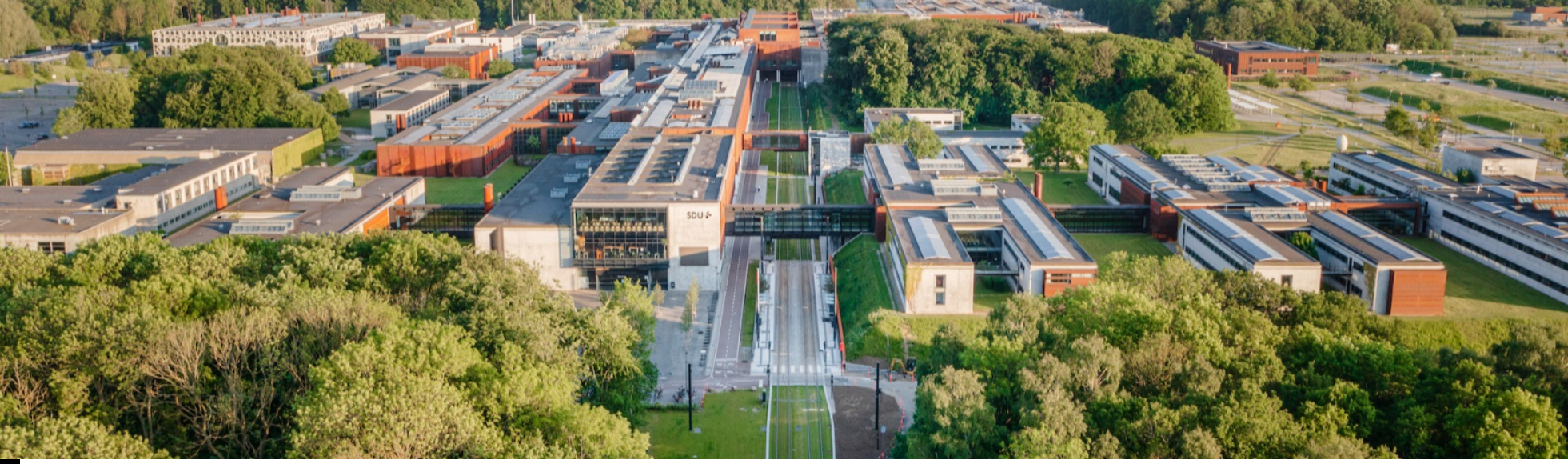


Brush-Up Maths for Data Science (2025)

Lecture Slides, Aug. 31th

Nicklas S. Andersen

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

Statistics

Statistics is the discipline of:

- collecting
- analyzing
- interpreting
- presenting data

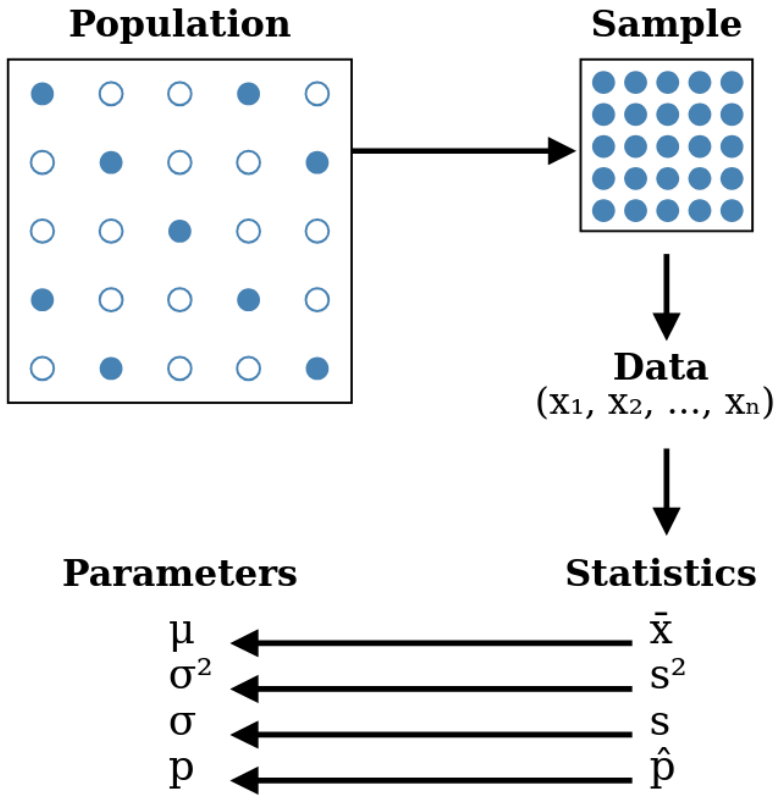
To then be able to:

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

Statistics

In this context, we are typically concerned with:

- A **Population**: It is the entire group of interest that the data aims to describe
- A **sample**: A subset of the **population**, ideally representative of the whole
- A **Parameter**: A true (but often unknown) value that describes the **population** (e.g. the mean μ)
- A **statistic**: A value calculated from the **sample** used to **estimate** the corresponding population parameter (e.g., the sample mean $\bar{x} \approx \mu$)



Populations

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

People:

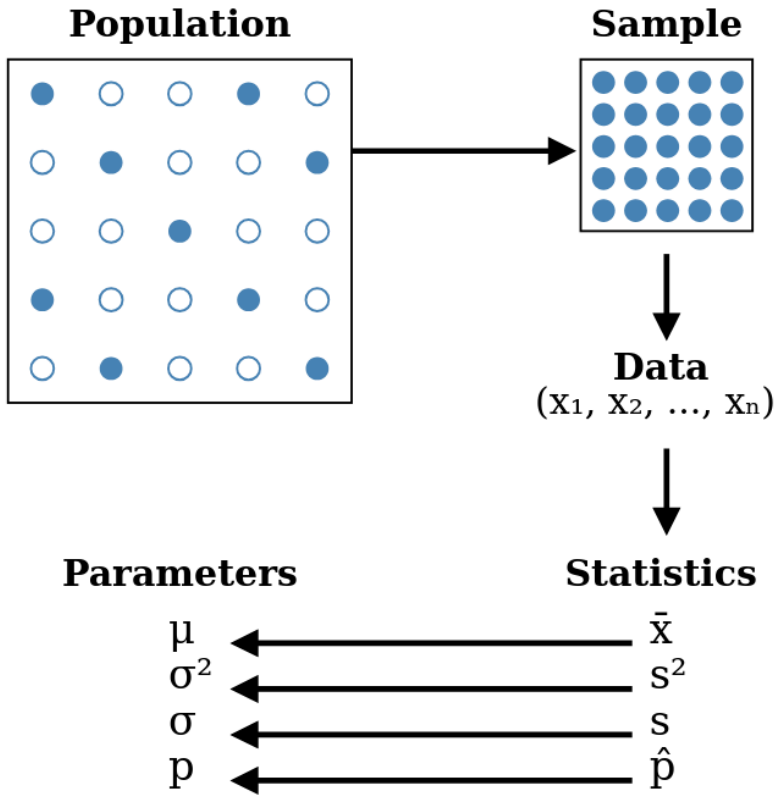
- Patients in a hospital
- Employees in a company

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

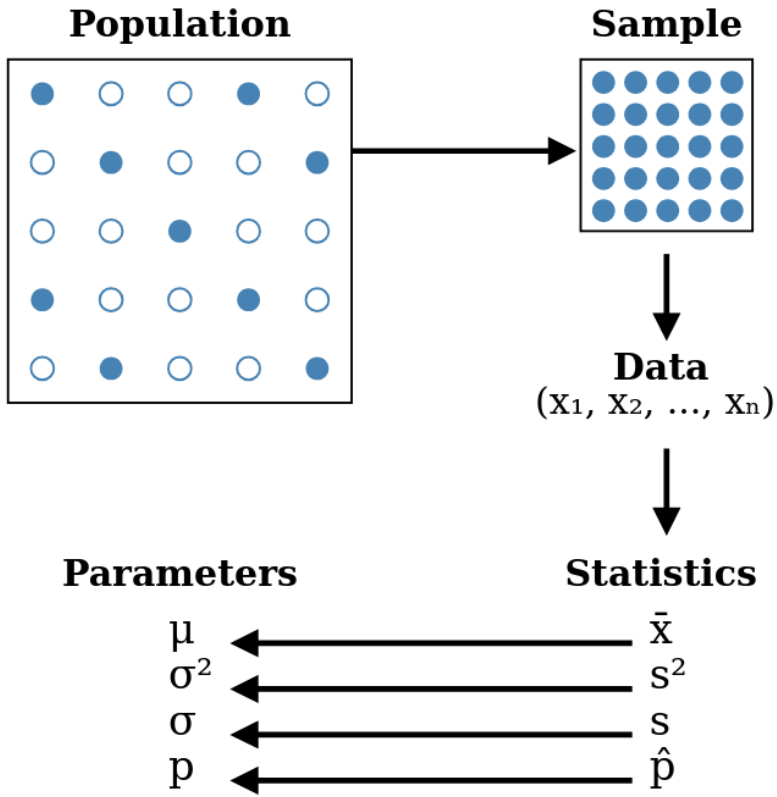
A representative sample reflects the characteristics of the population.

