

예시

# 기초 데이터 과학

교재 제작 참조 템플릿의 견본 원고. 주제와 수치, 인물, 기관은 모두 예시이며 실제 교육 과정과 관계없다.

|        |                           |
|--------|---------------------------|
| 지은이    | 편집부 (예시 저자)               |
| 판 · 날짜 | 제1판 v1.0.0 · 2026년 10월 3일 |
| 대상     | 대학 교양 · 1학기 15주 (예시)      |



[예시] 기초 데이터 과학



# [예시] 기초 데이터 과학

측정에서 회귀까지 — 교재 템플릿 견본

편집부 (예시 저자) 지음

예시출판 (가상)

|      |                                 |
|------|---------------------------------|
| 서명   | [예시] 기초 데이터 과학                  |
| 판·쇄  | 제1판 제1쇄 (v1.0.0) · 2026년 10월 3일 |
| 지은이  | 편집부 (예시 저자)                     |
| 펴낸곳  | 예시출판 (가상) · 주소와 전화번호는 예시 자리     |
| ISBN | 979-11-0000-000-0 93000 (예시 번호) |
| 정가   | 00,000원                         |

이 문서는 교재 제작용 템플릿의 견본이다. ISBN, 출판사, 저자, 수치 자료는 모두 예시이며 실제와 관계없다. 사진과 도면은 직접 만든 합성 자료다.

이 책의 저작권은 저자에게 있습니다. 저작권법에 따라 보호받는 저작물이므로 무단 전재와 복제를 금합니다. 잘못된 책은 구입하신 곳에서 바꿔 드립니다.

# 머리말

이 책은 교재 템플릿이 실제로 어떻게 보이는지 확인하기 위한 견본이다. 주제로 고른 ‘기초 데이터 과학’은 학습 요소를 골고루 보여 주기 좋아서 택한 예시일 뿐이다.

교재는 일반 도서와 달리 독자가 처음부터 끝까지 한 번 읽고 덮는 책이 아니다. 학습자는 장 표제지에서 목표를 확인하고, 본문과 예제를 오가며 개념을 익히고, 확인 문제로 이해를 점검한 뒤, 장 끝의 요약과 연습 문제로 돌아온다. 그래서 교재의 조판은 같은 종류의 정보가 언제나 같은 모양으로, 같은 자리에 나오는 것이 가장 중요하다.

이 견본은 「도서·교재 출판 디자인 시스템 & 빌드 표준」을 그대로 따른다. 판형은 46배판(188 × 257mm), 본문은 Pretendard 9.5/16pt, 코드는 JetBrains Mono 8/12.5pt다. 좌우 대칭 여백, 홀수 쪽 장 시작, 앞붙이 로마 숫자 쪽번호, 자동 차례와 표·그림 차례, 찾아보기, PDF 책갈피를 빌드가 만든다. 장·절·표·그림·식·예제·문제 번호는 원고에 쓰지 않는다.

원고를 새로 쓸 때는 다음 쪽의 「이 책의 구성과 활용법」과 저장소의 README.md를 함께 보기 바란다. 모든 학습 요소의 HTML 견본이 그곳에 있다.

2026년 10월  
편집부 (예시)



# 이 책의 구성과 활용법

모든 장은 같은 순서로 짜여 있다. 학습 요소는 언제나 같은 모양과 같은 자리로 나오므로, 한 번 익혀 두면 어느 장에서든 필요한 곳을 바로 찾을 수 있다.

|                                      |                                |                               |                                     |                               |
|--------------------------------------|--------------------------------|-------------------------------|-------------------------------------|-------------------------------|
| ① 장 표제지<br>학습 목표 · 핵심 용어 ·<br>생각해 보기 | ② 본문<br>절 · 예제와 풀이 · 팁 ·<br>심화 | ③ 확인 문제<br>절을 마칠 때마다 바로<br>점검 | ④ 장 마무리<br>요약 · 핵심 용어 정리 ·<br>연습 문제 | ⑤ 정답 및 해설<br>부록 A · 문제마다 쪽 안내 |
|--------------------------------------|--------------------------------|-------------------------------|-------------------------------------|-------------------------------|

## 학습 요소 범례

| 학습 목표            |
|------------------|
| 01 평균을 계산할 수 있다. |

### 학습 목표 .objectives

장 표제지에 3~5개를 ‘~할 수 있다’ 문장으로 씀. 번호는 장 끝 ‘요약’의 번호와 1:1로 대응한다.

| 핵심 용어     |
|-----------|
| ■ 평균 ■ 분산 |

### 핵심 용어 .keyterms

장에서 새로 정의하는 용어. 본문에서 처음 나올 때 굵게 쓰고, 장 끝 ‘핵심 용어 정리’와 찾아보기에 다시 실린다.

| 생각해 보기                     |
|----------------------------|
| 반 평균이 같으면 두 반의 성적도<br>같을까? |

### 생각해 보기 .warmup

장을 시작하기 전에 떠올려 볼 질문. 정답이 없으며, 장을 마친 뒤 다시 답해 보도록 본문에서 한 번 되짚는다.

| 예제 2-1 표본 평균 |
|--------------|
| 풀이<br>  답 14 |

### 예제와 풀이 .example

문제 → 단계별 풀이 → 답 순서. 번호 ‘예제 n-m’은 빌드가 매기며, 본문에서 xref로 참조한다.

| 팁              |
|----------------|
| 계산기보다 표로 정리하자. |

### 팁 .callout.tip

시간을 아끼는 요령이나 기억법. 강조색 세로줄로 표시한다.

### 주의

$n$ 과  $n - 1$ 을 혼동하지 말 것.

### 주의 `.callout.warn`

자주 틀리는 곳, 오해하기 쉬운 개념. 콜아웃은 연달아 두 개를 넘기지 않는다.

### 심화

왜  $n - 1$ 로 나눌까?

### 심화 `.sidebar.deep`

시험 범위를 넘는 깊은 내용. 건너뛰어도 다음 절을 읽는 데 지장이 없다.

### 읽을거리

통계라는 말의 유래

### 읽을거리 `.sidebar.read`

역사, 사례, 직업 이야기처럼 흥미를 돋우는 글. 본문 흐름 밖에 있다.

### 확인 문제

**2-1** 중앙값은? 정답 쪽 안내

### 확인 문제 `.check`

절을 마칠 때 바로 푸는 1~3문항. 문제 끝의 ‘정답  $n$ 쪽’은 부록 A의 해설로 이어지는 링크다.

### 요약

**01** 평균은 자료의 합을 개수로 나눈 값이다.

### 요약 · 핵심 용어 정리 `.summary .termlist`

장 끝에서 학습 목표 번호 순서대로 요약하고, 핵심 용어를 정의와 함께 다시 정리한다.

### 응용

### 연습 문제 `.exercises .level`

기본 · 응용 · 심화 세 단계. 강조색 막대 수(1~3개)가 난이도다. 모든 문제의 정답과 해설은 부록 A에 있다.

