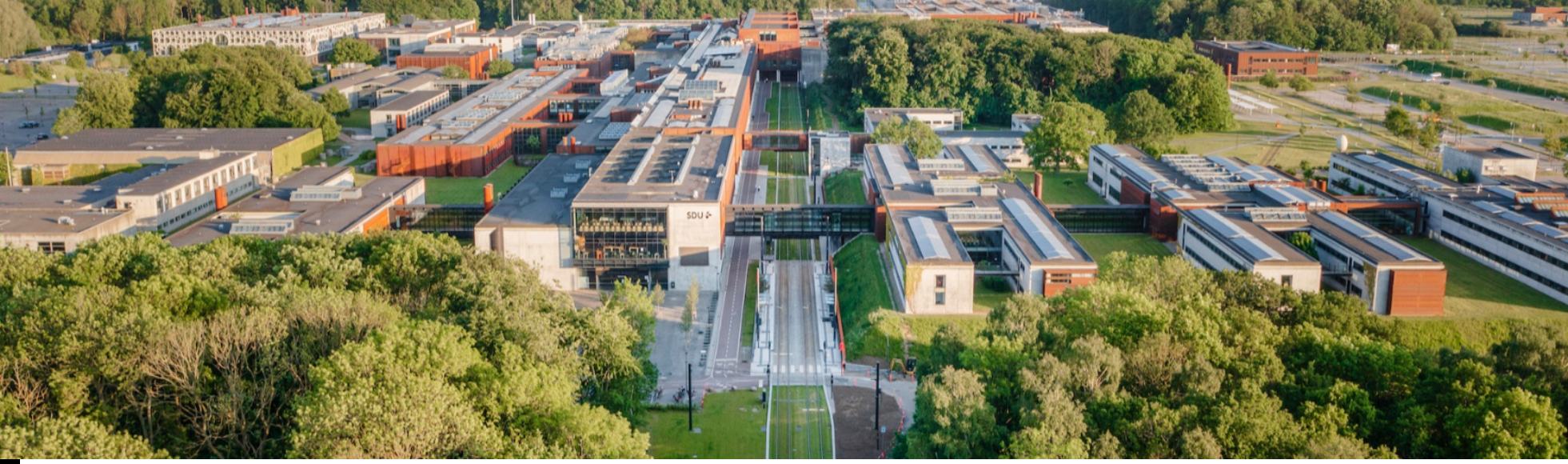


Mathematics Brush-Up for Data Science

Lecture Slides 16: Statistics

Nicklas S. Andersen

University of Southern Denmark (SDU)
Department of Mathematics & Computer Science (IMADA)

Updated: 2026-09-04 14:50 CEST

Statistics

Statistics is the discipline of:

- collecting
- analyzing
- interpreting
- presenting data

To then be able to:

- uncover patterns
- make predictions
- support decision-making under uncertainty

We will focus:

- on the vocabulary used
- sampling methods
- variable types
- numerical summaries
- graphs

Populations & Samples

- Definitions

The population is the entire group that a study aims to describe:

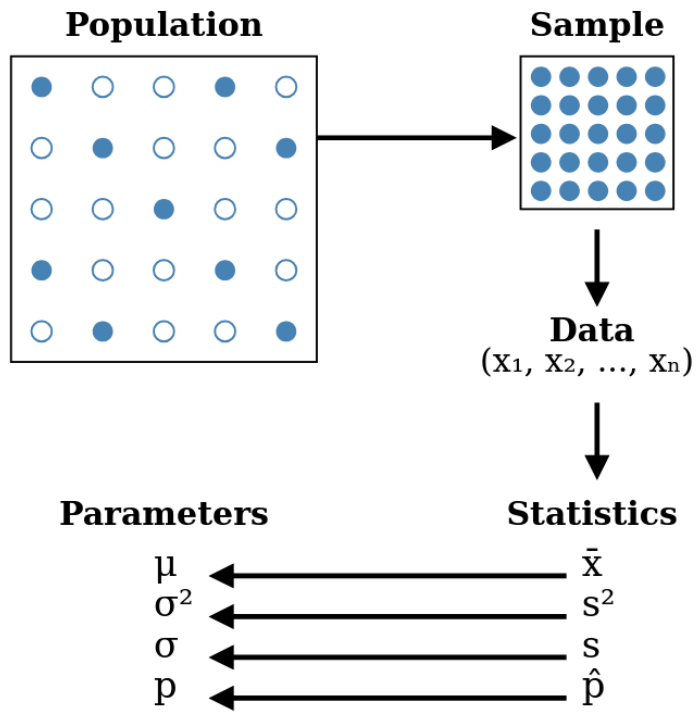
$$\mathcal{P} = \{u_1, u_2, \ldots, u_N\}, \qquad |\mathcal{P}| = N.$$

A sample is a smaller subset of the population. Its observed values are often written

$$x_1, x_2, \ldots, x_n,$$

where n is the sample size.

A representative sample lets us learn about the population without observing every unit.



Parameters & Statistics

- Definitions

A parameter is a fixed numerical characteristic of a population. For example,

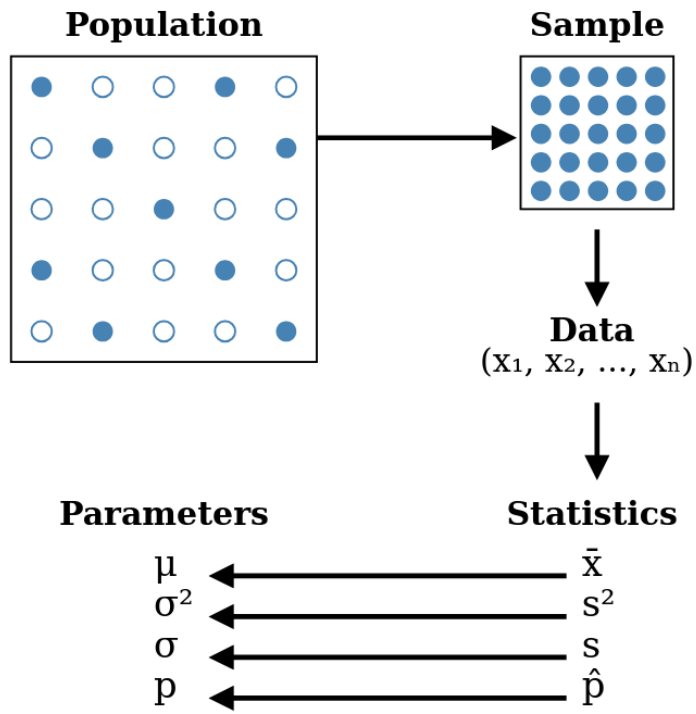
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is the population mean.

A statistic is calculated from sample data. For example,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is the sample mean and may estimate μ.



Statistical Notation

- Common Conventions

Statistical notation often distinguishes population quantities from sample-based quantities.

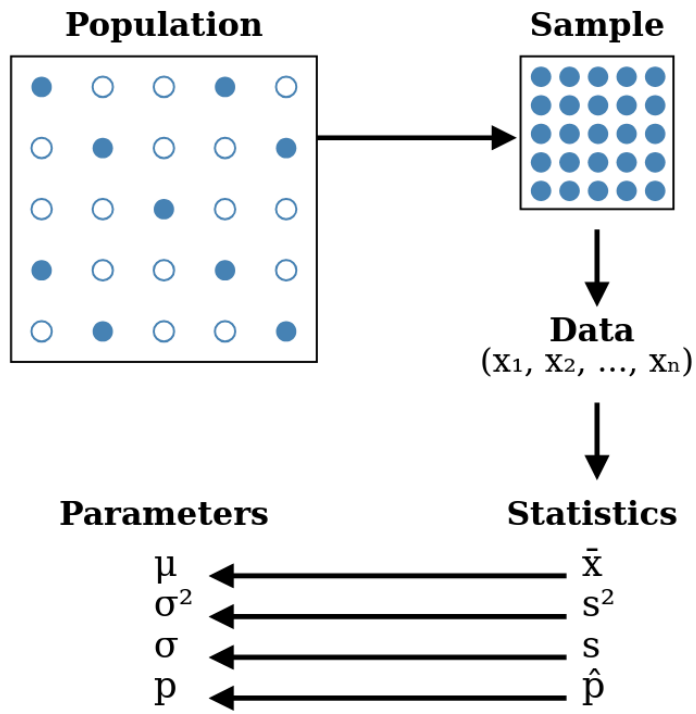
Sizes and symbols

- N denotes the population size
- n denotes the sample size
- Greek letters often denote population parameters
- Latin letters often denote sample statistics

Marks added to symbols

- A bar often denotes an arithmetic mean, as in $\overline{x}$
- A hat denotes an estimate: $\hat{\theta}$ estimates θ

In all cases, define each symbol in context.



