TEXTBOOK TEMPLATE · v1.0.0

Example

# Fundamentals of Data Science

Sample content for a textbook design template. All topics, figures, people and institutions are fictional examples and are not tied to any real curriculum.

| | |
|---|---|
| Author | Editorial Team (example author) |
| Edition · Date | First edition v1.0.0 · October 3, 2026 |
| Audience | Undergraduate general education · one 15-week semester (example) |

For print, the cover is supplied as a separate file (front cover + spine + back cover + 3 mm bleed).

# [Example] Fundamentals of Data Science

# [Example]
# Fundamentals of Data Science

From Measurement to Regression — a textbook template sample

Editorial Team (example author)

Example Press (fictional)

| | |
|---|---|
| Title | [Example] Fundamentals of Data Science |
| Edition | First edition, first printing (v1.0.0) · October 3, 2026 |
| Author | Editorial Team (example author) |
| Publisher | Example Press (fictional) · address and phone number placeholder |
| ISBN | 978-0-00-000000-0 (example number) |
| Price | $00.00 |


This document is a sample made with a textbook design template. The ISBN, publisher, author and all data are examples with no connection to real ones. Photographs and drawings are synthetic material created for this sample.

All rights reserved. No part of this publication may be reproduced or transmitted in any form without the prior written permission of the copyright holder.

# Preface

This book is a sample that shows what the textbook template looks like in practice. Its subject, “fundamentals of data science,” was chosen only because it exercises every learning element of a textbook.

Unlike a general book, a textbook is not read once from cover to cover. Learners check the goals on the chapter opener, move back and forth between the text and the worked examples, test their understanding with check questions, and return to the summary and exercises at the end of the chapter. The most important rule of textbook typesetting is therefore that the same kind of information always appears in the same form and in the same place.

The sample follows the publishing design system and build standard bundled with the template. The trim size is 188 × 257 mm, body text is set at 9.5/16 pt in the theme’s Latin text face, and code is set in JetBrains Mono at 8/12.5 pt. The build produces symmetrical margins, chapters that start on a right-hand page, roman folios for the front matter, an automatic table of contents and list of figures, an index and PDF bookmarks. Chapter, section, table, figure, equation, example and question numbers are never typed into the manuscript.

When writing a new manuscript, read “How to Use This Book” on the next page together with the repository’s `README.md`. HTML samples of every learning element are collected there.

October 2026

Editorial Team (example)

# How to Use This Book

Every chapter follows the same sequence. Learning elements always appear in the same form and in the same place, so once you know them you can find what you need in any chapter.

**① Chapter opener**
Learning goals · key terms · warm-up

**② Main text**
Sections · worked examples · tips · deeper dives

**③ Check questions**
A quick self-test after each section

**④ Chapter wrap-up**
Summary · key terms · exercises

**⑤ Answers**
Appendix A · page reference for each question

## Learning elements

★ Learning goals
**01** Calculate a mean.

**Learning goals** `.objectives`
Three to five goals on the chapter opener, each phrased as something the reader will be able to do. Their numbers match the numbered points of the chapter summary one to one.

Key terms
mean · variance

**Key terms** `.keyterms`
Terms defined in the chapter. They are set in bold where they first appear and are listed again in the key-term review at the end of the chapter and in the index.

Warm-up
If two classes have the same average, are their results the same?

**Warm-up** `.warmup`
A question to think about before the chapter. It has no single right answer; the text comes back to it so you can answer it again at the end.

Example 2-1 Sample mean
Solution
Answer 14

**Worked example** `.example`
Problem → step-by-step solution → answer. The build assigns the number "Example n-m"; the text refers to it with an `xref`.

Tip

Organize the data in a table before you calculate.

**Tip** `.callout.tip`

A time-saving trick or memory aid, marked with a vertical bar in the accent color.

Caution

Don't confuse n with n − 1.

**Caution** `.callout.warn`

Common mistakes and easily misunderstood ideas. Never place more than two callouts in a row.

Deeper dive

Why divide by n − 1?

**Deeper dive** `.sidebar.deep`

Material beyond the exam syllabus. You can skip it without losing the thread of the next section.

Further reading

Where the word "statistics" comes from

**Further reading** `.sidebar.read`

History, case studies and career stories that spark interest. It sits outside the main flow of the text.

Check

2-1 What is the median?
answer page

**Check questions** `.check`

One to three questions to answer at the end of a section. The "Answer p. n" note after each question links to its solution in Appendix A.

Summary

01 The mean is the sum of the values divided by their count.

**Summary · key-term review** `.summary` `.termlist`

At the end of the chapter the summary follows the order of the learning goals, and the key terms are listed again with their definitions.

Applied

**Exercises** `.exercises` `.level`

Three levels: basic, applied and challenge. The number of accent bars (one to three) shows the difficulty. Answers and explanations for every question are in Appendix A.