## 번호와 참조 읽는 법

| 표기         | 뜻                        | 예             |
|------------|--------------------------|---------------|
| 2.3, 2.3.1 | 2장 3절, 2장 3절 1항          | 2.3 분산과 표준편차  |
| 표·그림 n-m   | n장의 m번째 표·그림 (도면도 같은 방식) | 표 2-1, 그림 3-1 |
| 식 (n-m)    | n장의 m번째 번호 붙은 식          | 식 (2-2)       |
| 확인 문제 n-m  | n장의 m번째 확인 문제            | 확인 문제 1-2     |
| 연습 문제 n-m  | n장의 m번째 연습 문제            | 연습 문제 3-4     |



# 차례

|                                 |           |
|---------------------------------|-----------|
| 표·그림 차례 .....                   | xiii      |
| <b>1 데이터와 측정 .....</b>          | <b>1</b>  |
| 1.1 데이터란 무엇인가 .....             | 2         |
| 1.2 변수의 종류와 척도 .....            | 4         |
| 1.3 측정 결과를 표로 정리하기 .....        | 5         |
| <b>2 자료 요약하기: 기술 통계 .....</b>   | <b>9</b>  |
| 2.1 도수분포와 히스토그램 .....           | 10        |
| 2.2 중심 경향: 평균과 중앙값 .....        | 11        |
| 2.3 퍼짐: 분산과 표준편차 .....          | 12        |
| <b>3 두 변수의 관계: 상관과 회귀 .....</b> | <b>15</b> |
| 3.1 산점도와 상관 .....               | 16        |
| 3.2 단순 선형 회귀 .....              | 18        |
| 3.3 예측과 그 한계 .....              | 19        |
| <b>부록 A 정답 및 해설 .....</b>       | <b>23</b> |
| A.1 1장 데이터와 측정 .....            | 23        |
| A.2 2장 자료 요약하기 .....            | 24        |
| A.3 3장 두 변수의 관계 .....           | 24        |
| <b>용어집 .....</b>                | <b>27</b> |
| <b>참고 문헌 .....</b>              | <b>29</b> |
| <b>찾아보기 .....</b>               | <b>31</b> |



## 표·그림 차례

### 표

#### 1장

|       |                   |   |
|-------|-------------------|---|
| 표 1-1 | 변수의 네 가지 척도 ..... | 4 |
|-------|-------------------|---|

#### 2장

|       |                                        |    |
|-------|----------------------------------------|----|
| 표 2-1 | 6월 일 최고 기온의 도수분포 (가상 자료, n = 30) ..... | 10 |
|-------|----------------------------------------|----|

#### 3장

|       |                       |    |
|-------|-----------------------|----|
| 표 3-1 | 기온과 판매량 (가상 자료) ..... | 16 |
|-------|-----------------------|----|

### 그림

#### 1장

|        |                                          |   |
|--------|------------------------------------------|---|
| 그림 1-1 | 관측 장비가 설치된 들판 (3/4 폭 108.5mm, 3:2) ..... | 2 |
|--------|------------------------------------------|---|

#### 2장

|        |                      |    |
|--------|----------------------|----|
| 그림 2-1 | 일 최고 기온의 히스토그램 ..... | 10 |
|--------|----------------------|----|

#### 3장

|        |                                   |    |
|--------|-----------------------------------|----|
| 그림 3-1 | 기온과 판매량의 산점도와 회귀 직선 (3/4 폭) ..... | 17 |
|--------|-----------------------------------|----|

### 도면

#### 1장

|        |                        |   |
|--------|------------------------|---|
| 도면 1-1 | 기상 관측 노장 배치도(가상) ..... | 3 |
|--------|------------------------|---|



# 01



## 데이터와 측정

데이터 과학의 출발점은 ‘무엇을, 어떤 단위로, 어떻게 잴는가’를 분명히 하는 일이다. 이 장에서는 관측 단위와 변수, 척도의 종류를 익히고 측정 결과를 표로 정리하는 방법을 배운다.

### 학습 목표

- 01** 관측 단위와 변수를 구별하여 설명할 수 있다.
- 02** 변수를 명목·순서·구간·비율 척도로 분류할 수 있다.
- 03** 단위를 변환하고 척도에 맞는 연산을 고를 수 있다.
- 04** 측정 결과를 변수 열과 관측 행으로 이루어진 표로 정리할 수 있다.

### 핵심 용어

■ 관측 단위 ■ 변수 ■ 척도 ■ 정형 자료 ■ 결측값

### 생각해 보기

“오늘은 어제보다 두 배 더웠다.” 어제 10°C, 오늘 20°C였다면 이 말은 옳을까? 장을 마친 뒤 다시 답해 보자.

## 1.1 데이터란 무엇인가

**데이터**는 관찰하거나 측정한 결과를 기록한 것이다. 기상 관측소가 매일 기록하는 기온, 강수량, 풍속이 대표적인 예다. 이 책은 가상의 관측소 자료를 예로 들어 설명한다.

### 1.1.1 관측 단위와 변수

측정의 대상이 되는 하나하나를 **관측 단위**라 하고, 관측 단위마다 값이 달라질 수 있는 특성을 **변수**라 한다. 관측소 자료에서는 ‘하루’가 관측 단위이고, 그날의 최고 기온과 강수량이 변수다.