Populations

- Examples

A population can be any complete group that we want to study or describe. Examples include:

People:

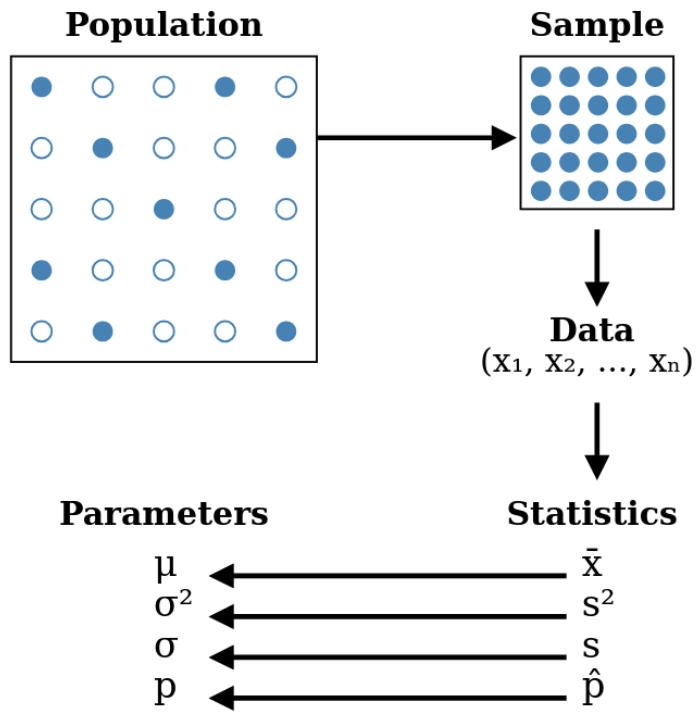
- Patients in a hospital
- Employees in a company

Objects or items:

- All cars produced by a factory in a year
- Books in a library

Events or measurements:

- Daily temperatures in a city over a decade
- Earthquake occurrences worldwide



Samples

- Sampling Methods

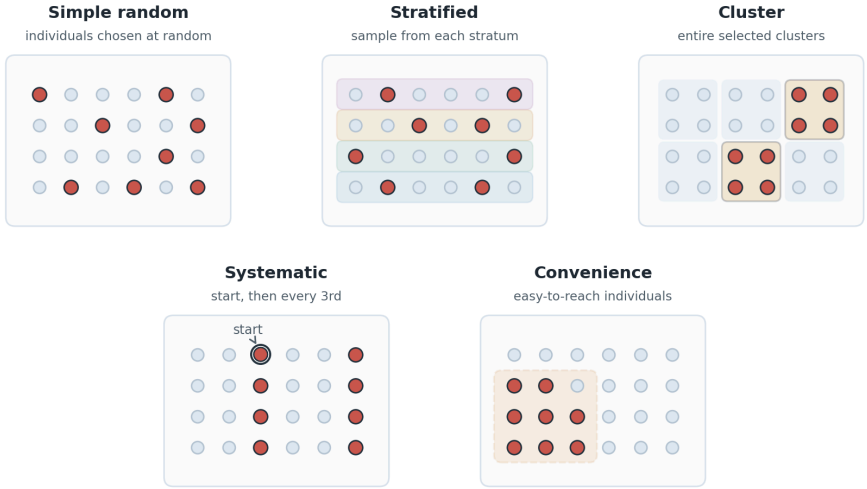
A representative sample reflects the characteristics of the population.

To achieve this, we use sampling methods, such as:

- Simple Random: Equal chance for all
- Stratified: Proportional sampling from groups
- Cluster: Random clusters, survey all members
- Systematic: Random start, every n-th member
- Convenience: Easy but typically biased

Furthermore:

- Studying the full population is impractical
- Well-designed samples support valid conclusions
- Proper methods reduce selection bias



Categorizing Data

- Data Points & Variables

To compute statistics, we start with a dataset: A collection of values called data points.

- Each data point represents an observation about an object or individual.
- These observations are organized into variables, the characteristics being measured or described.

Once we have gathered data, we can classify variables into two broad types:

- Qualitative: Descriptive or categorical values
- Quantitative: Numeric, measurable values

ID	Region	Product	Satisfaction	Purchases	Spending (EUR)
1	East	B	4	6	294
2	West	A	3	8	380
3	North	A	4	5	45
4	East	A	4	7	38
5	East	A	2	11	55
6	West	B	2	8	224
7	North	B	3	7	595
8	North	B	5	7	371
9	East	B	3	4	216
10	South	A	2	10	125
11	East	C	4	10	90
12	East	B	4	8	88
13	East	B	3	5	165
14	East	A	4	6	207
15	West	A	1	5	245

Categorizing Data

- Qualitative Variables

Qualitative variables describe categories or characteristics rather than measurable quantities:

- They cannot be added, subtracted, or averaged
- Best summarized using counts or percentages
- Typically visualized with: Bar charts

Examples from our dataset include:

- Region (North, South, East, West)
- Product (A, B, C)
- Satisfaction rating (1–5)

ID	Region	Product	Satisfaction	Purchases	Spending (EUR)
1	East	*B*	4	6	294
2	West	*A*	3	8	380
3	North	*A*	4	5	45
4	East	*A*	4	7	38
5	East	*A*	2	11	55
6	West	*B*	2	8	224
7	North	*B*	3	7	595
8	North	*B*	5	7	371
9	East	*B*	3	4	216
10	South	*A*	2	10	125
11	East	*C*	4	10	90
12	East	*B*	4	8	88
13	East	*B*	3	5	165
14	East	*A*	4	6	207
15	West	*A*	1	5	245

Categorizing Data

- Quantitative Variables

Quantitative variables measure numeric quantities:

- They can be added, subtracted, and averaged
- Summarized with means, totals, or ranges
- Typically visualized with: Histograms

Additionally, they may be:

- Discrete: take on separate, countable values
- Continuous: can take any value in an interval and are typically measured

Examples from our dataset include:

- Purchases: Discrete numeric
- Spending (EUR): Continuous numeric

ID	Region	Product	Satisfaction	Purchases	Spending (EUR)
1	East	*B*	4	6	294
2	West	*A*	3	8	380
3	North	*A*	4	5	45
4	East	*A*	4	7	38
5	East	*A*	2	11	55
6	West	*B*	2	8	224
7	North	*B*	3	7	595
8	North	*B*	5	7	371
9	East	*B*	3	4	216
10	South	*A*	2	10	125
11	East	*C*	4	10	90
12	East	*B*	4	8	88
13	East	*B*	3	5	165
14	East	*A*	4	6	207
15	West	*A*	1	5	245

Levels of Measurement

- Meaningful Comparisons

The levels of measurement describe what comparisons or calculations are meaningful for a variable.

Level	Natural order?	Meaningful differences?	Meaningful ratios?
Nominal	No	No	No
Ordinal	Yes	Not necessarily	No
Interval	Yes	Yes	No
Ratio	Yes	Yes	Yes

Examples from our dataset:

- Region: nominal
- Satisfaction rating: ordinal
- Calendar year: interval
- Purchases and spending: ratio

ID	Region	Product	Satisfaction	Purchases	Spending (EUR)
1	East	*B*	4	6	294
2	West	*A*	3	8	380
3	North	*A*	4	5	45
4	East	*A*	4	7	38
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Exercise Round 1

- Variable Types

For each variable described below:

- Determine whether it is qualitative or quantitative.
- If quantitative, state whether it is discrete or continuous.

1. Daily highest temperature in a city (°C)
2. Number of online orders made by a customer in a month
3. Favorite coffee type (latte, espresso, cappuccino, etc.)
4. Time taken to complete an online quiz (seconds)
5. Annual salary of employees in a company (EUR)
6. Brand of smartphone used by survey respondents
7. Number of books borrowed from a library each week
8. Eye color of students at a college

Presenting Data Graphically

- Frequency & Relative Frequency

Once data is collected, the next step is analyzing and summarizing it.

One of the most effective ways to do this is by using graphs:

- The type of graph we choose depends on the type of data (qualitative or quantitative)
- Before creating graphs, we often create a frequency distribution

In this context:

- Frequency: How many times a value (or group of values) occurs
- Relative frequency: The proportion of the dataset that each value (or group) represents

In particular, using the relative frequency allows comparison between different samples.