## Reading Numbers and References

| Notation | Meaning | Example |
|---|---|---|
| 2.3, 2.3.1 | Chapter 2, section 3; chapter 2, section 3, subsection 1 | 2.3 Spread |
| Table/Figure n-m | The m-th table or figure in chapter n (drawings work the same way) | Table 2-1, Figure 3-1 |
| Equation (n-m) | The m-th numbered equation in chapter n | Equation (2-2) |
| Check n-m | The m-th check question in chapter n | Check 1-2 |
| Exercise n-m | The m-th exercise in chapter n | Exercise 3-4 |

# Contents

List of Tables and Figures ..... xiii

1 Data and Measurement ..... 1

1.1 What Are Data? ..... 2
1.2 Types of Variables and Scales ..... 4
1.3 Organizing Measurements in a Table ..... 6

2 Descriptive Statistics ..... 11

2.1 Frequency Distributions and Histograms ..... 12
2.2 Central Tendency: Mean and Median ..... 13
2.3 Spread: Variance and Standard Deviation ..... 14

3 Correlation and Regression ..... 19

3.1 Scatter Plots and Correlation ..... 20
3.2 Simple Linear Regression ..... 22
3.3 Prediction and Its Limits ..... 24

Appendix A Answers and Explanations ..... 27

A.1 Chapter 1 Data and Measurement ..... 27
A.2 Chapter 2 Descriptive Statistics ..... 28
A.3 Chapter 3 Correlation and Regression ..... 29

Glossary ..... 31

References ..... 33

Index ..... 35

# List of Tables and Figures

## Tables


**Chapter 1**

**Table 1-1** The four scales of measurement ..... 4

**Chapter 2**

**Table 2-1** Frequency distribution of daily maximum temperatures in June (fictional data, n = 30) ..... 12

**Chapter 3**

**Table 3-1** Temperature and sales (fictional data) ..... 20


## Figures


**Chapter 1**

**Figure 1-1** The field where the instruments are installed (3/4 width, 108.5 mm, 3:2) ..... 2

**Chapter 2**

**Figure 2-1** Histogram of daily maximum temperatures ..... 13

**Chapter 3**

**Figure 3-1** Scatter plot of temperature and sales with the regression line (3/4 width) ..... 21


## Drawings


**Chapter 1**

**Drawing**

**1-1** Site plan of the weather-station enclosure (fictional) ..... 3

CHAPTER 01

# 01 Data and Measurement

Data science starts with knowing what was measured, and how. This chapter covers units, variables, scales and tables.

## Learning goals

**01** Tell observational units from variables.

**02** Classify variables as nominal, ordinal, interval or ratio.

**03** Convert units and pick operations that suit the scale.

**04** Organize measurements in a table.

## Key terms

- observational unit
- variable
- scale
- structured data
- missing value

## Warm-up

"Today was twice as hot as yesterday." If it was 10 °C yesterday and 20 °C today, is that really true?

## 1.1 What Are Data?

**Data** are recorded results of observations or measurements. The temperature, rainfall and wind speed that a weather station records every day are typical examples. This book uses data from a fictional weather station throughout.

### 1.1.1 Observational Units and Variables

Each individual thing that is measured is an **observational unit**, and a characteristic whose value can differ from unit to unit is a **variable**. In the weather-station data, one day is the observational unit, and that day's maximum temperature and rainfall are variables.

Figure 1-1 shows the field where the instruments stand. Their exact layout is given in the site plan in Drawing 1-1 (p. 3). A photograph conveys atmosphere and context; a drawing conveys position and dimensions.