**그림 1-1**은 관측 장비가 놓인 들판의 모습이다. 장비의 배치는 **도면 1-1**(3쪽)의 배치도로 정확히 나타낸다. 사진은 분위기와 맥락을, 도면은 위치와 치수를 전한다.

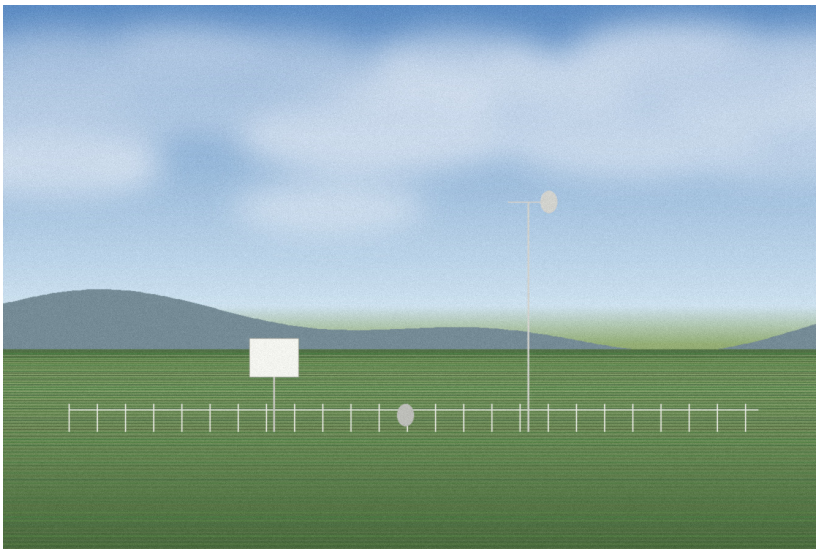


**그림 1-1** 관측 장비가 설치된 들판 (3/4 폭 108.5mm, 3:2) 합성 이미지(자리표시)

### 1.1.2 자료의 형태

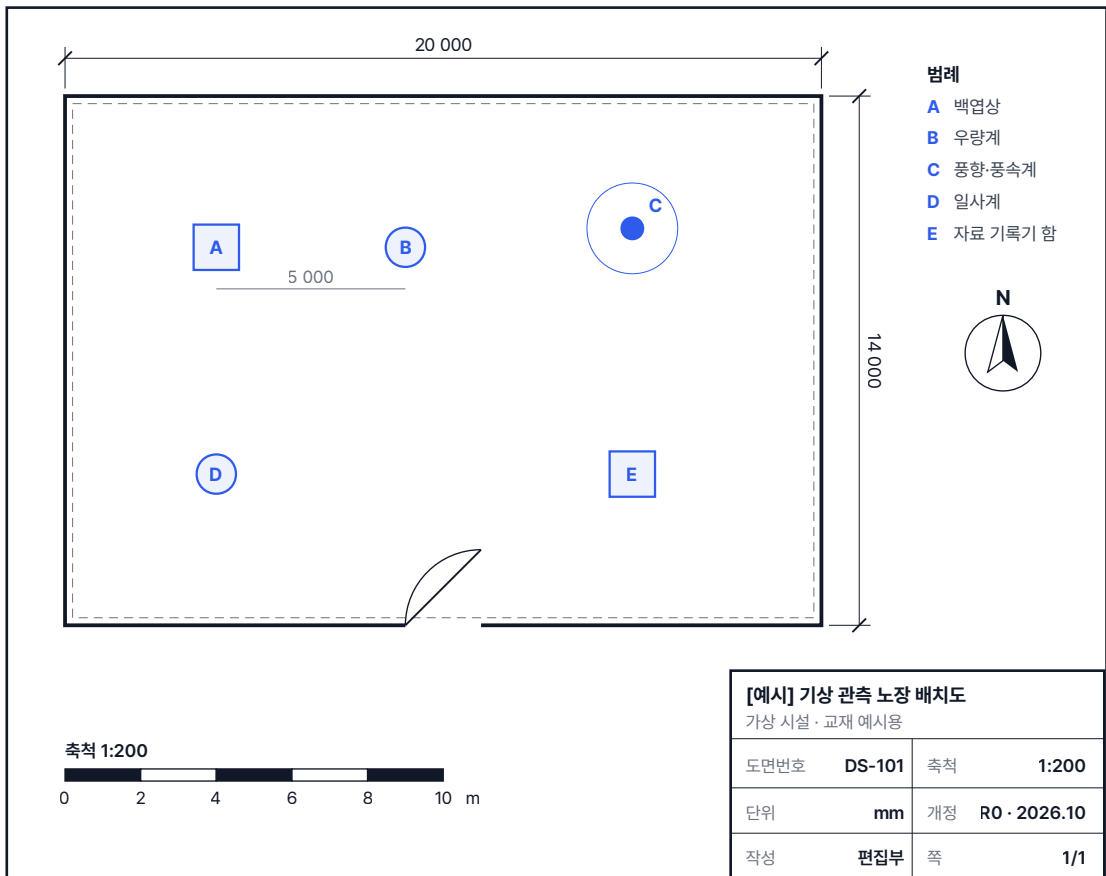
자료는 담기는 모양에 따라 두 가지로 나눌 수 있다.

#### 정형 자료

행과 열이 정해진 표에 담기는 자료를 **정형 자료**라 한다. 각 행은 관측 단위, 각 열은 변수다. 이 책에서 다루는 자료는 모두 정형 자료다.

## 비정형 자료

사진, 음성, 자유 서술처럼 정해진 표 구조가 없는 자료를 비정형 자료라 한다. 분석하려면 먼저 필요한 특성을 뽑아 정형 자료로 바꾼다.



**도면 1-1** 기상 관측 노장 배치도(가상). 축척 1:200 (원도 1:100), 단위 mm, 도면번호 DS-101

### 확인 문제

**1-1** 관측소가 30일 동안 매일 최고 기온과 강수량을 기록했다. 관측 단위와 변수를 각각 쓰시오.

정답 23쪽

**1-2** 설문지의 자유 서술 응답은 정형 자료인가, 비정형 자료인가? 정답 23쪽

## 1.2 변수의 종류와 척도

변수는 값의 성질에 따라 네 가지 **척도**로 나눈다(표 1-1). 척도에 따라 쓸 수 있는 연산과 요약 방법이 달라진다.

표 1-1 변수의 네 가지 척도

| 척도 | 값의 성질                | 가능한 연산    | 예              |
|----|----------------------|-----------|----------------|
| 명목 | 이름만 구별한다. 순서가 없다     | 같다·다르다    | 혈액형, 관측소 이름    |
| 순서 | 순서가 있지만 간격이 같지 않다    | + 크다·작다   | 학년, 만족도(상·중·하) |
| 구간 | 간격이 같지만 0이 '없음'이 아니다 | + 더하기·빼기  | 섭씨온도, 연도       |
| 비율 | 간격이 같고 0이 '없음'이다     | + 곱하기·나누기 | 강수량, 몸무게       |

주: 아래 척도일수록 위 척도의 연산을 모두 포함한다.

### 예제 1-1 변수의 척도 분류

관측소 자료의 변수 '관측소 코드', '구름 양(맑음·조금 흐림·흐림)', '최고 기온(°C)', '강수량(mm)'의 척도를 각각 쓰시오.

풀이

- 1 관측소 코드는 이름표일 뿐 크기나 순서가 없다 → 명목 척도.
- 2 구름 양은 순서가 있지만 '맑음'과 '조금 흐림'의 차이가 다른 단계 차이와 같다고 할 수 없다 → 순서 척도.
- 3 섭씨 0°C는 '온도 없음'이 아니다 → 구간 척도. 강수량 0mm는 '비가 없음'이다 → 비율 척도.

답 명목 · 순서 · 구간 · 비율

팁

척도가 헷갈리면 “0이 ‘없음’을 뜻하는가?”와 “두 배라고 말할 수 있는가?”를 차례로 물어보자. 둘 다 ‘예’면 비율 척도다.

### 1.2.1 단위 변환

구간 척도인 섭씨온도와 화씨온도는 식 (1-1)로 바꾼다. 두 단위의 0점이 다르므로 곱하기만으로는 바꿀 수 없다.

$$F = \frac{9}{5}C + 32 \quad (1-1)$$

### 예제 1-2 섭씨를 화씨로 바꾸기

관측소의 6월 평균 최고 기온은 26.4°C였다. 화씨로 바꾸시오.

풀이

- 1 식 (1-1)에  $C = 26.4$ 를 넣는다:  $9/5 \times 26.4 = 47.52$ .
- 2 32를 더한다:  $47.52 + 32 = 79.52$ .

**답 약 79.5°F**

### 읽을거리

#### 켈빈 온도는 비율 척도다

과학에서 쓰는 켈빈(K) 온도는 분자 운동이 멈추는 절대 영도를 0으로 둔다. 0K가 ‘열이 없음’을 뜻하므로 켈빈 온도는 비율 척도이고, “두 배 뜨겁다”는 말은 켈빈으로 잴 때만 뜻이 있다. 섭씨온도에 273.15를 더하면 켈빈 온도가 된다.