Presenting Data Graphically

- Example: Online Customer Dataset

Using EUR100 intervals, we calculate the frequency and relative frequency of the "Spending" column:

ID	Region	Product	Satisfaction	Purchases	Spending (EUR)
1	East	B	4	6	294
2	West	A	3	8	380
3	North	A	4	5	45
4	East	A	4	7	38
5	East	A	2	11	55
6	West	B	2	8	224
7	North	B	3	7	595
8	North	B	5	7	371
9	East	B	3	4	216
10	South	A	2	10	125
11	East	C	4	10	90
12	East	B	4	8	88
13	East	B	3	5	165
14	East	A	4	6	207
15	West	A	1	5	245

Frequency distribution (bins of EUR100):

Spending (EUR) Bin	Frequency	Values
0–99	5	38, 45, 55, 88, 90
100–199	2	125, 165
200–299	5	207, 216, 224, 245, 294
300–399	2	371, 380
400–499	0	– –
500–599	1	595
Total	15	

Relative frequencies:

Spending (EUR) Bin	Relative Frequency (%)
0–99	$\frac{5}{15} = 33.3\%$
100–199	$\frac{2}{15} = 13.3\%$
200–299	$\frac{5}{15} = 33.3\%$
300–399	$\frac{2}{15} = 13.3\%$
400–499	$\frac{0}{15} = 0.0\%$
500–599	$\frac{1}{15} = 6.7\%$
Total	100%

Presenting Data Graphically

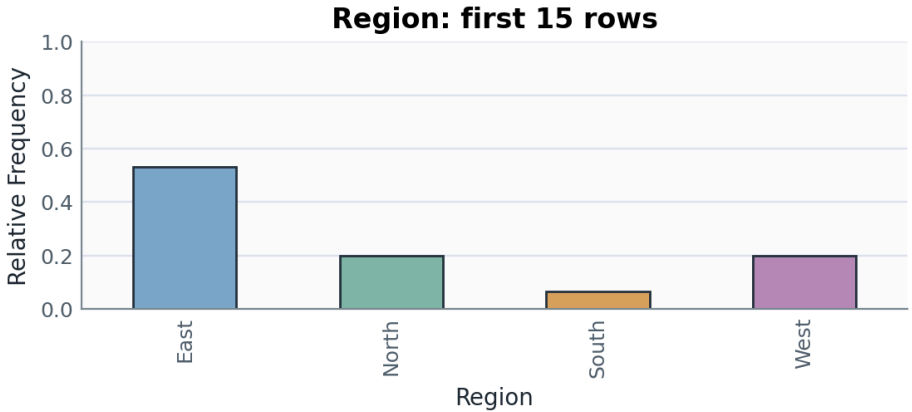
- Bar Charts

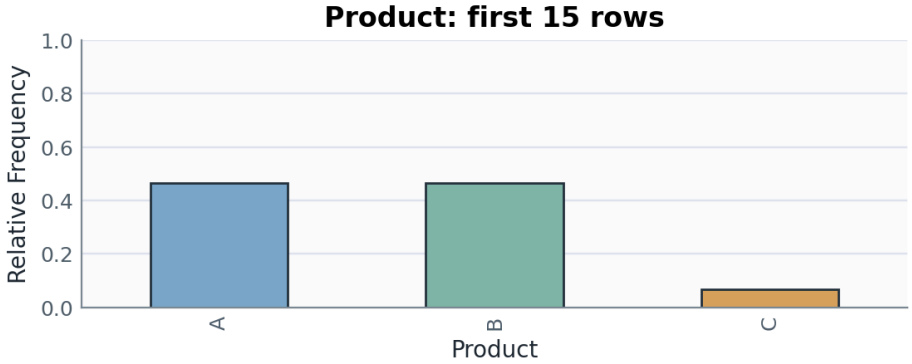
Bar charts are used to summarize qualitative data:

- Each bar represents a category
- The length of a bar shows the:
  - frequency or;
  - relative frequency

Examples with our dataset:

- Region: North, South, East, West
- Product: A, B, C





Presenting Data Graphically

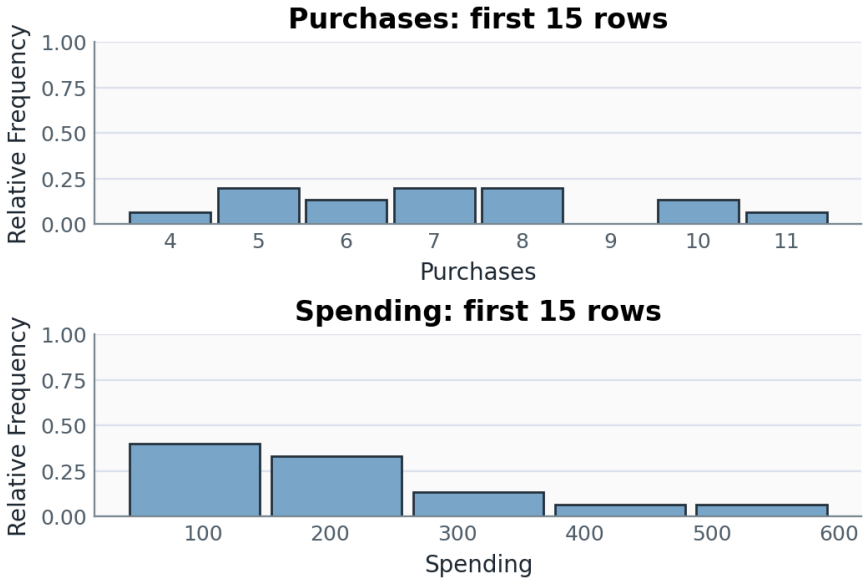
- Histograms

Histograms are used for quantitative data:

- Histograms group numerical data into ranges (bins)
- They show how many data points fall in each range

Examples with our dataset:

- Purchases: Discrete numeric
- Spending (EUR): Continuous numeric



Presenting Data Graphically

- Sample Size & Bin Width

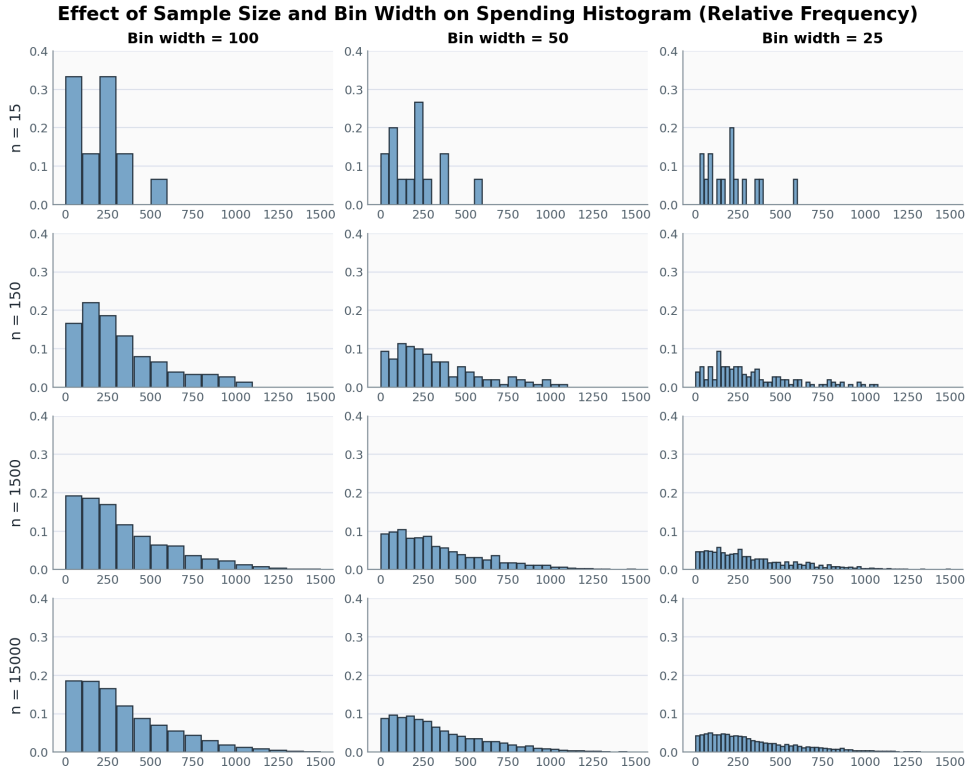
The look of a histogram depends on two key factors:

Sample size (n):

- Small n: Unstable and less detail
- Larger n: Smoother shape
- The Law of Large Numbers: Relative frequencies converges to true probabilities as $n \to \infty$

Bin width:

- Wide bins: Result in fewer bars and less detail
- Narrow bins: Result in more detail, but can look noisy, if the sample (of size n) is small
- Choose a bin width that balances detail and clarity



Exercise Round 2

- Frequency Distributions

Using the dataset of commute times (rounded to the nearest minute) for 15 students:

ID	Commute Time (minutes)
1	8
2	15
3	12
4	25
5	5
6	30
7	20
8	14
9	10
10	18
11	22
12	6
13	28
14	16
15	13

1. Create a frequency table using the 10-minute intervals

$$0–9,$$
$$10–19,$$
$$20–29,$$
$$30–39.$$

2. Calculate the relative frequency for each interval.

3. Identify which interval contains the highest proportion of students.

Measures of Central Tendency

- Building Intuition

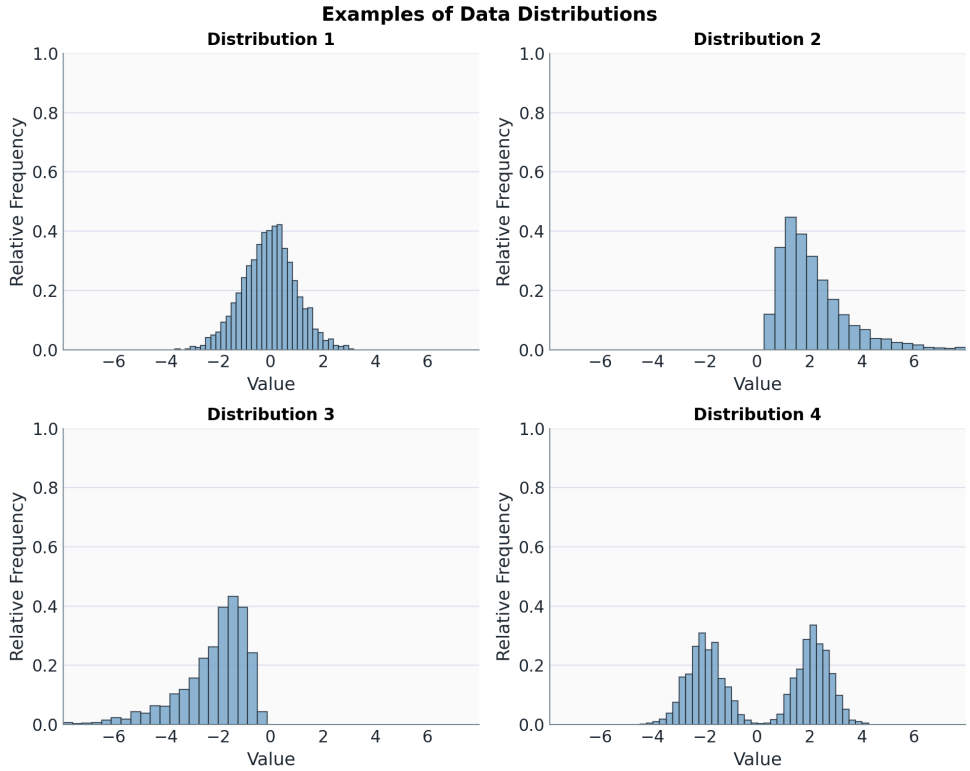
Quantitative data can be described not only verbally and graphically, but also with numbers.

When summarizing a distribution, we want to know:

- A representative value (some central value)
- how spread out the data values are

In the following, we thus:

- focus on measures of central tendency
- cover measures of spread subsequently



The Arithmetic Mean

- Definition

The arithmetic mean, or simply the mean, is found by dividing the sum of the data values by the number of values:

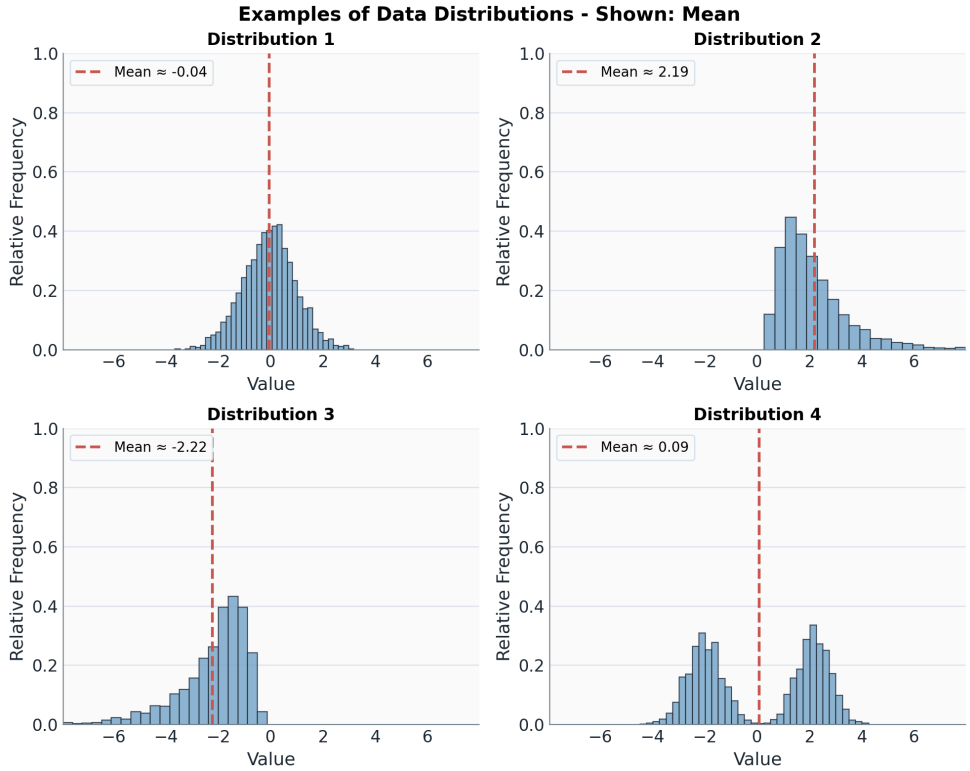
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where:

- x_i is the i-th data value
- n is the number of data values

Furthermore:

- The symbol $\overline{x}$ represents the mean
- x represents a single data value



The Arithmetic Mean

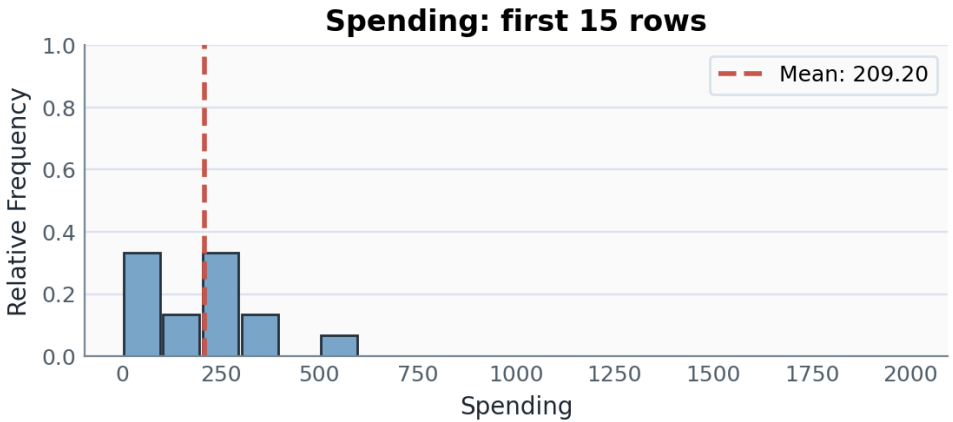
- Example: Online Customer Dataset

Suppose we want the average amount spent by the first 15 customers.

To do so, we sum the values in the "Spending" column and divide by 15:

$$\begin{aligned}\overline{x} &= \frac{294 + 380 + 45 + \cdots + 245}{15} \\ &\approx 209.2\end{aligned}$$

We can therefore conclude that the mean spending is about 209.2 euros.