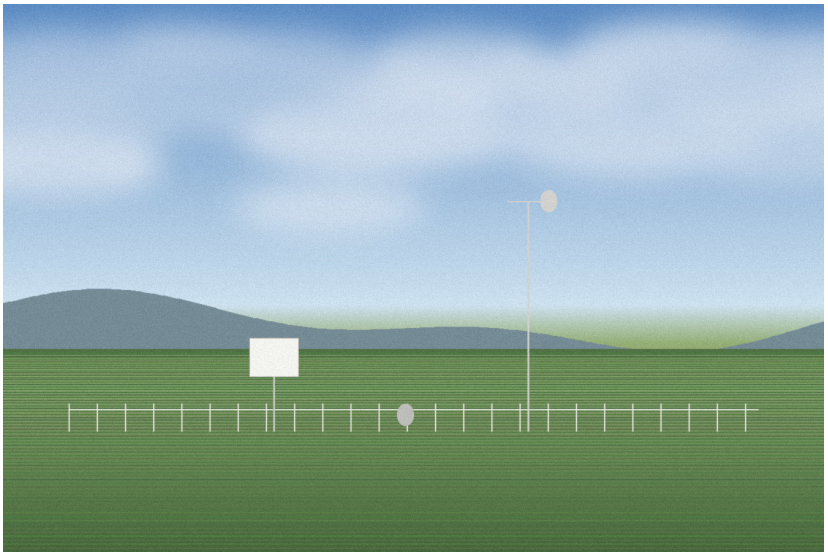


**Figure 1-1** The field where the instruments are installed (3/4 width, 108.5 mm, 3:2)
Synthetic image (placeholder)

### 1.1.2 Forms of Data

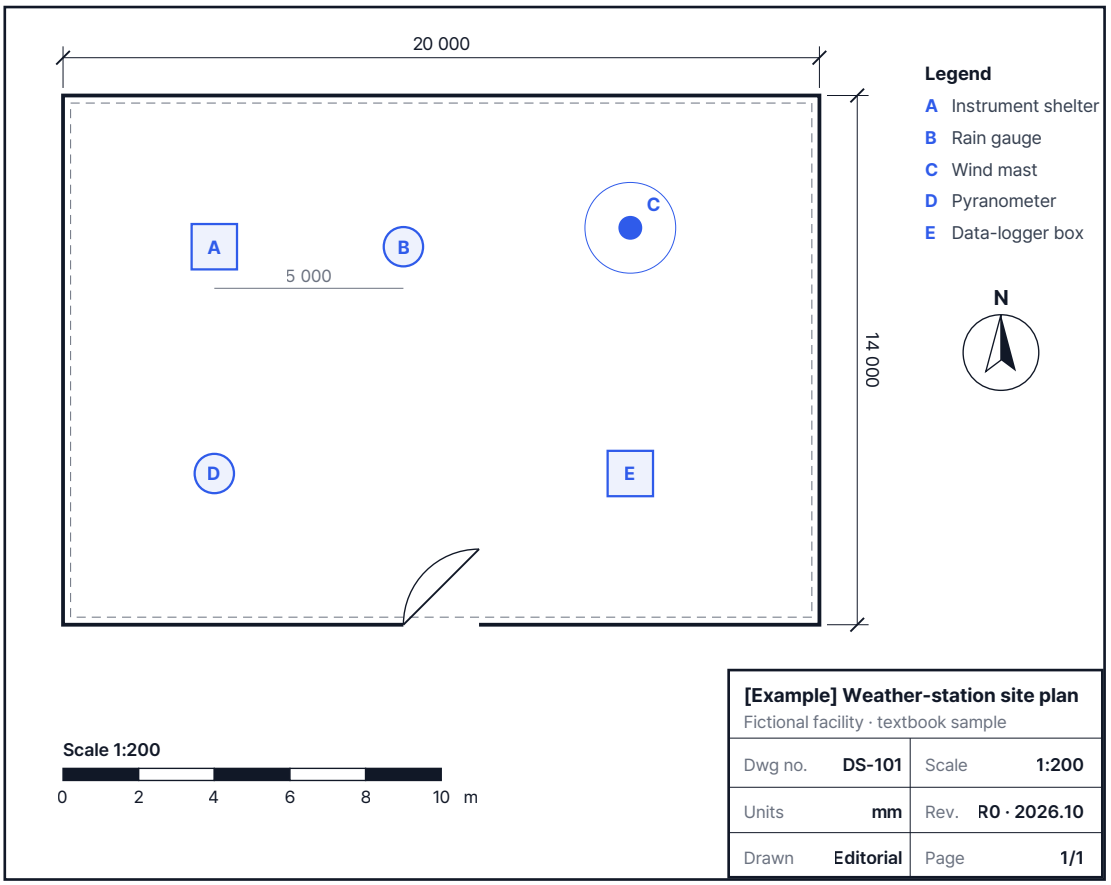
Data fall into two groups according to the shape they are stored in.

#### Structured data

Data that fit a table with fixed rows and columns are **structured data**. Each row is an observational unit and each column is a variable. All the data in this book are structured.

#### Unstructured data

Data with no fixed table structure, such as photographs, audio or free text, are unstructured data. To analyze them, you first extract the characteristics you need and turn them into structured data.




**Drawing 1-1** Site plan of the weather-station enclosure (fictional). Scale 1:200 (original 1:100), dimensions in mm, drawing no. DS-101

✔ Check

1-1 A station recorded the maximum temperature and rainfall every day for 30 days. Name the observational unit and the variables. Answer p. 27

1-2 Are free-text answers on a questionnaire structured or unstructured data? Answer p. 27

## 1.2 Types of Variables and Scales

Variables are divided into four **scales** according to the nature of their values (Table 1-1). The scale determines which operations and summaries you can use.

**Table 1-1** The four scales of measurement

| Scale | Nature of the values | Valid operations | Examples |
|---|---|---|---|
| **Nominal** | Names only; no order | equal / not equal | blood type, station name |
| **Ordinal** | Ordered, but intervals are not equal | + greater / less | school year, satisfaction (high, medium, low) |
| **Interval** | Equal intervals, but zero does not mean “none” | + add / subtract | degrees Celsius, calendar year |
| **Ratio** | Equal intervals and zero means “none” | + multiply / divide | rainfall, body weight |

Note: each scale allows all the operations of the scales above it.

**Example 1-1 Classifying variables by scale**

Give the scale of each weather-station variable: “station code,” “cloud cover (clear, partly cloudy, overcast),” “maximum temperature (°C)” and “rainfall (mm).”

Solution

1. A station code is only a label with no size or order → nominal.
2. Cloud cover has an order, but the step from “clear” to “partly cloudy” is not necessarily the same as the other steps → ordinal.
3. 0 °C does not mean “no temperature” → interval. 0 mm of rainfall means “no rain” → ratio.

Answer **nominal · ordinal · interval · ratio**

**Tip**

If you are unsure of the scale, ask two questions in turn: “Does zero mean none?” and “Can I say twice as much?” If both answers are yes, it is a ratio scale.

### 1.2.1 Converting Units

Celsius and Fahrenheit, both interval scales, are converted with **Equation (1-1)**. Because the two units have different zero points, multiplication alone is not enough.

$$F = \frac{9}{5}C + 32 \tag{1-1}$$

### Example 1-2 Converting Celsius to Fahrenheit

The station's average maximum temperature in June was 26.4 °C. Convert it to Fahrenheit.

Solution

1. Substitute $C$ = 26.4 into **Equation (1-1)**: 9/5 × 26.4 = 47.52.
2. Add 32: 47.52 + 32 = 79.52.

Answer **about 79.5 °F**

Further reading

***The Kelvin scale is a ratio scale***

The kelvin (K) scale used in science puts zero at absolute zero, where molecular motion stops. Because 0 K means "no heat," kelvin is a ratio scale, and "twice as hot" only makes sense when temperature is measured in kelvins. Add 273.15 to a Celsius temperature to get kelvins.

