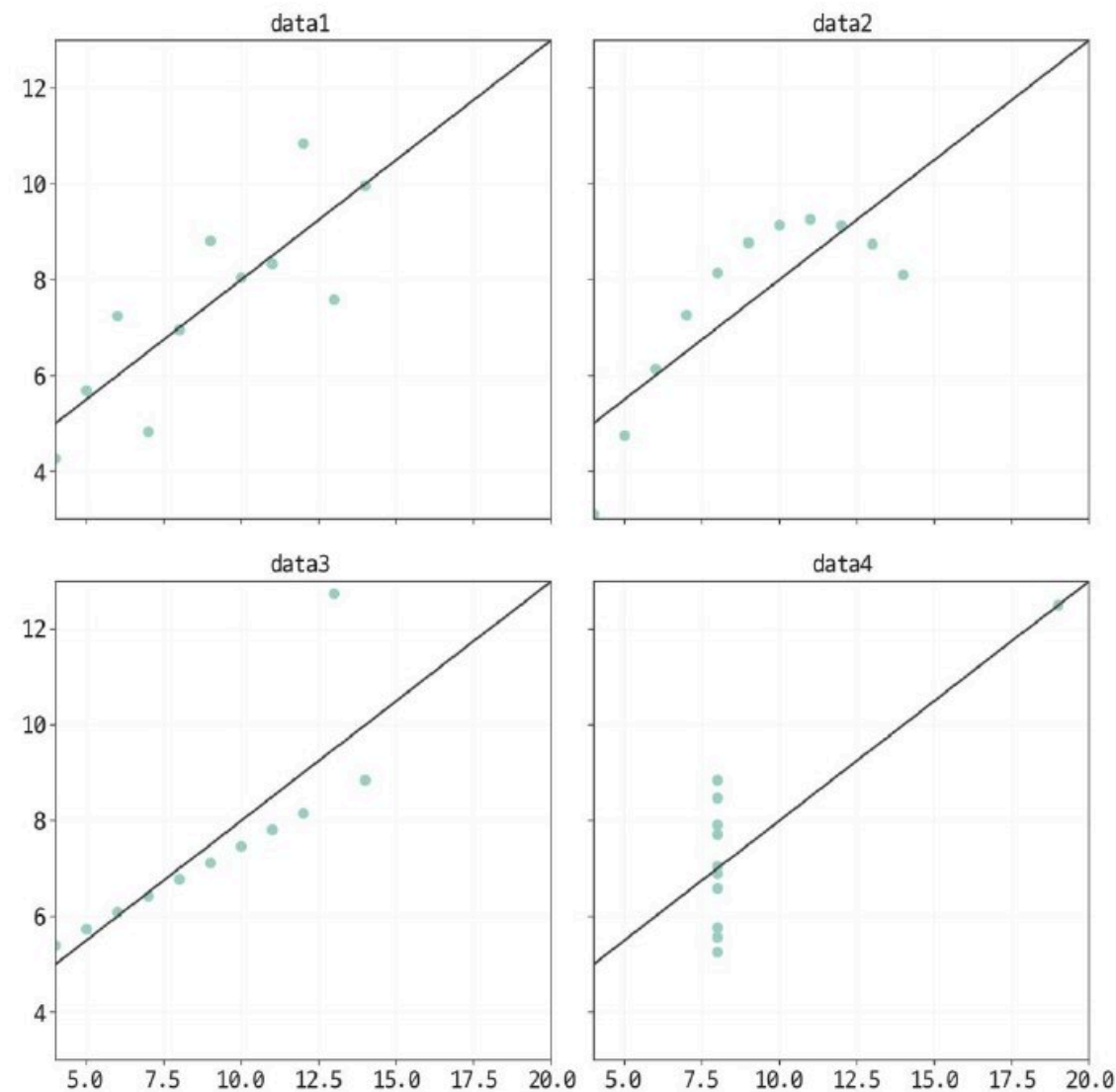
[그림 3-6] 히트맵

동일한 지표를 가지고 있지만 그림으로 표현하면 전혀 다른 데이터

	data1	data2	data3	data4
X_mean	9.00	9.00	9.00	9.00
X_variance	10.00	10.00	10.00	10.00
Y_mean	7.50	7.50	7.50	7.50
Y_variance	3.75	3.75	3.75	3.75
X&Y_correlation	0.82	0.82	0.82	0.82
X&Y_regression line	3.00+0.50x	3.00+0.50x	3.00+0.50x	3.00+0.50x