Outliers

- Example: Online Customer Dataset

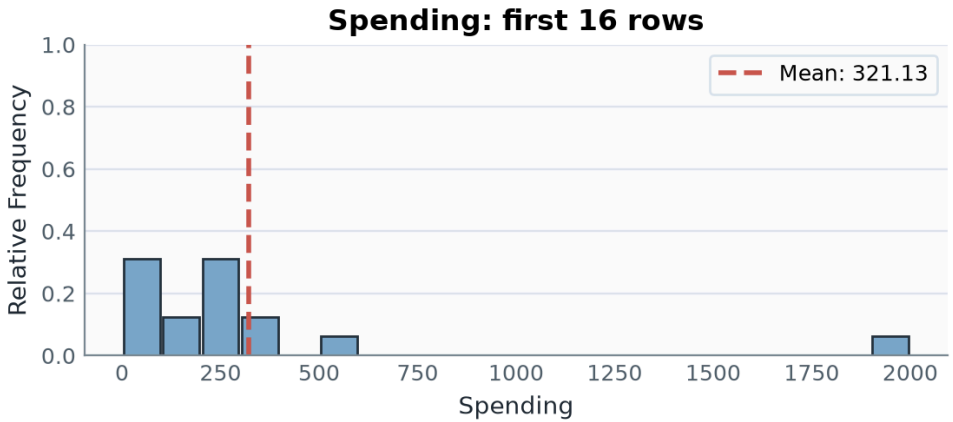
Now suppose a new customer spends 2000 euros. Including this value, the mean becomes:

$$\overline{x} = \frac{294 + 380 + 45 + \cdots + 245 + 2000}{16}$$
$$= \frac{5138}{16} \approx 321.13$$

While about 321.13 euros is mathematically correct, it no longer represents a typical value.

A value much higher or lower than the rest is called an outlier. Outliers may signal unusual but valid behavior or data entry errors.

When outliers are present, the mean is pulled toward them. In such cases, another measure of center, the median, is often more useful.



The Median

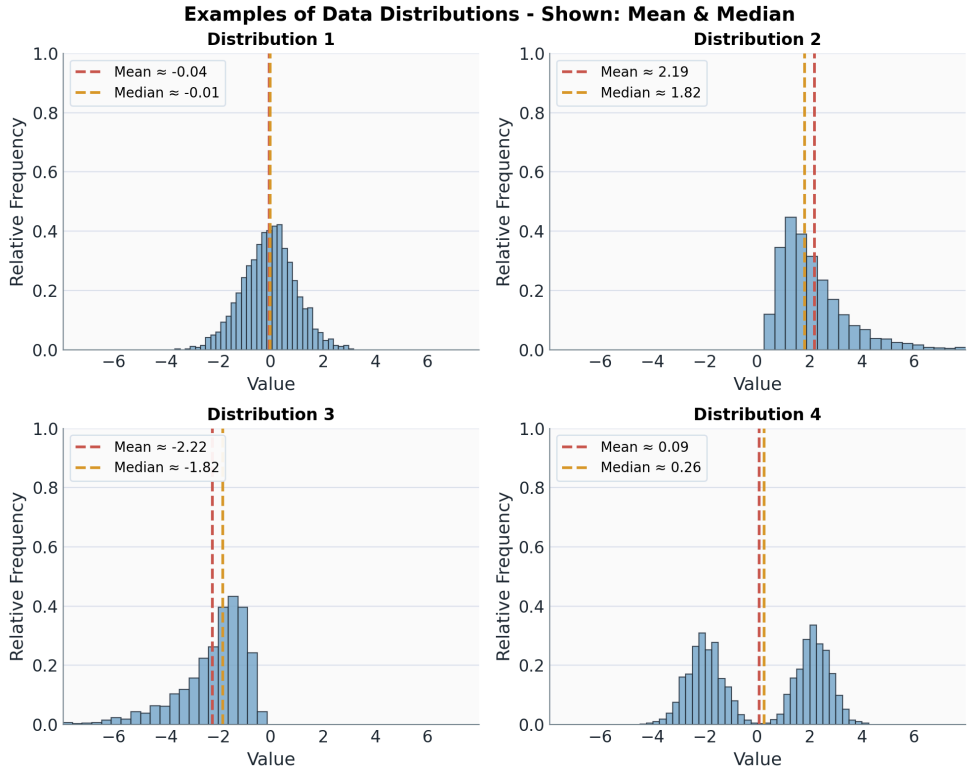
- Definition

The median is the middle value of an ordered data set.

To find the median:

- Order the data from smallest to largest.
- If there is an odd number of values: The middle value is the median
- If there is an even number of values: The median is the mean of the two middle values

There is no special formula or symbol for the median, but we denote the median by m, in the following.



The Median

- Example: Online Customer Dataset

Find the median of the "Spending" amount:

$$\text{”Spending”:} \quad 294, 380, 45, 38, 55, 224, 595, 371, 216, 125, 90, 88, 165, 207, 245, 2000$$

$$\text{sorted ”Spending”:} \quad 38, 45, 55, 88, 90, 125, 165, 207, 216, 224, 245, 294, 371, 380, 595, 2000$$

There are 16 data values (an even number), so the median is the mean of the two middle values:

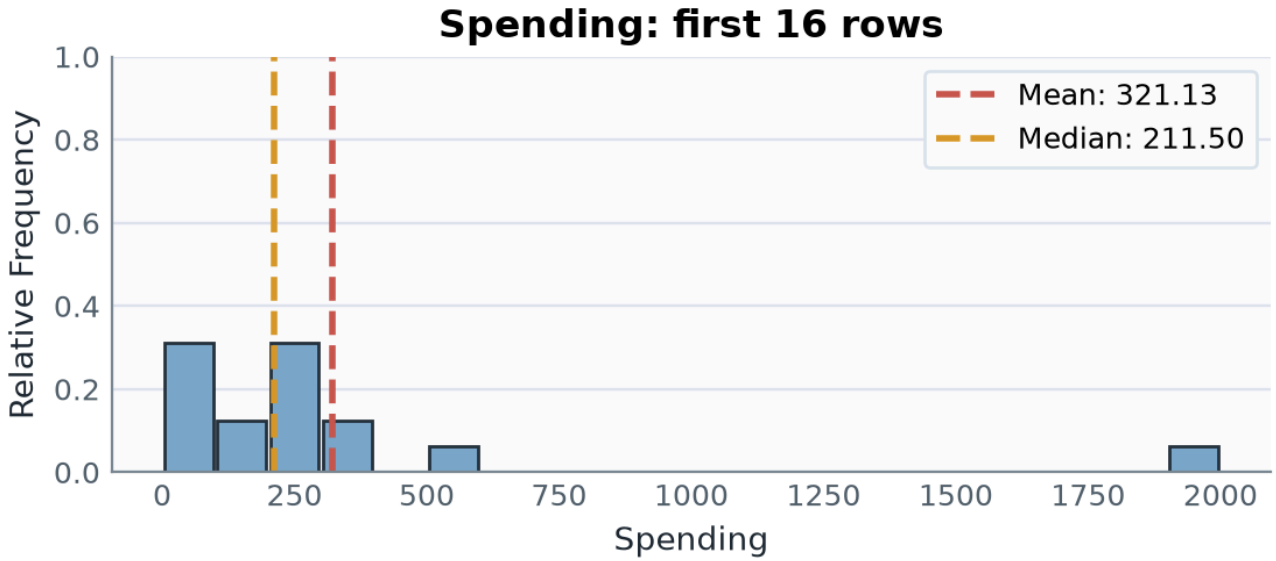
$$\underbrace{38, 45, 55, 88, 90, 125, 165,}_{\text{lower half}} \ \underbrace{207, 216}_{\text{middle}} \ \underbrace{224, 245, 294, 371, 380, 595, 2000}_{\text{upper half}}$$

The median is therefore:

$$m = \frac{207 + 216}{2} = 211.5$$

The Median

- Example: Online Customer Dataset (Continued)