## 1.3 Organizing Measurements in a Table

Structured data put one observational unit in each row and one variable in each column. Data in this form can be read directly by most analysis tools. **Listing 1-1** reads the temperature column of a CSV file and counts the missing values.

**Listing 1-1** Reading the temperature column of a CSV file (Python standard library)

```python
import csv

temps, missing = [], 0
with open("station_june.csv", encoding="utf-8") as f:
    for row in csv.DictReader(f):      # one row = one day (observational unit)
        value = row["max_temp"].strip()   # one column = one variable
        if value == "":
            missing += 1 # a blank is a missing value
            continue
        temps.append(float(value))

print(len(temps), "days recorded,", missing, "days missing")
```

**Caution**

Never fill missing values with 0. A temperature of 0 °C is a real measurement, so replacing blanks with 0 drags the mean down sharply. Count missing values separately.

**Deeper dive**

***Measurement error***

Measuring the same thing several times gives slightly different values. Error that comes from the limits of the instrument is called measurement error, and the instrument's resolution (for example 0.1 °C) should be recorded with the data. Never report results to more digits than the resolution allows.

**Check**

1-3 In **Listing 1-1**, what does the line that counts a blank as a missing value test? Answer p. 27

## Summary

**01** An observational unit is each individual thing measured; a variable is a characteristic that differs from unit to unit.

**02** There are four scales — nominal, ordinal, interval and ratio — and each lower scale allows more operations.

**03** Units with different zero points, such as Celsius and Fahrenheit, are converted by multiplying and adding, as in **Equation (1-1)**.

**04** Structured data put an observational unit in each row and a variable in each column; missing values are never filled with 0.

## Key Terms

| | |
|---|---|
| **observational unit** | Each individual thing that is measured; one row of the table. |
| **variable** | A characteristic whose value can differ between observational units; one column of the table. |
| **scale of measurement** | A classification of variables by the nature of their values: nominal, ordinal, interval, ratio. |
| **structured data** | Data stored in a table with fixed rows and columns. |
| **missing value** | A value that was not recorded. Counted separately and never treated as 0. |

## Exercises

Basic Apply the concept directly

1-1 Give the scale of each variable: (1) blood type (2) school year (3) temperature in °C (4) body weight. Answer p. 27

**1-2** Convert 25 °C to Fahrenheit. Answer p. 27

**Applied** Combine two concepts

**1-3** In **Drawing 1-1**, the centers of instrument shelter A and rain gauge B are 25 mm apart on the page. What is the real distance in meters? Answer p. 28

**1-4** Use scales of measurement to explain why the warm-up claim "it went from 10 °C to 20 °C, so it is twice as hot" is wrong, and find the true ratio in kelvins. Answer p. 28

**Challenge** Choose your own method

**1-5** Two days are missing from a 30-day temperature record. Describe two ways to handle them when reporting the mean temperature, with the pros and cons of each. Answer p. 28

CHAPTER 02

# 02

# Descriptive Statistics

Thirty raw numbers reveal little. Tables, charts and a few statistics show the shape, center and spread of data.

★ Learning goals

**01** Build a frequency table and a histogram.

**02** Compare the mean and median and the effect of outliers.

**03** Calculate and interpret variance and standard deviation.

**04** Compute statistics in code and check them by hand.

Key terms

- frequency table
- histogram
- mean
- median
- variance
- standard deviation

Warm-up

Two classes both have a mathematics average of 70. Can we say their score distributions are the same? What else would we need to know?

## 2.1 Frequency Distributions and Histograms

A table that divides the range of values into classes and counts the values in each class is a **frequency table**. Table 2-1 groups the 30 daily maximum temperatures recorded in June at the fictional station into classes 2 °C wide.

**Table 2-1** Frequency distribution of daily maximum temperatures in June (fictional data, n = 30)

| Class (°C) | Frequency (days) | Relative frequency | Cumulative frequency |
|---|---|---|---|
| 21 to under 23 | 3 | 0.100 | 3 |
| 23 to under 25 | 5 | 0.167 | 8 |
| 25 to under 27 | 10 | 0.333 | 18 |
| 27 to under 29 | 7 | 0.233 | 25 |
| 29 to under 31 | 4 | 0.133 | 29 |
| 31 to under 33 | 1 | 0.033 | 30 |
| **Total** | **30** | **1.000** | — |

Source: fictional data created for this book. Relative frequencies are rounded to three decimal places.

A frequency table drawn as bars is a **histogram** (Figure 2-1). The bars touch because the classes are consecutive intervals of a continuous quantity.