## 1.3 측정 결과를 표로 정리하기

정형 자료는 한 행에 관측 단위 하나, 한 열에 변수 하나를 둔다. 이 형식을 지키면 대부분의 분석 도구에서 바로 읽을 수 있다. 코드 1-1은 CSV 파일에서 기온 열을 읽어 결측값을 세는 예다.

**코드 1-1** CSV 파일에서 기온 열 읽기 (Python 표준 라이브러리)

```
1 import csv
2
3 temps, missing = [], 0
4 with open("station_june.csv", encoding="utf-8") as f:
5     for row in csv.DictReader(f):      # 한 행 = 하루(관측 단위)
6         value = row["max_temp"].strip() # 한 열 = 변수
7         if value == "":
8             missing += 1 # 빈칸은 결측값
9             continue
10        temps.append(float(value))
11
12 print(len(temps), "일 기록,", missing, "일 결측")
```

**주의**

결측값을 0으로 채우면 안 된다. 기온 0°C는 실제로 측정된 값이므로, 빈칸을 0으로 바꾸면 평균이 크게 낮아진다. 결측은 결측대로 따로 센다.

**심화**

**측정 오차**

같은 대상을 여러 번 재도 값은 조금씩 다르다. 측정 장비의 한계에서 오는 오차를 측정 오차라 하며, 기록에는 장비의 분해능(예: 0.1°C)을 함께 적어 둔다. 분해능보다 더 자세한 자릿수로 결과를 보고하지 않는다.

**확인 문제**

**1-3** 코드 1-1에서 빈칸을 결측값으로 세는 줄은 무엇을 검사하는가? 정답 23쪽

## 요약

- 01** 관측 단위는 측정 대상 하나하나이고, 변수는 관측 단위마다 달라지는 특성이다.
- 02** 척도는 명목·순서·구간·비율의 네 가지이며, 아래 척도일수록 쓸 수 있는 연산이 많다.
- 03** 섭씨와 화씨처럼 0점이 다른 단위는 식 **(1-1)**처럼 곱하기와 더하기로 바꾼다.
- 04** 정형 자료는 한 행에 관측 단위, 한 열에 변수를 두며, 결측값은 0으로 채우지 않는다.

## 핵심 용어 정리

|                                    |                                 |
|------------------------------------|---------------------------------|
| <b>관측 단위</b><br>observational unit | 측정의 대상이 되는 하나하나. 표의 한 행.        |
| <b>변수</b><br>variable              | 관측 단위마다 값이 달라질 수 있는 특성. 표의 한 열. |
| <b>척도</b><br>scale of measurement  | 변수 값의 성질에 따른 분류. 명목·순서·구간·비율.   |
| <b>정형 자료</b><br>structured data    | 행과 열이 정해진 표에 담기는 자료.            |
| <b>결측값</b><br>missing value        | 기록되지 않은 값. 0과 구별하여 따로 센다.       |

## 연습 문제

■ ■ ■ **기본** 개념을 그대로 적용한다

**1-1** 다음 변수의 척도를 쓰시오. (1) 혈액형 (2) 학년 (3) 섭씨온도 (4) 몸무게   정답 23쪽

**1-2** 25°C를 화씨온도로 바꾸시오.   정답 23쪽

■ ■ ■ **응용** 두 개념을 함께 쓴다

**1-3** 도면 **1-1**에서 백엽상(A)과 우량계(B)의 중심 사이는 지면에서 25mm다. 실제 거리는 몇 m인가?  
정답 23쪽

**1-4** ‘생각해 보기’의 주장 “10°C에서 20°C가 되었으니 두 배 더워졌다”가 옳지 않은 까닭을 척도로 설명하고, 켈빈 온도로 비를 구하시오.   정답 23쪽

■■■ 심화 스스로 방법을 고른다

---

**1-5** 30일 기온 기록 중 2일이 결측이다. 평균 기온을 보고할 때 취할 수 있는 처리 방법 두 가지와 각각의 장단점을 쓰시오. 정답 23쪽

# 02



## 자료 요약하기: 기술 통계

숫자 서른 개를 그대로 늘어놓으면 특징이 보이지 않는다. 이 장에서는  
도수분포표와 히스토그램으로 자료의 모양을 보고, 평균·중앙값·분산·표준편차로  
중심과 퍼짐을 한두 개의 수로 요약한다.

### 학습 목표

- 01 도수분포표를 만들고 히스토그램으로 나타낼 수 있다.
- 02 평균과 중앙값을 계산하고 이상값의 영향을 비교할 수 있다.
- 03 표본 분산과 표준편차를 계산하고 그 뜻을 해석할 수 있다.
- 04 코드로 기술 통계량을 구하고 손 계산과 대조할 수 있다.

### 핵심 용어

■ 도수분포 ■ 히스토그램 ■ 평균 ■ 중앙값 ■ 분산 ■ 표준편차

### 생각해 보기

두 반의 수학 평균이 똑같이 70점이다. 그렇다면 두 반의 성적 분포도 같다고 말할 수 있을까? 어떤 정보가 더 있어야 할까?

## 2.1 도수분포와 히스토그램

값의 범위를 몇 개의 계급으로 나누고 각 계급에 속하는 값의 개수를 센 표를 **도수분포표**라 한다. 표 2-1은 가상 관측소의 6월 일 최고 기온 30일을 폭 2°C의 계급으로 정리한 것이다.

표 2-1 6월 일 최고 기온의 도수분포 (가상 자료, n = 30)

| 계급 (°C)     | 도수 (일) | 상대도수  | 누적 도수 |
|-------------|--------|-------|-------|
| 21 이상 23 미만 | 3      | 0.100 | 3     |
| 23 이상 25 미만 | 5      | 0.167 | 8     |
| 25 이상 27 미만 | 10     | 0.333 | 18    |
| 27 이상 29 미만 | 7      | 0.233 | 25    |
| 29 이상 31 미만 | 4      | 0.133 | 29    |
| 31 이상 33 미만 | 1      | 0.033 | 30    |
| 합계          | 30     | 1.000 | —     |

출처: 이 책을 위해 만든 가상 자료. 상대도수는 소수 셋째 자리에서 반올림했다.

도수분포표를 막대로 그린 것이 **히스토그램**이다(그림 2-1). 막대 사이에 틈을 두지 않는 까닭은 계급이 연속된 수의 구간이기 때문이다.