The Mode

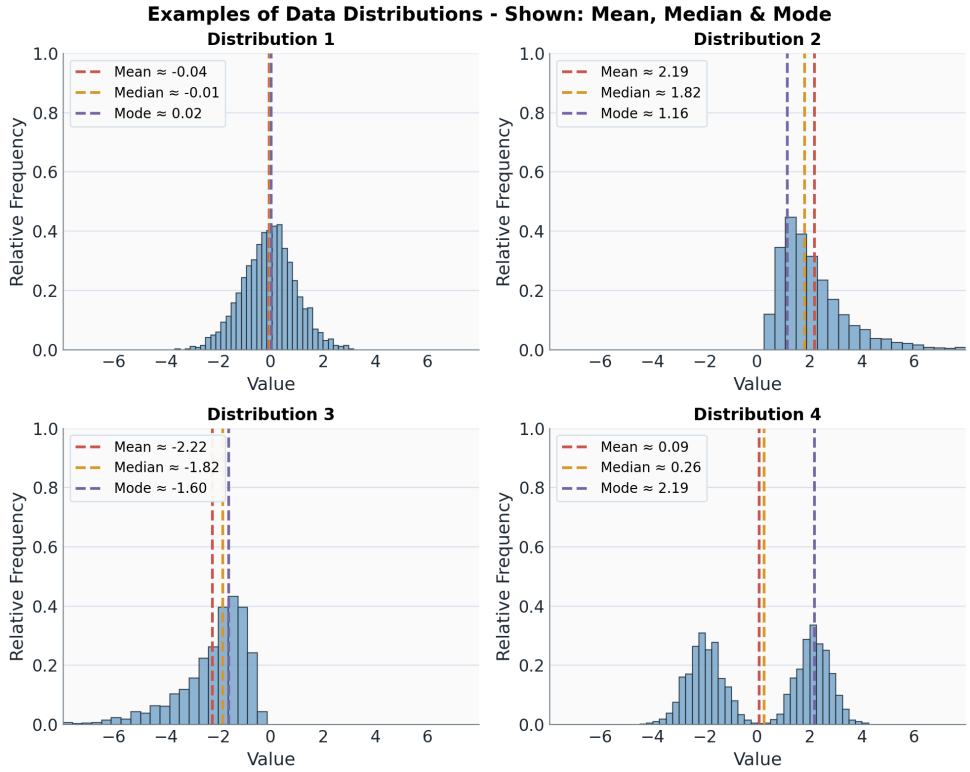
- Definition

The mode is the data value or category that occurs most frequently in a dataset.

A dataset can have:

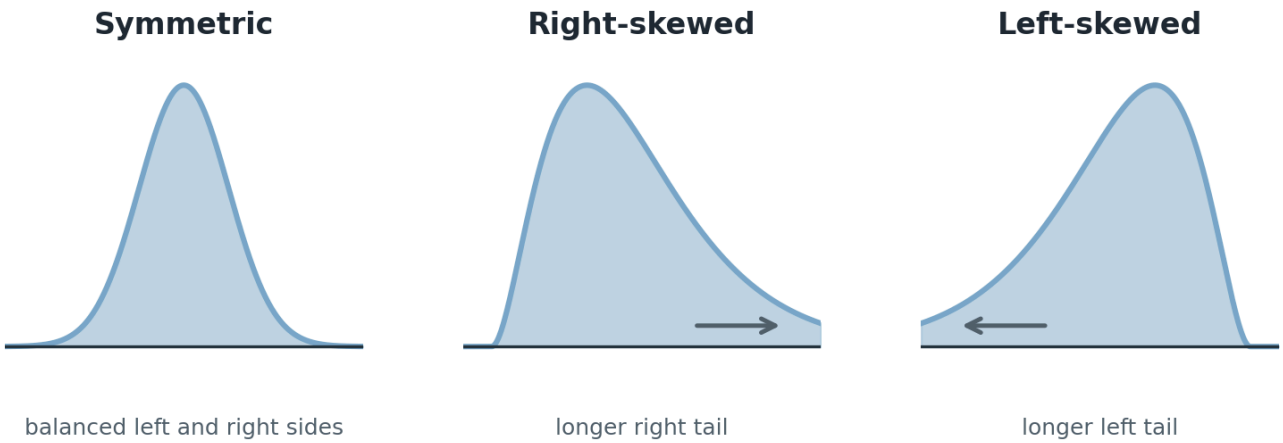
- no mode if all values occur equally often
- one mode (unimodal)
- two modes (bimodal)
- more than two modes (multimodal)

For grouped data, the bin or interval with the highest frequency is called the modal class.



Distribution Shape

- Symmetry & Skewness



- Symmetric: the two sides are roughly balanced
- Right-skewed: the longer tail extends toward larger values
- Left-skewed: the longer tail extends toward smaller values

The mean is pulled toward the longer tail; the median is usually less affected by extreme values.

Exercise Round 3

- Measures of Central Tendency

For the dataset

$$2, 3, 3, 4, 5, 7, 18.$$

1. Find the mean, median, and mode.

2. Which measure best describes a typical value? Explain.

3. Remove the value 18 and recompute the mean and median.

4. Which measure changes more when the outlier is removed?

Measures of Spread

- Building Intuition

Consider three lists of quiz scores (10-point quiz):

- Class A: $5, 5, 5, 5, 5, 5, 5, 5, 5, 5$
- Class B: $0, 0, 0, 0, 0, 10, 10, 10, 10, 10$
- Class C: $4, 4, 4, 5, 5, 5, 5, 6, 6, 6$

All three classes have a mean of 5 and a median of 5, yet the distributions are very different:

- Class A: No variation. Same values
- Class B: Extreme variation. Scores are split
- Class C: Moderate variation. Scores cluster around 5

This shows that, in addition to measures of center (mean, median), we also need measures of spread to describe data variation.

In the following, we will look at:

- Range
- Standard deviation
- Percentiles, quartiles, and interquartile range (IQR)
- Box plots as a graphical summary of spread

Range

- Definition

The range is the simplest way to measure spread.

It uses only two values:

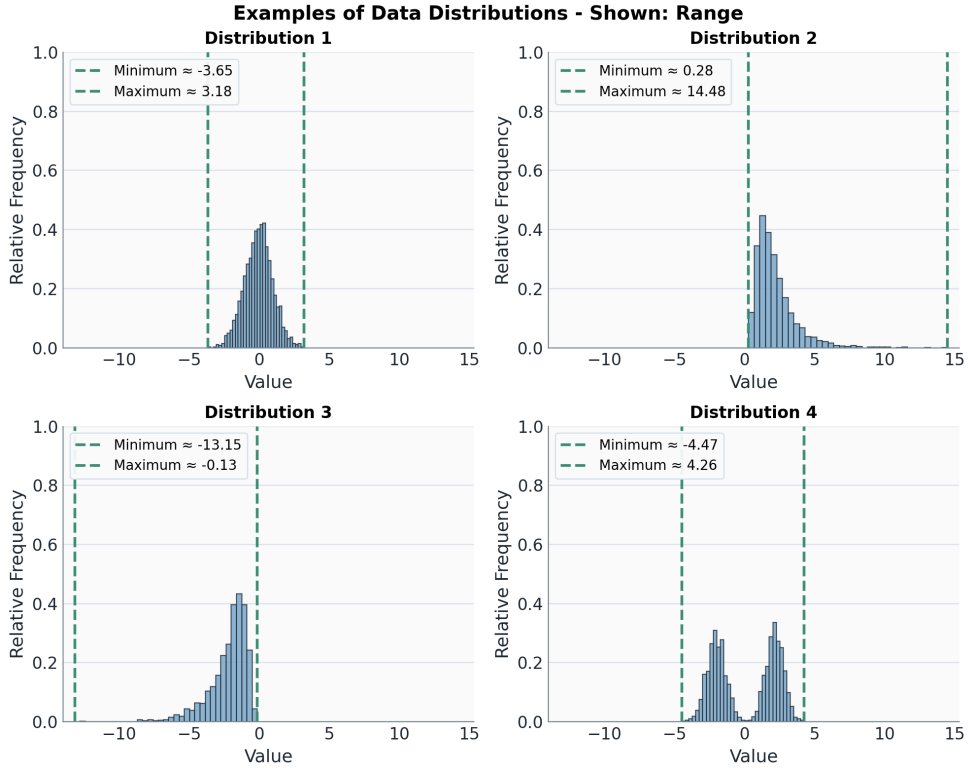
- The largest value (maximum)
- The smallest value (minimum)

The range is calculated as:

$$r = x_{\max} - x_{\min}$$

where:

- $x_{\max}$ is the largest value
- $x_{\min}$ is the smallest value



Range

- Example: Online Customer Dataset

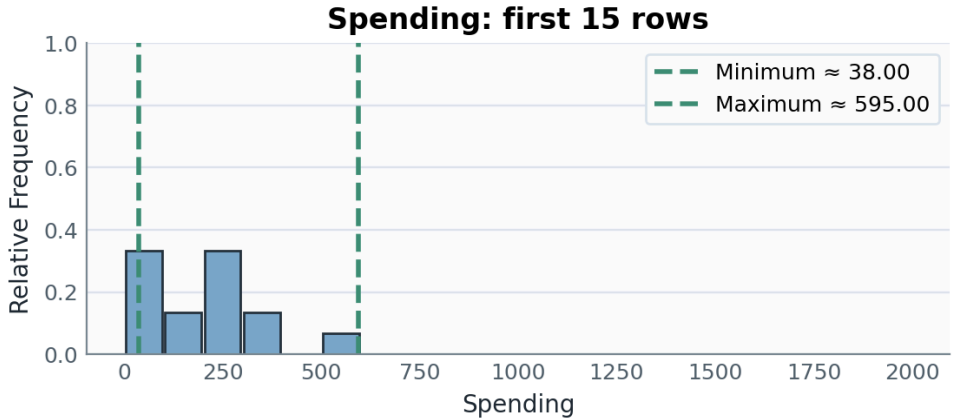
Consider the spending data for 15 customers:

$$38,\ 45,\ 55,\ 88,\ 90,\ 125,\ 165,\ 207,$$
$$216,\ 224,\ 245,\ 294,\ 371,\ 380,\ 595$$

- The minimum spending value ($x_{\min}$) is 38
- The maximum spending value ($x_{\max}$) is 595

The range is calculated as:

$$r = x_{\max} - x_{\min} = 595 - 38 = 557$$



Range

- Examples

Find the range for each dataset:

- X: $10, 20, 30, 40, 50$
- Y: $10, 35, 36, 37, 50$

For both datasets:

$$\text{Range} = 50 - 10 = 40$$

The ranges are the same, but the distributions are not:

- X: Values are evenly spread out
- Y: Values are close together except two extremes

Key observations:

- The range only considers the extremes
- The range ignores all the values in between

Consider the quiz score distributions from earlier:

- Class A: Range = $5 - 5 = 0$
- Class B: Range = $10 - 0 = 10$
- Class C: Range = $6 - 4 = 2$

Suppose we add Class D's quiz scores:

$$0, 5, 5, 5, 5, 5, 5, 5, 5, 10$$
$$\Rightarrow \text{Range} = 10 - 0 = 10$$

Notice again that Class D has the same range as Class B, but the distributions are very different:

- Class B: $0, 0, 0, 0, 0, 10, 10, 10, 10, 10$

This shows why we need more sophisticated measures to better describe variation.

Deviation

- Definition & Example

The difference between a data value x_i and the mean $\overline{x}$ is called the deviation d from the mean:

$$d = x_i - \overline{x}$$

Key points:

- Positive deviations indicate values above the mean
- Negative deviations indicate values below the mean
- The sum of all deviations is always zero (apart from small rounding errors)

This does not mean that data values are zero distance from the mean. It simply reflects that positive and negative deviations cancel each other out.

Spending (EUR)	x_i	$-\overline{x}$	$= d$
294	294	−209.2	84.8
380	380	−209.2	170.8
45	45	−209.2	−164.2
38	38	−209.2	−171.2
55	55	−209.2	−154.2
224	224	−209.2	14.8
595	595	−209.2	385.8
371	371	−209.2	161.8
216	216	−209.2	6.8
125	125	−209.2	−84.2
90	90	−209.2	−119.2
88	88	−209.2	−121.2
165	165	−209.2	−44.2
207	207	−209.2	−2.2
245	245	−209.2	35.8
Sum			≈ 0

Variance

- Definition & Example

Because positive and negative deviations cancel, we square the deviations before combining them:

- Squared deviations are nonnegative
- Larger deviations have a greater effect

The sample variance is the sum of the squared deviations divided by $n - 1$:

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

For this dataset:

$$s^2 = \frac{330{,}926.40}{15 - 1} = \frac{330{,}926.40}{14} \approx 23{,}637.60$$

Spending (EUR)	x_i	$-\overline{x}$	$= d$	d^2
294	294	−209.2	84.8	7,191.04
380	380	−209.2	170.8	29,172.64
45	45	−209.2	−164.2	26,961.64
38	38	−209.2	−171.2	29,309.44
55	55	−209.2	−154.2	23,777.64
224	224	−209.2	14.8	219.04
595	595	−209.2	385.8	148,841.64
371	371	−209.2	161.8	26,179.24
216	216	−209.2	6.8	46.24
125	125	−209.2	−84.2	7,089.64
90	90	−209.2	−119.2	14,208.64
88	88	−209.2	−121.2	14,689.44
165	165	−209.2	−44.2	1,953.64
207	207	−209.2	−2.2	4.84
245	245	−209.2	35.8	1,281.64
Sum			≈ 0	330,926.40

Standard Deviation

- Definition

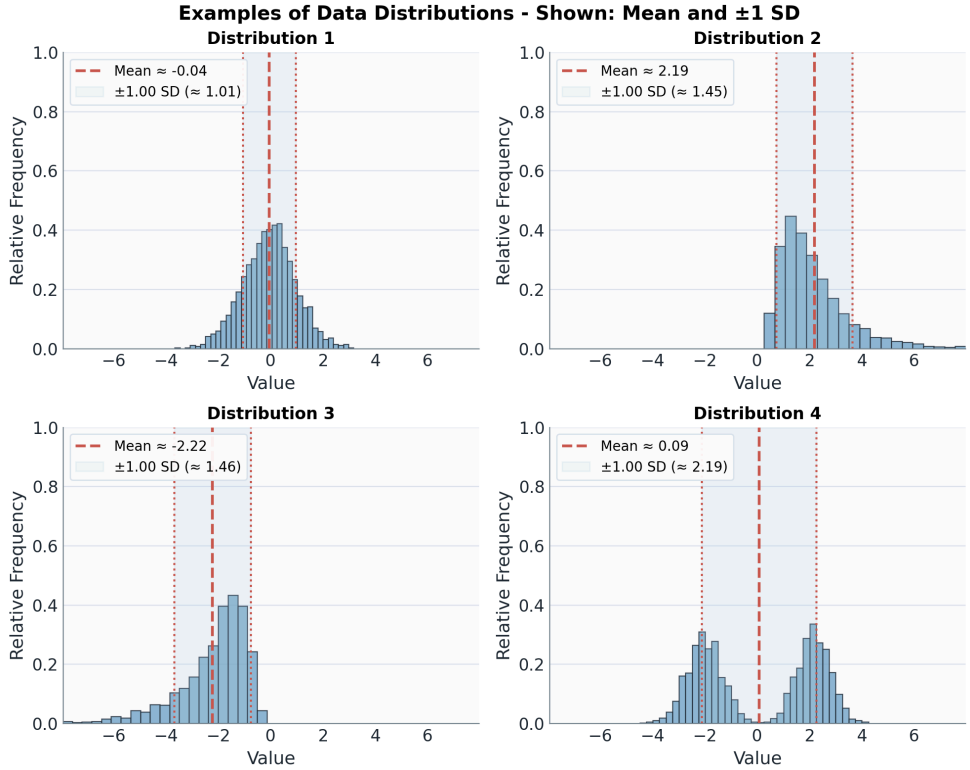
Variance is in squared units, which makes it less intuitive. On the other hand, the standard deviation expresses spread in the same units as the data:

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

- s measures typical distance from $\overline{x}$
- larger values of s indicate greater variability

Steps to compute s:

1. Find each deviation from the mean
2. Square the deviations
3. Sum the squares
4. Divide by $n-1$ (variance)
5. Take the square root

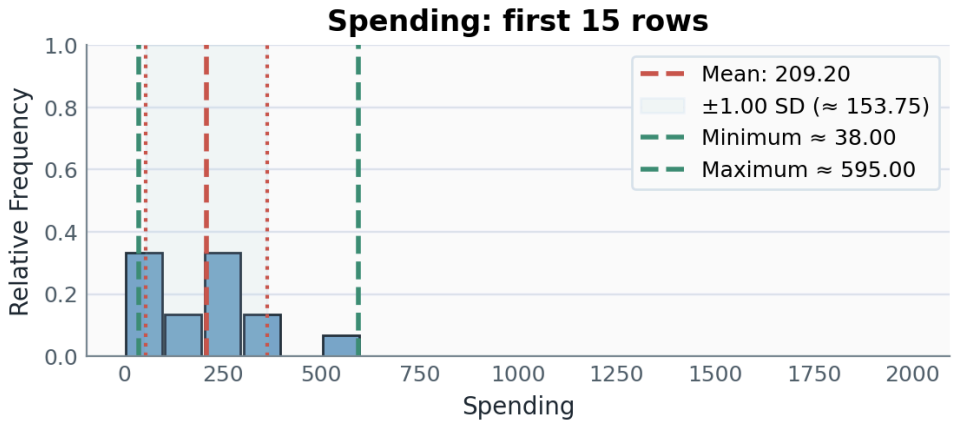


Standard Deviation

- Example: Online Customer Dataset

For the customer dataset the standard deviation of the spending is:

$$s = \sqrt{23{,}637.60} \approx 153.75$$



Exercise Round 4

- Range, Variance & Standard Deviation

For the sample

$$2, 4, 4, 6.$$

1. Find:

a. the range
b. the sample variance
c. the sample standard deviation

Percentiles

- Definition

A percentile describes relative position within an ordered data set.