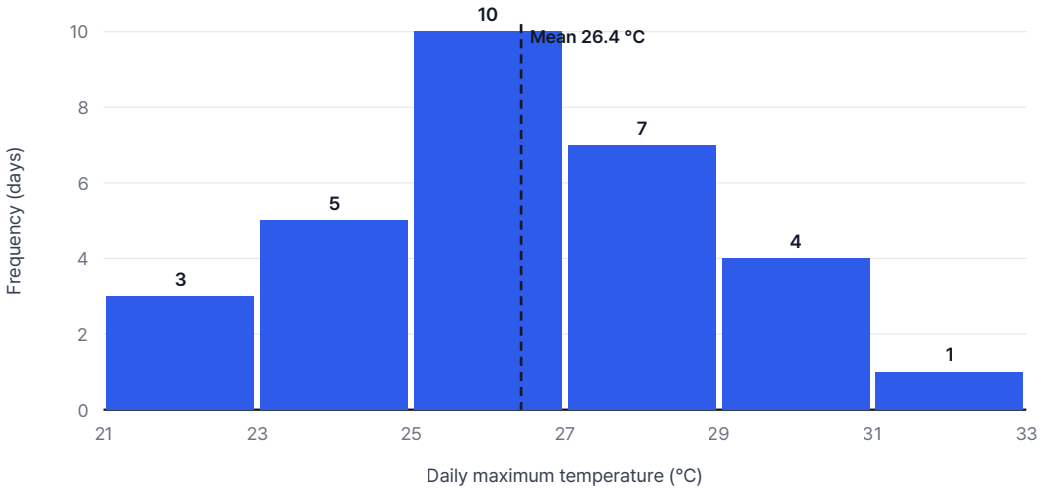



**Figure 2-1** Histogram of daily maximum temperatures. The dashed line is the mean

## 2.2 Central Tendency: Mean and Median

### 2.2.1 The Mean

The **mean** is the sum of all the values divided by their count. The mean of $n$ values $x_1, \ldots, x_n$ is given by Equation (2-1).

$$\bar{x} = \frac{1}{n} \Sigma_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{2-1}$$

### 2.2.2 The Median

When the values are sorted by size, the one in the middle is the **median**. If there is an even number of values, use the mean of the two middle values.

**Example 2-1 Finding the mean and median**

Five temperature readings are 12, 15, 11, 14 and 18 (°C). Find the mean and the median.

Solution

1. Sum: 12 + 15 + 11 + 14 + 18 = 70. Mean: 70 ÷ 5 = 14.
2. In order: 11, 12, **14**, 15, 18. The third of the five values is the median.

Answer **mean 14 °C, median 14 °C**

### The effect of outliers

A value far away from the others is an outlier. If the 18 in the example above were mistyped as 58, the mean would jump to 22 while the median would stay at 14. The median is less sensitive to outliers.

Tip

A large gap between the mean and the median signals skewed data or an outlier. Make a habit of reporting both.

Check

2-1 Find the median of the data 9, 3, 7, 5, 11. Answer p. 28

2-2 Which is less sensitive to outliers, the mean or the median? Answer p. 28

## 2.3 Spread: Variance and Standard Deviation

Two data sets with the same mean can differ in whether their values cluster near the mean or scatter widely. Square each deviation from the mean, add the squares and

divide by $n - 1$: the result is the sample **variance** (Equation (2-2)). The square root of the variance is the **standard deviation** (Equation (2-3)).

$$s^2 = \frac{1}{n - 1} \Sigma_{i=1}^{n} (x_i - \overline{x})^2 \tag{2-2}$$

$$s = \sqrt{s^2} \tag{2-3}$$

### Example 2-2 Finding the variance and standard deviation

Find the sample variance and standard deviation of the data in **Example 2-1**: 12, 15, 11, 14, 18.

Solution

1. The mean is 14, so the deviations are −2, 1, −3, 0 and 4.
2. Sum of squared deviations: 4 + 1 + 9 + 0 + 16 = 30.
3. Sample variance: 30 ÷ (5 − 1) = 7.5. Standard deviation: √7.5 ≈ 2.74.

Answer **variance 7.5, standard deviation about 2.74 °C**

### Caution

Calculators and software offer both the "population variance," which divides by $n$, and the "sample variance," which divides by $n - 1$. Every variance in this book is a sample variance.

Deeper dive

### Why divide by n − 1?

The sample mean sits as close as possible to the sample values, so the sum of squared deviations measured from it is slightly smaller than it would be from the true mean. Dividing by $n - 1$ instead of $n$ corrects this bias on average. The idea is often explained in terms of degrees of freedom.

Listing 2-1 uses the `statistics` module of the Python standard library to compute descriptive statistics for the 30 June temperatures. `variance` and `stdev` return the sample variance and sample standard deviation.

**Listing 2-1** Computing descriptive statistics, with the output

```python
import statistics as st

june = [21.4, 23.8, 25.1, 26.3, 24.7, 22.9, 27.5, 28.2, 26.8, 25.6,
        24.1, 29.3, 30.4, 27.9, 26.2, 25.4, 23.3, 28.7, 31.2, 29.8,
        27.1, 26.6, 24.9, 25.8, 28.4, 30.1, 27.4, 26.1, 22.6, 25.3]

for name, fn in [("mean", st.mean), ("median", st.median),
                 ("variance", st.variance), ("std dev", st.stdev)]:
    print(f"{name}: {fn(june):.2f}")
# mean: 26.43   median: 26.25   variance: 6.00   std dev: 2.45
```

Check

2-3 What is the variance of a data set in which every value is the same?

Answer p. 28

## Summary

**01** A frequency table counts the values in each class; a histogram draws it as touching bars.

**02** The mean is the sum divided by the count; the median is the middle value. The median is less sensitive to outliers.

**03** The sample variance is the sum of squared deviations divided by $n - 1$; the standard deviation is its square root.

**04** Comparing the output of the `statistics` module with hand calculations shows the difference between the formulas ($n$ versus $n - 1$).