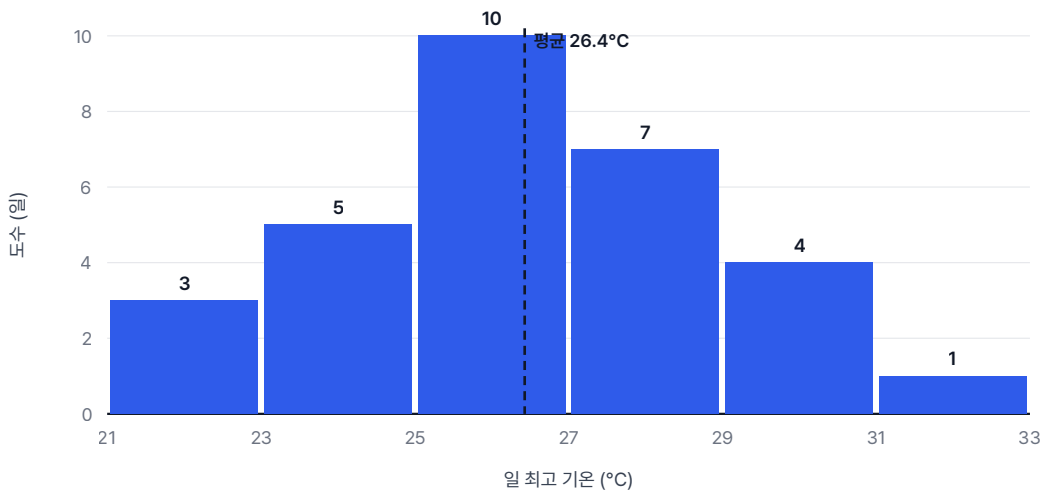


그림 2-1 일 최고 기온의 히스토그램. 점선은 평균

## 2.2 중심 경향: 평균과 중앙값

### 2.2.1 평균

평균은 모든 값을 더해 개수로 나눈 값이다.  $n$ 개의 값  $x_1, \dots, x_n$ 의 평균은 식 (2-1)과 같다.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n} \quad (2-1)$$

### 2.2.2 중앙값

값을 크기 순서로 늘어놓았을 때 한가운데 오는 값이 **중앙값**이다. 값의 개수가 짝수이면 가운데 두 값의 평균을 쓴다.

#### 예제 2-1 평균과 중앙값 구하기

다섯 번 잦 기온이 12, 15, 11, 14, 18(°C)이다. 평균과 중앙값을 구하시오.

풀이

- 1 합:  $12 + 15 + 11 + 14 + 18 = 70$ . 평균:  $70 \div 5 = 14$ .
- 2 크기 순서: 11, 12, **14**, 15, 18. 다섯 개 중 셋째 값이 중앙값이다.

답 평균 14°C, 중앙값 14°C

#### 이상값의 영향

다른 값들과 동떨어진 값을 이상값이라 한다. 위 예제의 18을 58로 잘못 적으면 평균은 22로 크게 뛰지만 중앙값은 14 그대로다. 중앙값은 이상값에 덜 민감하다.

#### 팁

평균과 중앙값이 크게 다르면 자료가 한쪽으로 치우쳤거나 이상값이 있다는 신호다. 둘을 함께 보고하는 습관을 들이자.

#### 확인 문제

**2-1** 자료 9, 3, 7, 5, 11의 중앙값을 구하시오. 정답 24쪽

**2-2** 평균과 중앙값 가운데 이상값에 덜 민감한 것은 무엇인가? 정답 24쪽

## 2.3 퍼짐: 분산과 표준편차

평균이 같아도 값이 평균 근처에 모여 있는지, 넓게 흩어져 있는지는 다를 수 있다. 각 값과 평균의 차이인 편차를 제곱해 더한 뒤  $n - 1$ 로 나눈 값을 표본 **분산**이라 한다(식 (2-2)). 분산의 제곱근이 **표준편차**다(식 (2-3)).

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

$$s = \sqrt{s^2} \quad (2-3)$$

### 예제 2-2 분산과 표준편차 구하기

예제 2-1의 자료 12, 15, 11, 14, 18의 표본 분산과 표준편차를 구하시오.

풀이

- 1 평균은 14이므로 편차는 -2, 1, -3, 0, 4다.
- 2 편차 제곱의 합:  $4 + 1 + 9 + 0 + 16 = 30$ .
- 3 표본 분산:  $30 \div (5 - 1) = 7.5$ . 표준편차:  $\sqrt{7.5} \approx 2.74$ .

**답 분산 7.5, 표준편차 약 2.74°C**

### 주의

계산기나 프로그램에는  $n$ 으로 나누는 ‘모분산’과  $n - 1$ 로 나누는 ‘표본 분산’이 따로 있다. 이 책의 분산은 모두 표본 분산이다.

### 심화

#### 왜 $n - 1$ 로 나눌까?

표본 평균은 표본 값들에 가장 가깝게 놓이므로, 표본 평균으로 잰 편차 제곱의 합은 참 평균으로 잰 것보다 조금 작아진다.  $n$  대신  $n - 1$ 로 나누면 이 치우침이 평균적으로 바로잡힌다. 이를 자유도로 설명하기도 한다.

**코드 2-1**은 Python 표준 라이브러리의 `statistics` 모듈로 6월 기온 30일의 기온 통계량을 구한다. `variance`와 `stdev`는 표본 분산과 표본 표준편차를 돌려준다.

**코드 2-1** 기술 통계량 계산과 실행 결과

```

1 import statistics as st
2
3 june = [21.4, 23.8, 25.1, 26.3, 24.7, 22.9, 27.5, 28.2, 26.8, 25.6,
4         24.1, 29.3, 30.4, 27.9, 26.2, 25.4, 23.3, 28.7, 31.2, 29.8,
5         27.1, 26.6, 24.9, 25.8, 28.4, 30.1, 27.4, 26.1, 22.6, 25.3]
6
7 for name, fn in [("평균", st.mean), ("중앙값", st.median),
8                 ("분산", st.variance), ("표준편차", st.stdev)]:
9     print(f"{name}: {fn(june):.2f}")
10 # 평균: 26.43   중앙값: 26.25   분산: 6.00   표준편차: 2.45

```

**확인 문제****2-3** 모든 값이 똑같은 자료의 분산은 얼마인가? 정답 24쪽**요약**

- 01** 도수분포표는 계급별 도수를 센 표이고, 히스토그램은 이를 틈 없는 막대로 그린 그림이다.
- 02** 평균은 합 ÷ 개수, 중앙값은 한가운데 값이다. 중앙값은 이상값에 덜 민감하다.
- 03** 표본 분산은 편차 제곱의 합을  $n - 1$ 로 나눈 값이고, 표준편차는 그 제곱근이다.
- 04** statistics 모듈의 결과를 손 계산과 대조하면 공식의 차이( $n$ 과  $n - 1$ )를 확인할 수 있다.

**핵심 용어 정리**

**도수분포표**                      계급마다 속하는 값의 개수(도수)를 정리한 표.  
frequency table

**히스토그램**                      도수분포를 틈 없는 막대로 나타낸 그림.  
histogram

**평균**                              모든 값의 합을 개수로 나눈 값. 식 (2-1).  
mean

**중앙값**                              크기 순서로 늘어놓았을 때 한가운데 값.  
median

## 분산

variance

편차 제곱의 평균적인 크기. 표본 분산은  $n - 1$ 로 나눈다.

## 표준편차

standard deviation

분산의 제곱근. 원자료와 단위가 같다.