The k-th percentile is a value that places approximately $k\%$ of the ordered data at or below it.

Quartiles are common percentiles that conceptually divide the ordered data into four approximately equal parts:

- Q_1: 25th percentile
- Q_2: 50th percentile (median)
- Q_3: 75th percentile

Together with the minimum and maximum, quartiles form the five-number summary, a more comprehensive view of data spread.

How to find quartiles:

1. Order the data
2. Find the median (Q_2)
3. Median of the lower half (Q_1)
4. Median of the upper half (Q_3)

The Five-Number Summary & IQR

- Definition

The quartiles conceptually divide the ordered data into four approximately equal parts:

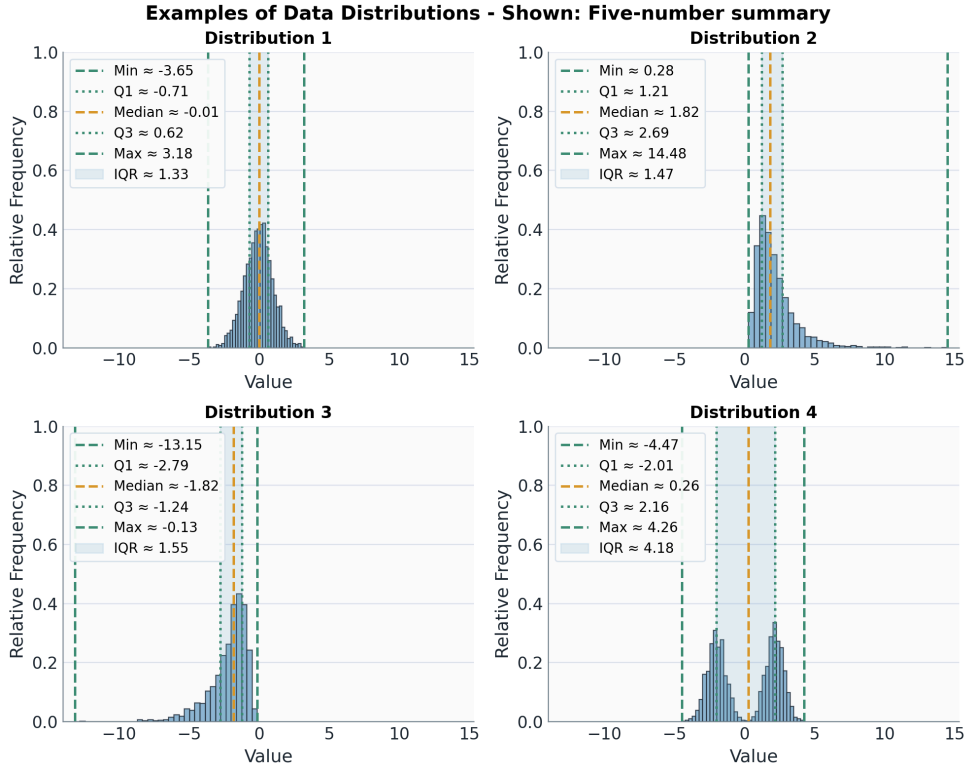
- Approximately 25% between Minimum and Q_1
- Approximately 25% between Q_1 and Median
- Approximately 25% between Median and Q_3
- Approximately 25% between Q_3 and Maximum

The middle 50% of the data lies between Q_1 and Q_3.

The interquartile range (IQR) measures this spread:

$$IQR = Q_3 - Q_1$$

A larger IQR indicates more variability in the central half of the data.



The Five-Number Summary & IQR

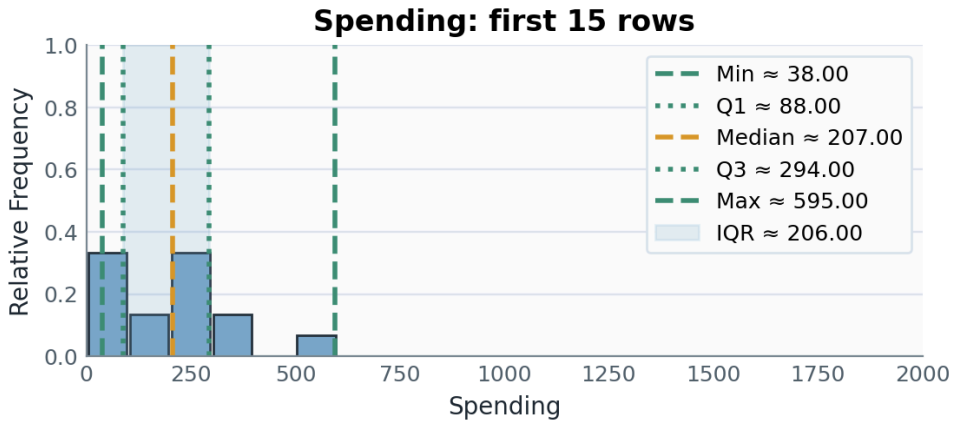
- Example: Online Customer Dataset

The following are the sorted values from our dataset's "Spending" column:

$$38, 45, 55, 88, 90, 125, 165,$$
$$207,$$
$$216, 224, 245, 294, 371, 380, 595$$

Breaking this into the lower half, median, and upper half gives us:

- Minimum: 38 (smallest value)
- First Quartile (Q_1): 88 (median of lower half)
- Median (Q_2): 207 (middle value)
- Third Quartile (Q_3): 294 (median of upper half)
- Maximum: 595 (largest value)
- IQR $Q_3 - Q_1 = 294 - 88 = 206$



Box-and-Whisker Plots

- Five-Number Summaries & Outliers

A box plot, or box-and-whisker plot, is a graphical representation of the five-number summary, that marks potential outliers.

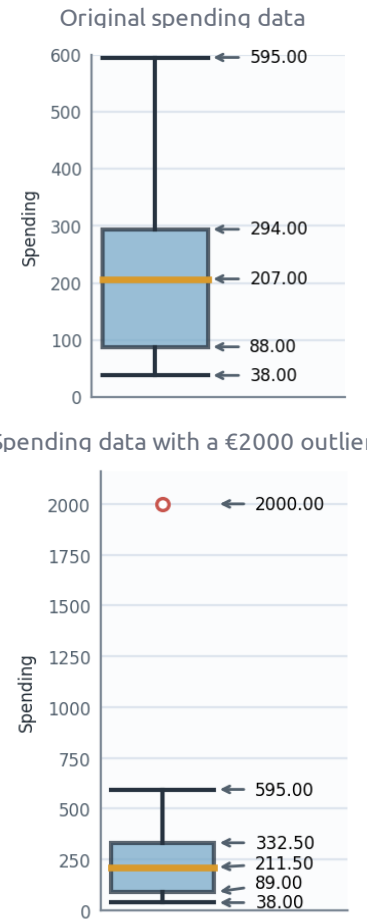
Its key parts are:

- The box spans from Q_1 to Q_3
- A vertical line inside the box marks the median (Q_2)
- Whiskers reach the most extreme non-outliers
- Separate points mark potential outliers

Potential outliers satisfy the $1.5\,IQR$ rule:

$$x < Q_1 - 1.5\,IQR \quad \text{or} \quad x > Q_3 + 1.5\,IQR.$$

Thus, a box plot summarizes center, spread, skewness, and potential outliers.



Exercise Round 5

- Spread, Position & Outliers

For the ordered dataset

$$12, 14, 15, 16, 16, 17, 18, 19, 20, 55.$$

1. Find the five-number summary and IQR using the median-of-halves method.

2. Use the $1.5 \cdot IQR$ rule to identify potential outliers.

3. Is the data symmetrically distributed or skewed?

4. Compare the range and IQR. What does each reveal about the spread of the data?