## Key Terms

| | |
|---|---|
| **frequency table** | A table of the number of values (frequency) in each class. |
| **histogram** | A frequency distribution drawn as touching bars. |
| **mean** | The sum of all values divided by their count; **Equation (2-1)**. |
| **median** | The middle value when the data are sorted by size. |
| **variance** | The typical size of the squared deviations. The sample variance divides by $n - 1$. |
| **standard deviation** | The square root of the variance, in the same units as the data. |

## Exercises

Basic Apply the concept directly

2-1 Find the mean and median of the data 4, 8, 6, 5, 7. Answer p. 28

2-2 Using **Table 2-1**, find the relative frequency of days with a maximum temperature from 27 °C to under 29 °C. Answer p. 28

**Applied** Combine two concepts

**2-3** Find the sample variance and standard deviation of the data 2, 4, 4, 4, 5, 5, 7, 9 to three decimal places. Answer p. 28

**2-4** The monthly incomes of five households (in thousands of dollars) are 10, 12, 11, 13 and 54. Find the mean and the median, and say which better represents a "typical income." Answer p. 28

**Challenge** Choose your own method

**2-5** Each value in **Example 2-2** is increased by 3 and then multiplied by 2. Without calculating, give the mean and standard deviation of the new data and explain why. Answer p. 29

CHAPTER 03

# 03

# Correlation and Regression

Do warmer days sell more ice cream? This chapter measures relationships and fits a line for prediction.

★ Learning goals

**01** Read direction and strength from a scatter plot.

**02** Calculate $r$ and tell correlation from causation.

**03** Fit a least-squares line and use it to predict.

**04** Explain the limits of $r^2$ and extrapolation.

Key terms

- scatter plot
- correlation coefficient
- regression line
- extrapolation

Warm-up

Days with high ice-cream sales are said to have more swimming accidents. Would selling less ice cream reduce accidents?

## 3.1 Scatter Plots and Correlation

Table 3-1 lists the temperature and ice-cream sales recorded over eight days by a fictional shop. Each row is one day (the observational unit) and the two columns are variables.

**Table 3-1** Temperature and sales (fictional data)

| Day | Temperature $x$ (°C) | Sales $y$ (units) |
|---|---|---|
| 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 Reading a Scatter Plot

A plot that marks each pair of values as a point is a **scatter plot** (Figure 3-1). Points rising to the right show a positive relationship; points falling to the right show a negative one. The closer the points lie to a straight line, the stronger the relationship.

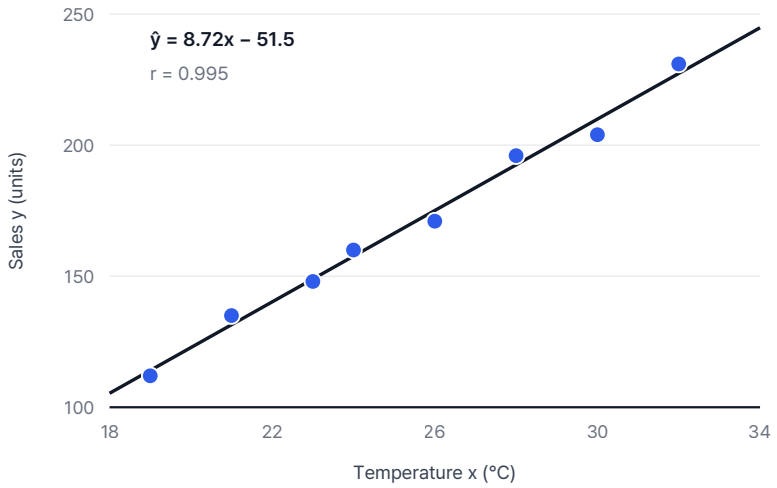



**Figure 3-1** Scatter plot of temperature and sales with the regression line (3/4 width)

### 3.1.2 The Correlation Coefficient

The **correlation coefficient** $r$ expresses the direction and strength of a linear relationship as a number between −1 and 1 (**Equation (3-1)**). Here $S_{xy}$ is the sum of the products of the two deviations, and $S_{xx}$ and $S_{yy}$ are the sums of squared deviations.

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

**Example 3-1 Calculating a correlation coefficient**

Find the correlation coefficient of the five pairs $(x, y)$ = (1, 2), (2, 4), (3, 5), (4, 4), (5, 5).

Solution

1. Means: $\bar{x} = 3$, $\bar{y} = 4$. Deviations: for $x$, −2, −1, 0, 1, 2; for $y$, −2, 0, 1, 0, 1.
2. $S_{xy} = 4 + 0 + 0 + 0 + 2 = 6$, $S_{xx} = 10$, $S_{yy} = 4 + 0 + 1 + 0 + 1 = 6$.
3. $r = 6 \div \sqrt{(10 \times 6)} = 6 \div \sqrt{60} \approx 0.775$.

Answer **$r \approx 0.775$ (a fairly strong positive correlation)**

**Caution**

Correlation is not causation. Two variables can move together without either one causing the other; a third variable, such as temperature, may affect both.

**Check**

3-1 If the correlation coefficient is 0, can we say there is no relationship at all between the two variables? Answer p. 29

3-2 In Table 3-1, how many units were sold on the hottest day? Answer p. 29

## 3.2 Simple Linear Regression

The straight line $\hat{y} = a + bx$ that best represents the points of a scatter plot is the **regression line**. The method of least squares chooses $a$ and $b$ so that the sum of the squared vertical distances (residuals) between the points and the line is as small as possible; the solution is given by Equation (3-2).

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

**Example 3-2 Predicting sales with a regression line**

For the data in **Table 3-1**, $\bar{x}$ = 25.375, $\bar{y}$ = 169.625, $S_{xy}$ = 1219.125 and $S_{xx}$ = 139.875. Find the regression line and predict sales on a 27 °C day.