## 연습 문제

■ ■ ■ 기본 개념을 그대로 적용한다

**2-1** 자료 4, 8, 6, 5, 7의 평균과 중앙값을 구하시오. 정답 24쪽

**2-2** 표 2-1에서 최고 기온이 27°C 이상 29°C 미만인 날의 상대도수를 구하시오. 정답 24쪽

■ ■ ■ 응용 두 개념을 함께 쓴다

**2-3** 자료 2, 4, 4, 4, 5, 5, 7, 9의 표본 분산과 표준편차를 소수 셋째 자리까지 구하시오. 정답 24쪽

**2-4** 다섯 가구의 월 소득(백만 원)이 10, 12, 11, 13, 54다. 평균과 중앙값을 구하고, ‘대표적인 소득’으로 어느 값이 알맞은지 쓰시오. 정답 24쪽

■ ■ ■ 심화 스스로 방법을 고른다

**2-5** 예제 2-2의 각 값에 3을 더한 뒤 2를 곱했다. 새 자료의 평균과 표준편차를 계산 없이 구하고 그 까닭을 설명하시오. 정답 24쪽

# 03



## 두 변수의 관계: 상관과 회귀

기온이 오르면 아이스크림이 더 팔릴까? 두 변수가 함께 움직이는 정도를 수로 나타내고, 한 변수로 다른 변수를 예측하는 직선을 구한다. 그리고 그 직선을 어디까지 믿을 수 있는지 따져 본다.

### 학습 목표

- 01 산점도를 그리고 두 변수 관계의 방향과 강도를 읽을 수 있다.
- 02 상관계수를 계산하고 상관과 인과를 구별할 수 있다.
- 03 최소제곱법으로 단순 회귀 직선을 구하고 예측에 쓸 수 있다.
- 04 결정계수와 외삽의 위험을 들어 예측의 한계를 설명할 수 있다.

### 핵심 용어

■ 산점도   ■ 상관계수   ■ 회귀 직선   ■ 결정계수   ■ 외삽

### 생각해 보기

아이스크림 판매량이 많은 날에는 물놀이 사고도 많다고 한다. 그렇다면 아이스크림 판매를 줄이면 사고가 줄어들까?

### 3.1 산점도와 상관

가상의 매장이 8일 동안 기록한 기온과 아이스크림 판매량이 표 3-1에 있다. 한 행이 하루(관측 단위)이고, 두 열이 변수다.

표 3-1 기온과 판매량 (가상 자료)

| 날 | 기온 $x$ (°C) | 판매량 $y$ (개) |
|---|-------------|-------------|
| 1 | 19          | 112         |
| 2 | 21          | 135         |
| 3 | 23          | 148         |
| 4 | 24          | 160         |
| 5 | 26          | 171         |
| 6 | 28          | 196         |
| 7 | 30          | 204         |
| 8 | 32          | 231         |

#### 3.1.1 산점도 읽기

두 변수의 값을 좌표로 삼아 점을 찍은 그림이 **산점도**다(그림 3-1). 점들이 오른쪽 위로 올라가면 양의 관계, 오른쪽 아래로 내려가면 음의 관계다. 점들이 직선에 가깝게 모일수록 관계가 강하다.

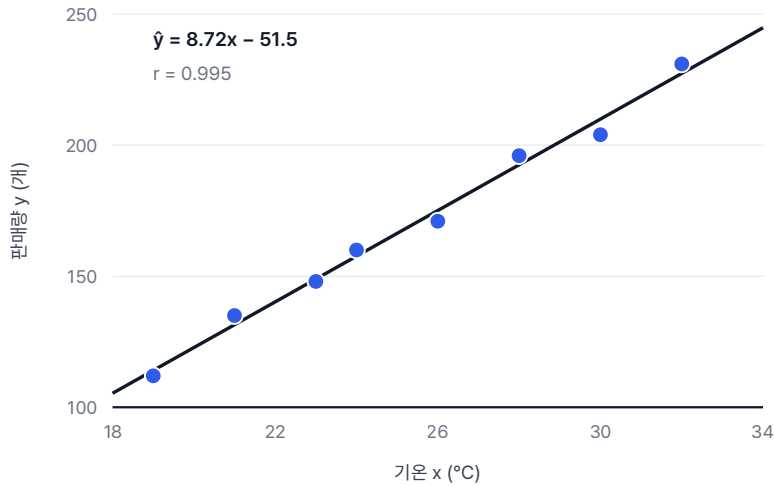


그림 3-1 기온과 판매량의 산점도와 회귀 직선 (3/4 쪽)

### 3.1.2 상관계수

직선 관계의 방향과 강도를 -1에서 1 사이의 수로 나타낸 것이 **상관계수**  $r$ 이다(식 (3-1)). 여기서  $S_{xy}$ 는 두 편차를 곱해 더한 값이고  $S_{xx}$ ,  $S_{yy}$ 는 각 편차 제곱의 합이다.

$$r = \frac{S_{xy}}{\sqrt{(S_{xx} S_{yy})}}, S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \quad (3-1)$$

#### 예제 3-1 상관계수 계산하기

다섯 쌍의 자료  $(x, y) = (1, 2), (2, 4), (3, 5), (4, 4), (5, 5)$ 의 상관계수를 구하시오.

풀이

- 1 평균:  $\bar{x} = 3, \bar{y} = 4$ . 편차:  $x$ 는 -2, -1, 0, 1, 2이고  $y$ 는 -2, 0, 1, 0, 1이다.
- 2  $S_{xy} = 4 + 0 + 0 + 0 + 2 = 6, S_{xx} = 10, S_{yy} = 4 + 0 + 1 + 0 + 1 = 6$ .
- 3  $r = 6 \div \sqrt{(10 \times 6)} = 6 \div \sqrt{60} \approx 0.775$ .

답  $r \approx 0.775$  (꽤 강한 양의 상관)

**주의**

상관은 인과가 아니다. 두 변수가 함께 움직여도 한쪽이 다른 쪽의 원인이라는 보장은 없다. 기온 같은 제3의 변수가 둘 모두에 영향을 줄 수 있다.

**확인 문제**

**3-1** 상관계수가 0이면 두 변수 사이에 아무 관계도 없다고 할 수 있는가? 정답 24쪽

**3-2** 표 3-1에서 기온이 가장 높은 날의 판매량은 몇 개인가? 정답 24쪽

**3.2 단순 선형 회귀**

산점도의 점들을 가장 잘 대표하는 직선  $\hat{y} = a + bx$ 를 **회귀 직선**이라 한다. 각 점과 직선의 세로 거리(잔차)를 제곱해 더한 값이 가장 작아지도록  $a$ 와  $b$ 를 정하는 방법이 최소제곱법이며, 그 해는 식 (3-2)와 같다.

$$b = \frac{S_{xy}}{S_{xx}}, a = \bar{y} - b\bar{x} \quad (3-2)$$

**예제 3-2 회귀 직선으로 판매량 예측하기**

표 3-1의 자료에서  $\bar{x} = 25.375$ ,  $\bar{y} = 169.625$ ,  $S_{xy} = 1219.125$ ,  $S_{xx} = 139.875$ 다. 회귀 직선을 구하고 기온 27°C인 날의 판매량을 예측하시오.

풀이

- 기울기:  $b = 1219.125 \div 139.875 \approx 8.716$ .
- 절편:  $a = 169.625 - 8.716 \times 25.375 \approx -51.54$ .
- 예측:  $\hat{y} = -51.54 + 8.716 \times 27 \approx 183.8$ .