동일한 지표를 가지고 있지만 그림으로 표현하면 전혀 다른 데이터

	data1	data2	data3	data4
X_mean	9.00	9.00	9.00	9.00
X_variance	10.00	10.00	10.00	10.00
Y_mean	7.50	7.50	7.50	7.50
Y_variance	3.75	3.75	3.75	3.75
X&Y_correlation	0.82	0.82	0.82	0.82
X&Y_regression line	$3.00+0.50x$	$3.00+0.50x$	$3.00+0.50x$	$3.00+0.50x$



[그림 3-7] 앤스컴의 예

확인 문제 — 공분산·상관계수와 산점도

시험에 이런 형태로 나올 수 있는 문제입니다. 문제를 띄우면 답 장으로 넘기기 전에 먼저 손으로 풀어 보세요.

1 공분산이 음수라는 것은 무슨 뜻인가?

2 상관계수로 나올 수 없는 값은?

3 키의 단위를 cm 에서 m 로 바꾸면?

4 앤스컴의 예가 보여 주는 것은?

공분산이 음수라는 것은 무슨 뜻인가?

하루 스마트폰 사용 시간과 시험 점수의 공분산이 -12 로 나왔습니다.

- ① 두 변수는 아무 관계가 없다
- ② 한쪽이 크면 다른 쪽은 작은 경향이 있다
- ③ 관계가 매우 약하다
- ④ 계산이 틀렸다 — 공분산은 음수가 될 수 없다

공분산이 음수라는 것은 무슨 뜻인가?

2번 — 한쪽이 클수록 다른 쪽은 작은 경향

공분산의 부호는 방향만 말해 줍니다. 음수면 산점도가 오른쪽 아래로 향합니다. -12 라는 크기로는 관계가 강한지 약한지 판단할 수 없습니다 — 단위에 따라 값이 바뀌기 때문입니다.

상관계수로 나올 수 없는 값은?

- ① -0.95
- ② 0
- ③ 0.3
- ④ 1.2

상관계수로 나올 수 없는 값은?

4번 — 1.2

상관계수는 공분산을 두 표준편차의 곱으로 나눈 값이라 언제나 -1 이상 1 이하입니다. 1.2 가 나왔다면 계산 어딘가가 틀린 것입니다.

키의 단위를 cm 에서 m 로 바꾸면?

키(cm)와 몸무게(kg)의 공분산과 상관계수를 구했습니다. 키를 m 로 바꿔 다시 구합니다.

- ① 공분산·상관계수 모두 그대로
- ② 공분산은 $1/100$ 이 되고 상관계수는 그대로
- ③ 공분산은 그대로, 상관계수는 $1/100$
- ④ 둘 다 $1/100$

키의 단위를 cm 에서 m 로 바꾸면?

2번 — 공분산만 1/100, 상관계수는 그대로

키의 편차가 모두 1/100 이 되므로 편차의 곱인 공분산도 1/100 이 됩니다. 상관계수는 분모의 키 표준편차도 함께 1/100 이 되어 약분되므로 그대로입니다. 상관계수가 단위와 무관하다는 것이 이 문제의 요점입니다.

앤스컴의 예가 보여 주는 것은?

네 데이터의 평균·분산·상관계수가 모두 같았습니다.

- ① 지표가 같으면 데이터도 같다
- ② 상관계수가 같으면 산점도 모양도 같다
- ③ 지표가 같아도 그림으로 그려 보면 전혀 다를 수 있다
- ④ 데이터가 많을수록 지표가 정확하다

앤스컴의 예가 보여 주는 것은?

3번 — 지표만 보지 말고 그림으로 확인해야 한다

앤스컴의 네 데이터는 평균·분산·상관계수가 거의 같지만 산점도는 직선, 곡선, 이상값 하나가 끌어당긴 모양 등으로 제각각입니다. 요약 지표는 데이터의 모양을 다 담지 못합니다.

Q&A