Solution

1. Slope: $b$ = 1219.125 ÷ 139.875 ≈ 8.716.
2. Intercept: $a$ = 169.625 − 8.716 × 25.375 ≈ −51.54.
3. Prediction: $\hat{y}$ = −51.54 + 8.716 × 27 ≈ 183.8.

Answer **$\hat{y}$ = 8.72$x$ − 51.5; about 184 units at 27 °C**

Tip

A regression line always passes through the point of means $(\bar{x}, \bar{y})$. Plugging the means into your line is a quick way to catch mistakes.

**Listing 3-1** translates **Equation (3-2)** directly into code. No external library is needed to find the regression line and the correlation coefficient.

**Listing 3-1** Least-squares regression and correlation (no external libraries)

```
from math import sqrt

x = [19, 21, 23, 24, 26, 28, 30, 32]          # temperature (°C)
y = [112, 135, 148, 160, 171, 196, 204, 231]  # sales (units)
mx, my = sum(x) / len(x), sum(y) / len(y)
sxy = sum((a - mx) * (b - my) for a, b in zip(x, y))
sxx = sum((a - mx) ** 2 for a in x)
syy = sum((b - my) ** 2 for b in y)

b = sxy / sxx
a = my - b * mx
r = sxy / sqrt(sxx * syy)
print(f"y = {a:.2f} + {b:.3f}x, r = {r:.3f}")   # y = -51.54 + 8.716x, r = 0.995
```

Further reading

### Where "regression" gets its name

"Regression" means "going back." In a late-nineteenth-century study comparing the heights of parents and children, the children of very tall parents tended to "go back" toward the average, and the line describing this effect was called a regression line. Today the word covers any model that explains one variable with another.

## 3.3 Prediction and Its Limits

### 3.3.1 The Coefficient of Determination

The square of the correlation coefficient, $r^2$, is the **coefficient of determination**: the share of the total variation in sales that the regression line explains. For the data in Table 3-1, $r \approx 0.995$, so $r^2 \approx 0.991$ — temperature explains about 99% of the variation.

### 3.3.2 The Risks of Extrapolation

A regression line can be trusted only within the range of the data (here 19–32 °C). Extending the line beyond that range to make predictions is called **extrapolation**; it calls for caution, because nothing guarantees that the relationship stays linear outside the range.

Deeper dive

### Residual plots

Plotting each point's residual $(y - \hat{y})$ against $x$ shows whether a straight-line model is appropriate. Residuals scattered evenly around 0 are a good sign; a curved pattern means you should consider a non-linear model.

✓ Check

3-3 If the correlation coefficient is 0.9, what is the coefficient of determination? Answer p. 29

## Summary

**01** A scatter plot shows the relationship between two variables as points; read its direction (positive or negative) and strength (how tightly the points follow a line).

**02** The correlation coefficient $r$ lies between −1 and 1, and correlation alone never proves causation.

**03** The least-squares regression line minimizes the sum of squared residuals and passes through the point of means.

**04** The coefficient of determination $r^2$ is the share of variation explained; predictions outside the range of the data (extrapolation) are unreliable.

## Key Terms

| | |
|---|---|
| **scatter plot** | A plot of points whose coordinates are the values of two variables. |
| **correlation coefficient** | A number from −1 to 1 giving the direction and strength of a linear relationship; **Equation (3-1)**. |
| **regression line** | The line $\hat{y} = a + bx$ with the smallest sum of squared residuals. |
| **coefficient of determination** | The share of variation explained by the regression line, $r^2$. |

| | |
|---|---|
| **extrapolation** | Extending a model beyond the range of the data to make predictions. |

## Exercises

**Basic** Apply the concept directly

**3-1** What does a correlation coefficient of $r = -0.9$ tell you about the relationship between two variables? Answer p. 29

**3-2** Using the regression line $\hat{y} = 2.2 + 0.6x$, find the predicted value when $x = 4$. Answer p. 29

**Applied** Combine two concepts

**3-3** Find the regression line and the correlation coefficient for the data (2, 3), (4, 7), (6, 5), (8, 9). Answer p. 29

**3-4** Use the line from **Example 3-2** to predict sales on a 40 °C day, and explain why this prediction is hard to trust. Answer p. 29

**Challenge** Choose your own method

**3-5** Answer the warm-up question. Suggest another variable that could explain the correlation between ice-cream sales and swimming accidents, and describe what additional data you would need to establish a causal relationship. Answer p. 29

APPENDIX A

# Answers and Explanations

Answers to the check questions and exercises. Select a number to return to the question. Where a question can be solved in several ways, one representative solution is given.

## A.1 Chapter 1 Data and Measurement

### Check questions

1-1 **Observational unit: one day. Variables: maximum temperature and rainfall.** Each row is a day and each column is a variable.

1-2 **Unstructured data.** There is no fixed column structure, so to analyze the answers you first extract characteristics such as keywords and turn them into structured data.

1-3 **It tests whether the value is an empty string.** If `value == ""`, the day is counted as missing and not added to the list.

### Exercises

1-1 **(1) nominal (2) ordinal (3) interval (4) ratio.**

1-2 **77 °F.** 9/5 × 25 + 32 = 45 + 32 = 77.

1-3 **5 m.** At a scale of 1:200, 25 mm × 200 = 5,000 mm = 5 m. This matches the "5 000" dimension line on the drawing.