**답**  $\hat{y} = 8.72x - 51.5$ , 27°C에서 약 184개

**팁**

회귀 직선은 언제나 두 평균의 점  $(\bar{x}, \bar{y})$ 을 지난다. 계산한 직선에 평균을 넣어 보면 실수를 빨리 찾을 수 있다.

코드 3-1은 식 (3-2)를 그대로 코드로 옮긴 것이다. 외부 라이브러리 없이도 회귀 직선과 상관계수를 구할 수 있다.

**코드 3-1** 최소제곱 회귀와 상관계수 (외부 라이브러리 없음)

```

1  from math import sqrt
2
3  x = [19, 21, 23, 24, 26, 28, 30, 32]          # 기온(°C)
4  y = [112, 135, 148, 160, 171, 196, 204, 231] # 판매량(개)
5  mx, my = sum(x) / len(x), sum(y) / len(y)
6  sxy = sum((a - mx) * (b - my) for a, b in zip(x, y))
7  sxx = sum((a - mx) ** 2 for a in x)
8  syy = sum((b - my) ** 2 for b in y)
9
10 b = sxy / sxx
11 a = my - b * mx
12 r = sxy / sqrt(sxx * syy)
13 print(f"y = {a:.2f} + {b:.3f}x, r = {r:.3f}")    # y = -51.54 + 8.716x, r = 0.995

```

#### 읽을거리

##### ‘회귀’라는 이름

회귀(regression)는 ‘되돌아감’이라는 뜻이다. 19세기 말 부모와 자녀의 키를 비교한 연구에서, 키가 아주 큰 부모의 자녀는 부모보다 평균 쪽으로 ‘되돌아가는’ 경향이 관찰되었고, 이 현상을 설명하는 직선에 회귀라는 이름이 붙었다. 지금은 한 변수로 다른 변수를 설명하는 모형 전체를 가리킨다.

## 3.3 예측과 그 한계

### 3.3.1 결정계수

상관계수를 제곱한  $r^2$ 을 **결정계수**라 한다. 판매량의 전체 변동 가운데 회귀 직선이 설명하는 비율이다. 표 3-1의 자료에서  $r \approx 0.995$ 이므로  $r^2 \approx 0.991$ , 곧 변동의 약 99%를 기온으로 설명한다.

### 3.3.2 외삽의 위험

회귀 직선은 자료가 있는 범위(여기서는 19~32°C) 안에서만 믿을 수 있다. 범위 밖으로 직선을 늘어 예측하는 것을 **외삽**이라 하며, 관계가 범위 밖에서도 직선이라는 보장이 없으므로 조심해야 한다.

## 심화

### 잔차 그림

각 점의 잔차( $y - \hat{y}$ )를  $x$ 에 대해 다시 찍어 보면 직선 모형이 맞는지 알 수 있다. 잔차가 0 주위에 고르게 흩어지면 좋고, 휘어진 모양이 보이면 직선이 아닌 모형을 생각해야 한다.

## 확인 문제

**3-3** 상관계수가 0.9이면 결정계수는 얼마인가? 정답 24쪽

## 요약

- 01** 산점도는 두 변수의 관계를 점으로 보여 준다. 방향(양·음)과 강도(직선에 모인 정도)를 읽는다.
- 02** 상관계수  $r$ 은  $-1 \sim 1$  사이 값이며, 상관이 있어도 인과라고 단정할 수 없다.
- 03** 최소제곱 회귀 직선은  $b = S_{xy}/S_{xx}$ ,  $a = \bar{y} - b\bar{x}$ 로 구하며 평균의 점을 지난다.
- 04** 결정계수  $r^2$ 은 설명된 변동의 비율이고, 자료 범위 밖의 예측(외삽)은 믿기 어렵다.

## 핵심 용어 정리

**산점도** 두 변수의 값을 좌표로 삼아 찍은 점 그림.  
scatter plot

**상관계수** 직선 관계의 방향과 강도를  $-1 \sim 1$ 로 나타낸 값. 식 (3-1).  
correlation coefficient

**회귀 직선** 잔차 제곱의 합이 가장 작은 직선  $\hat{y} = a + bx$ .  
regression line

**결정계수** 회귀 직선이 설명하는 변동의 비율  $r^2$ .  
coefficient of determination

**외삽** 자료 범위 밖으로 모형을 늘려 예측하는 일.  
extrapolation

## 연습 문제

■ ■ ■ 기본 개념을 그대로 적용한다

**3-1** 상관계수  $r = -0.9$ 는 두 변수의 관계에 대해 무엇을 말해 주는가? 정답 24쪽

**3-2** 회귀 직선  $\hat{y} = 2.2 + 0.6x$ 에서  $x = 4$ 일 때의 예측값을 구하시오. 정답 24쪽

■ ■ ■ 응용 두 개념을 함께 쓴다

**3-3** 자료 (2, 3), (4, 7), (6, 5), (8, 9)의 회귀 직선과 상관계수를 구하시오. 정답 25쪽

**3-4** 예제 3-2의 직선으로 기온 40°C인 날의 판매량을 예측하고, 이 예측을 그대로 믿기 어려운 까닭을 쓰시오. 정답 25쪽

■ ■ ■ 심화 스스로 방법을 고른다

**3-5** ‘생각해 보기’의 질문에 답하시오. 아이스크림 판매량과 물놀이 사고 건수의 상관을 설명할 수 있는 다른 변수를 제시하고, 인과 관계를 확인하려면 어떤 자료가 더 필요한지 쓰시오. 정답 25쪽



## 정답 및 해설

확인 문제와 연습 문제의 정답이다. 번호를 누르면 문제로 돌아간다. 풀이 과정이 여럿인 문제는 대표 풀이 하나만 실는다.

### A.1 1장 데이터와 측정

#### 확인 문제

- 1-1** 관측 단위: 하루. 변수: 최고 기온, 강수량. 한 행이 하루, 한 열이 변수가 된다.
- 1-2** 비정형 자료. 정해진 열 구조가 없으므로 분석하려면 주제어 같은 특성을 뽑아 정형 자료로 바꾼다.
- 1-3** 값이 빈 문자열인지 검사한다. `value == ""`이면 결측으로 세고 목록에 넣지 않는다.

#### 연습 문제

- 1-1** (1) 명목 (2) 순서 (3) 구간 (4) 비율.
- 1-2**  $77^{\circ}\text{F}$ .  $9/5 \times 25 + 32 = 45 + 32 = 77$ .
- 1-3** 5m. 축척 1:200이므로  $25\text{mm} \times 200 = 5,000\text{mm} = 5\text{m}$ . 도면의 치수선 '5 000'과도 일치한다.
- 1-4** 옳지 않다. 켈빈으로는 약 1.035배. 섭씨는 0이 '없음'이 아닌 구간 척도라 비를 말할 수 없다.  $283.15\text{K} \rightarrow 293.15\text{K}$ 이므로  $293.15 \div 283.15 \approx 1.035$ .
- 1-5** 예: ① 결측일을 빼고 28일로 평균 ② 앞뒤 날의 평균으로 채움. ①은 간단하고 값을 지어내지 않지만 결측이 특정 날씨에 몰리면 치우친다. ②는 30일을 유지하지만 채운 값이 실제 측정이 아니므로 보고서에 방법을 밝혀야 한다.