1-4 **The claim is wrong; in kelvins the ratio is about 1.035.** Celsius is an interval scale whose zero does not mean "none," so ratios are meaningless. 283.15 K → 293.15 K, and 293.15 ÷ 283.15 ≈ 1.035.

1-5 **For example: ① average the remaining 28 days; ② fill each gap with the mean of the days before and after.** ① is simple and invents no values, but it is biased if the missing days share a particular kind of weather. ② keeps all 30 days, but the filled values are not real measurements, so the report must state the method.

## A.2 Chapter 2 Descriptive Statistics

### Check questions

2-1 **7.** The middle value of 3, 5, 7, 9, 11.

2-2 **The median.**

2-3 **0.** Every deviation is 0.

### Exercises

2-1 **Mean 6, median 6.** Sum 30 ÷ 5 = 6. In order: 4, 5, 6, 7, 8.

2-2 **0.233.** 7 ÷ 30 ≈ 0.233.

2-3 **Variance 4.571, standard deviation 2.138.** Mean 5; sum of squared deviations 9 + 1 + 1 + 1 + 0 + 0 + 4 + 16 = 32; 32 ÷ 7 ≈ 4.571; √4.571 ≈ 2.138.

2-4 **Mean 20, median 12. The median is the better choice.** The outlier 54 pulls the mean up, so four of the five households earn less than the mean.

2-5 **Mean 34, standard deviation about 5.48.** Adding a constant shifts the mean but does not change the spread. Multiplying by a constant scales both the mean and the standard deviation. (14 + 3) × 2 = 34, 2.74 × 2 = 5.48.

## A.3 Chapter 3 Correlation and Regression

### Check questions

3-1 **No.** The correlation coefficient measures only linear relationships. Even with a curved relationship, such as a U shape, $r$ can be close to 0.

3-2 **231 units.** Day 8, 32 °C.

3-3 **0.81.**

### Exercises

3-1 **A strong negative linear relationship.** As one variable increases, the other generally decreases, and the points lie close to a straight line.

3-2 **4.6.** 2.2 + 0.6 × 4 = 4.6.

3-3 **$\hat{y}$ = 2 + 0.8$x$, $r$ = 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 ÷ 20.

3-4 **About 297 units — but this is extrapolation, so it is hard to trust.** 8.716 × 40 − 51.54 ≈ 297.1. The data cover only 19–32 °C, and on extremely hot days people may stay indoors, changing the relationship.

3-5 **The common cause is temperature (the season).** On hot days more ice cream is sold and more people swim, so accidents rise. To establish causation you need data that control for the third variable, such as comparisons between days with the same temperature or experimental data in which sales were varied deliberately.

# Glossary

Terms defined in this book, in alphabetical order. The chapter where each term is first defined is given in parentheses.

| | |
|---|---|
| **coefficient of determination** | The share of variation explained by the regression line. (Chapter 3) |
| **correlation coefficient** | The direction and strength of a linear relationship, from −1 to 1. (Chapter 3) |
| **extrapolation** | Prediction outside the range of the data. (Chapter 3) |
| **frequency table** | A table counting the values in each class. (Chapter 2) |
| **histogram** | A frequency distribution drawn as touching bars. (Chapter 2) |
| **mean** | The sum divided by the count. (Chapter 2) |
| **median** | The middle value when the data are sorted. (Chapter 2) |
| **missing value** | A value that was not recorded; kept distinct from 0. (Chapter 1) |
| **observational unit** | Each individual thing that is measured. (Chapter 1) |
| **regression line** | The line with the smallest sum of squared residuals. (Chapter 3) |
| **scale of measurement** | Nominal, ordinal, interval or ratio. (Chapter 1) |
| **scatter plot** | A plot of two variables as coordinates. (Chapter 3) |
| **standard deviation** | The square root of the variance. (Chapter 2) |
| **variable** | A characteristic that can differ between observational units. (Chapter 1) |

| | |
|---|---|
| **variance** | The sum of squared deviations divided by $n - 1$. (Chapter 2) |

# References

The works below are **fictional examples** that show the reference format. Replace them with real sources in an actual manuscript.


### Books

Example, K. (2024). *Statistics for first-time learners (example)*. Example Press.

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

### Articles

Specimen, L., & Park, J. (2025). Comparing methods for handling missing observations (example). *Journal of Examples*, 12(3), 45–67.

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

# Index


**C**

coefficient of determination 24
correlation coefficient 21

**D**

data 2
degrees of freedom 16
drawings
  site plan 3

**E**

extrapolation 24

**F**

frequency distribution 12

**H**

histogram 12

**L**

least squares 22

**M**

mean 13
measurement error 7
median 13
missing value 6

**O**

observational unit 2
outlier 14

**P**

photographs
  observation site 2

**R**

regression line 22

**S**

scale of measurement 4
scatter plot 20
standard deviation 15
structured data 3

**U**

unit conversion 5
unstructured data 3

**V**

variable 2
variance 15

**Colophon**

[Example] Fundamentals of Data Science — textbook template sample. Trim 188 × 257 mm, type area 146 × 209 mm. Body text 9.5/16 pt in the theme's Latin text face; code in JetBrains Mono 8/12.5 pt (all fonts licensed under the SIL OFL 1.1).

Typeset from HTML with WeasyPrint. Numbering, the table of contents, the lists of tables, figures and drawings, the index and the links between questions and answers were generated by the build tool (tools/build.py). Photographs and drawings are synthetic material made with tools/make_assets.py. Build v1.0.0 · October 3, 2026.