## A.2 2장 자료 요약하기

### 확인 문제

- 2-1** 7. 3, 5, 7, 9, 11에서 가운데 값.
- 2-2** 중앙값.
- 2-3** 0. 모든 편차가 0이다.

### 연습 문제

- 2-1** 평균 6, 중앙값 6. 합  $30 \div 5 = 6$ . 순서대로 4, 5, 6, 7, 8.
- 2-2** 0.233.  $7 \div 30 \approx 0.233$ .
- 2-3** 분산 4.571, 표준편차 2.138. 평균 5, 편차 제곱의 합  $9 + 1 + 1 + 1 + 0 + 0 + 4 + 16 = 32$ ,  $32 \div 7 \approx 4.571$ ,  $\sqrt{4.571} \approx 2.138$ .
- 2-4** 평균 20, 중앙값 12. 중앙값이 알맞다. 54라는 이상값이 평균을 끌어올려, 다섯 가구 중 네 가구가 평균보다 소득이 적다.
- 2-5** 평균 34, 표준편차 약 5.48. 더한 수는 평균만 옮기고 퍼짐은 바꾸지 않는다. 곱한 수는 평균과 표준편차를 모두 그 배수로 바꾼다.  $(14 + 3) \times 2 = 34$ ,  $2.74 \times 2 = 5.48$ .

## A.3 3장 두 변수의 관계

### 확인 문제

- 3-1** 아니다. 상관계수는 직선 관계만 잰다. U자 모양처럼 곡선 관계가 있어도  $r$ 은 0에 가까울 수 있다.
- 3-2** 231개. 8일째,  $32^{\circ}\text{C}$ .
- 3-3** 0.81.

### 연습 문제

- 3-1** 강한 음의 직선 관계. 한 변수가 커질 때 다른 변수는 대체로 작아지며, 점들이 직선 근처에 모여 있다.
- 3-2** 4.6.  $2.2 + 0.6 \times 4 = 4.6$ .

- 3-3**  $\hat{y} = 2 + 0.8x$ ,  $r = 0.8$ .  $\bar{x} = 5$ ,  $\bar{y} = 6$ ,  $S_{xy} = 16$ ,  $S_{xx} = 20$ ,  $S_{yy} = 20$ .  $b = 0.8$ ,  $a = 6 - 4 = 2$ ,  $r = 16 \div 20$ .
- 3-4** 약 297개. 외삽이므로 믿기 어렵다.  $8.716 \times 40 - 51.54 \approx 297.1$ . 자료는 19~32°C 범위뿐이며, 아주 더운 날에는 외출이 줄어 관계가 달라질 수 있다.
- 3-5** 공통 원인은 기온(계절). 더운 날에는 아이스크림도 많이 팔리고 물놀이도 늘어 사고가 늘어난다. 인과를 확인하려면 기온이 같은 날끼리 비교하거나, 판매를 조절한 실험 자료처럼 제3의 변수를 통제된 자료가 필요하다.



# 용어집

이 책에서 정의한 용어를 가나다순으로 모았다. 괄호 안은 처음 정의된 장이다.

|                                             |                                   |
|---------------------------------------------|-----------------------------------|
| <b>결정계수</b><br>coefficient of determination | 회귀 직선이 설명하는 변동의 비율. (3장)          |
| <b>결측값</b><br>missing value                 | 기록되지 않은 값. 0과 구별한다. (1장)          |
| <b>관측 단위</b><br>observational unit          | 측정의 대상이 되는 하나하나. (1장)             |
| <b>도수분포표</b><br>frequency table             | 계급마다 값의 개수를 센 표. (2장)             |
| <b>변수</b><br>variable                       | 관측 단위마다 달라질 수 있는 특성. (1장)         |
| <b>분산</b><br>variance                       | 편차 제곱의 합을 $n - 1$ 로 나눈 값. (2장)    |
| <b>산점도</b><br>scatter plot                  | 두 변수의 값을 좌표로 찍은 그림. (3장)          |
| <b>상관계수</b><br>correlation coefficient      | 직선 관계의 방향과 강도, $-1 \sim 1$ . (3장) |
| <b>외삽</b><br>extrapolation                  | 자료 범위 밖의 예측. (3장)                 |
| <b>중앙값</b><br>median                        | 크기 순서의 한가운데 값. (2장)               |
| <b>척도</b><br>scale of measurement           | 명목·순서·구간·비율. (1장)                 |
| <b>평균</b><br>mean                           | 합을 개수로 나눈 값. (2장)                 |
| <b>표준편차</b><br>standard deviation           | 분산의 제곱근. (2장)                     |
| <b>회귀 직선</b><br>regression line             | 잔차 제곱합이 최소인 직선. (3장)              |

---

## 히스토그램

도수분포를 틈 없는 막대로 그린 그림. (2장)

histogram

---

## 참고 문헌

아래 문헌은 서지 형식을 보여 주기 위한 **가상의 예시**다. 실제 원고에서는 실제 문헌으로 바꾼다.

### 국내 문헌

---

김예시. (2024). 『처음 배우는 통계 (예시)』. 예시출판.

이건본·박자리. (2025). 관측 자료의 결측 처리 방법 비교 (예시). 『예시 학회지』, 12(3), 45–67.

### 국외 문헌

---

Sample, A., & Placeholder, B. (2023). Introductory data science (example). Example Press.

Author, C. (2022). Regression for beginners (example). Journal of Examples, 8(2), 101–120.



# 찾아보기

## ㄱ

---

결정계수 19  
결측값 5  
관측 단위 2

## ㄴ

---

단위 변환 5  
데이터 2  
도면  
    배치도 3  
도수분포 10

## ㄷ

---

변수 2  
분산 12  
비정형 자료 3

## ㄹ

---

사진  
    관측 현장 2  
산점도 16  
상관계수 17

## ㅇ

---

외삽 19  
이상값 11

## ㅈ

---

자유도 12  
정형 자료 2  
중앙값 11

## ㅊ

---

척도 4  
최소제곱법 18  
측정 오차 6

## 표

---

평균 11  
표준편차 12

## ㅎ

---

회귀 직선 18  
히스토그램 10

## 콜로폰

[예시] 기초 데이터 과학 — 교재 템플릿 견본. 판형 46배판 188 × 257mm, 판면 146 × 209mm. 본문 Pretendard 9.5/16pt, 제목 Pretendard Bold, 코드 JetBrains Mono 8/12.5pt (두 서체 모두 SIL OFL 1.1).

HTML 원고를 WeasyPrint로 조판했다. 번호, 차례, 표·그림·도면 차례, 찾아보기, 문제와 정답의 상호 링크는 빌드 도구(tools/build.py)가 만들었다. 사진과 도면은 tools/make\_assets.py로 만든 합성 자료다. 빌드 v1.0.0 · 2026년 10월 3일.