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

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 (€)
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)

ID	Region	Product	Satisfaction	Purchases	Spending (€)
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
- Typicalls visualized with: Histograms

Examples from our dataset include:

- Satisfaction (1-5): Discrete numeric
- Purchases: Discrete numeric
- Spending (€): Continuous numeric

ID	Region	Product	Satisfaction	Purchases	Spending (€)
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

Presenting Data Graphically

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 €100 intervals, we calculate the frequency and relative frequency of the "Spending" column:

ID	Region	Product	Satisfaction	Purchases	Spending (€)
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 €100):

Spending (€) Bin	Frequency	Values
0–99	5	38, 45, 55, 88, 90
100–199	2	125, 165
200–299	4	207, 216, 224, 245
300–399	3	294, 371, 380
400–499	0	– –
500–599	1	595
Total	15	

Relative frequencies:

Spending (€) Bin	Relative Frequency (%)
0–99	$\frac{5}{15} = 33.3\%$
100–199	$\frac{2}{15} = 13.3\%$
200–299	$\frac{4}{15} = 26.7\%$
300–399	$\frac{3}{15} = 20.0\%$
400–499	$\frac{0}{15} = 0.0\%$
500–599	$\frac{1}{15} = 6.7\%$
Total	100.0%

Presenting Data Graphically

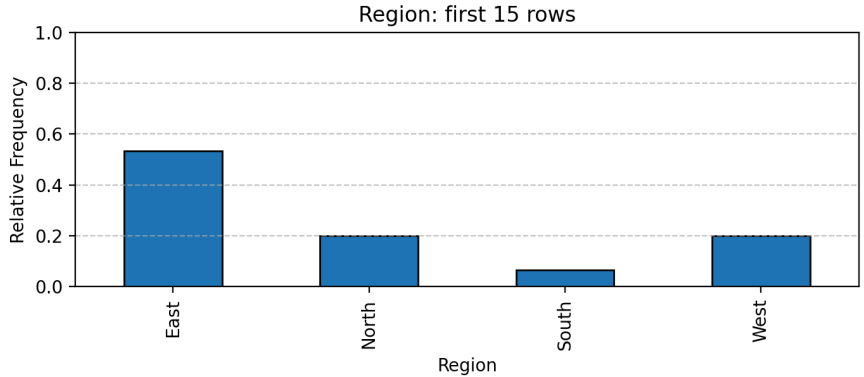
- Bar Charts

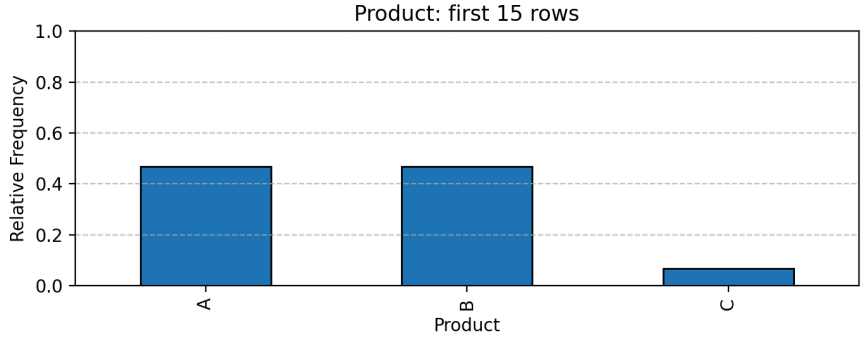
Bar charts are used to summarize qualitative data:

- 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

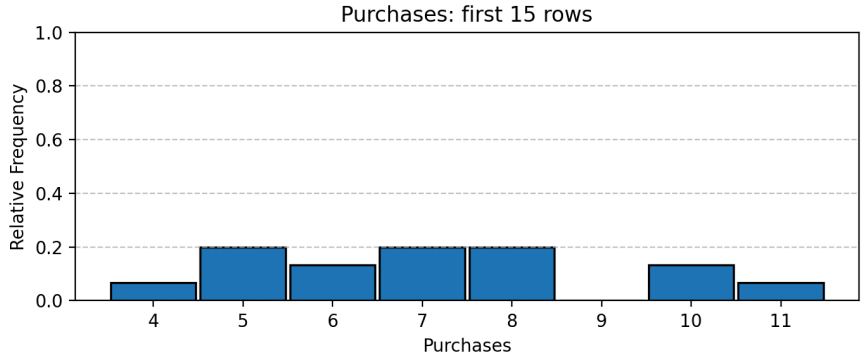
- Histograms

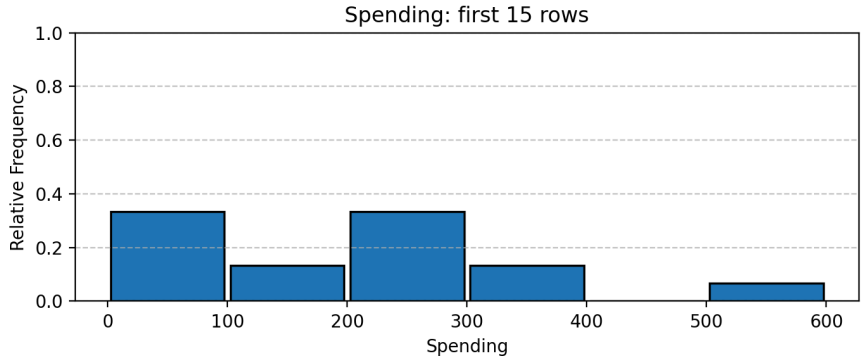
Histograms are used for quantitative data:

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

Examples with our dataset:

- Satisfaction (1-5): Discrete numeric
- Purchases: Discrete numeric
- Spending (€): Continuous numeric





Presenting Data Graphically

- Histograms

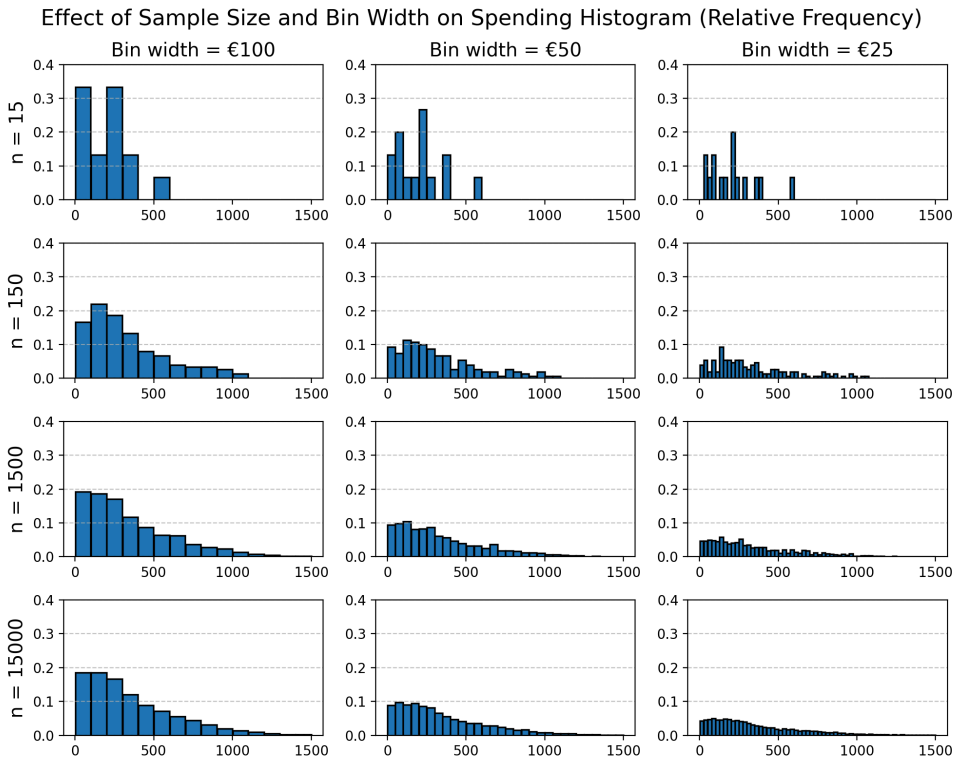
The look of a histogram depends on two key factors:

Sample size (n):

- Small n: Unstable and less detail
- Larger n: Smoother shape
- The Law of Large Numbers: Relative frequencies converge to true probabilities as n grows

Bin width:

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



Exercise Set

- Part 1

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 (€)
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

Exercise Set

- Part 2

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

9. Create a frequency table using bins of width 9 minutes:

$$0–9,\ 10–19,\ 20–29,\ 30–39$$

10. Calculate the relative frequency for each bin
11. Identify which interval contains the highest proportion of 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

Measures of Central Tendency

- Building Intuition

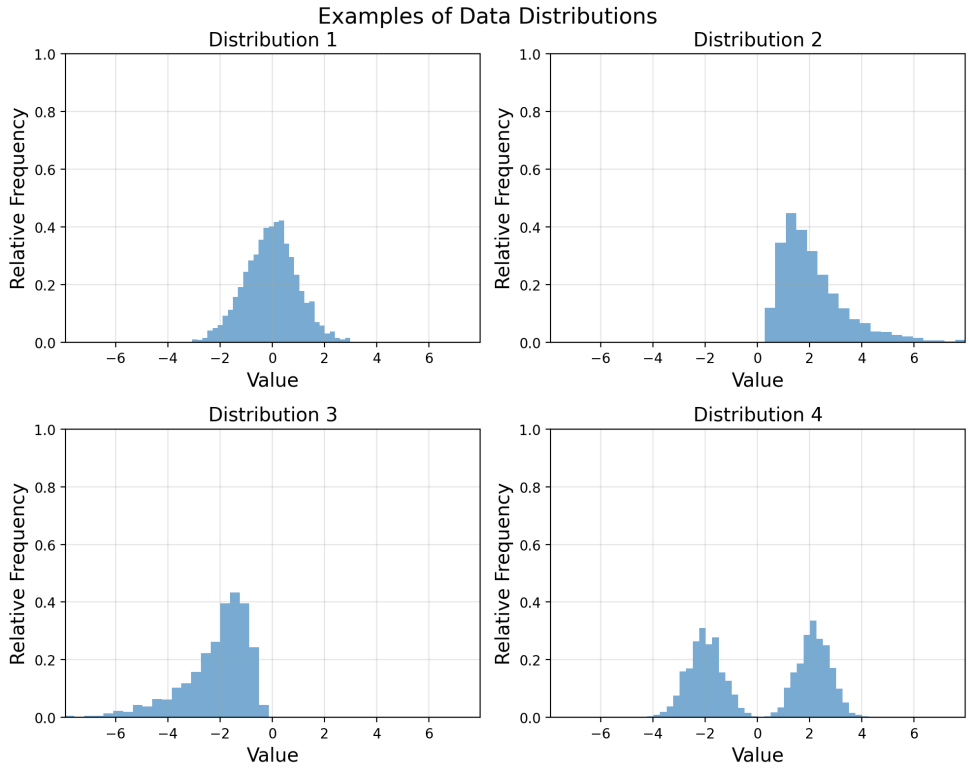
Quantitative data can be described not only verbally and graphically, but also with numbers.

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:

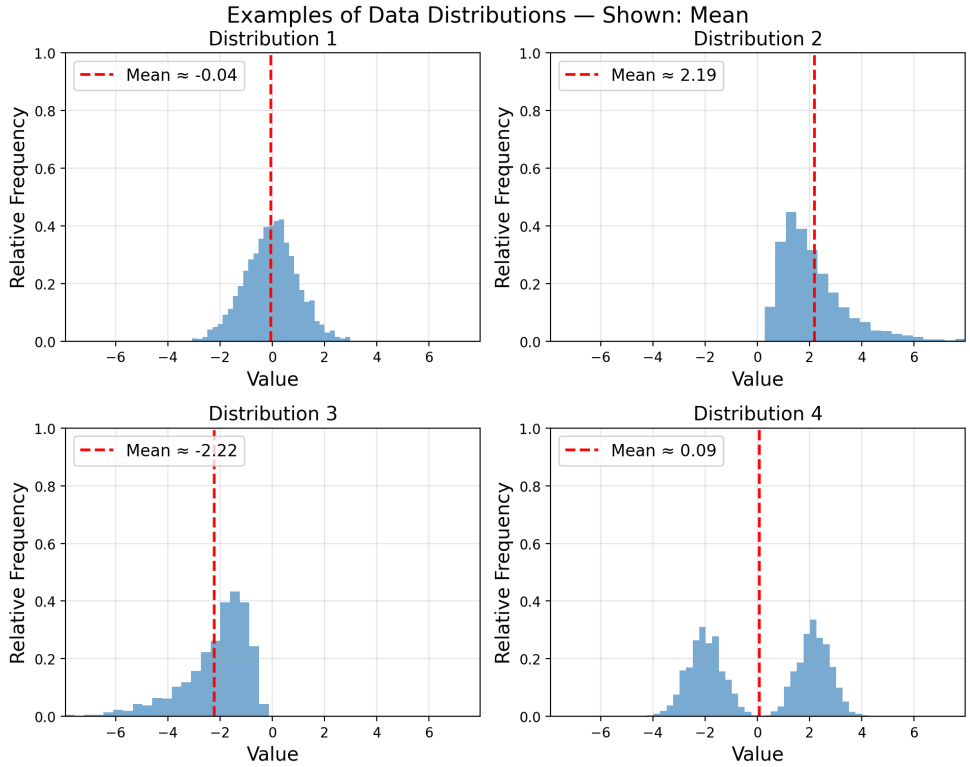
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

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

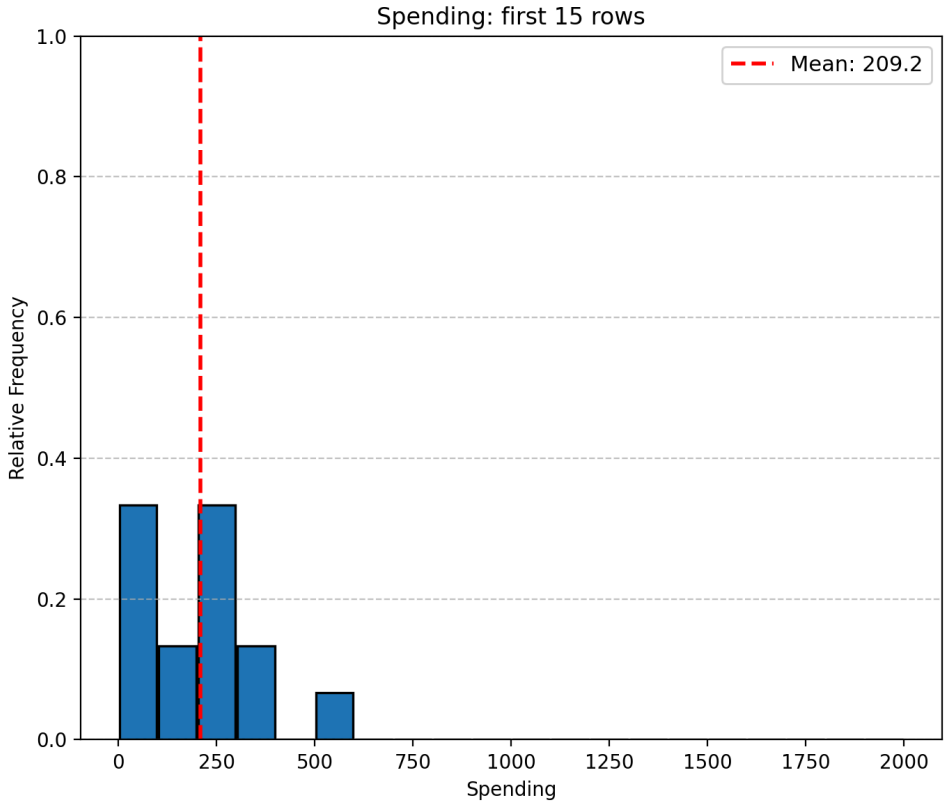
- Example: Online Customer Dataset

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

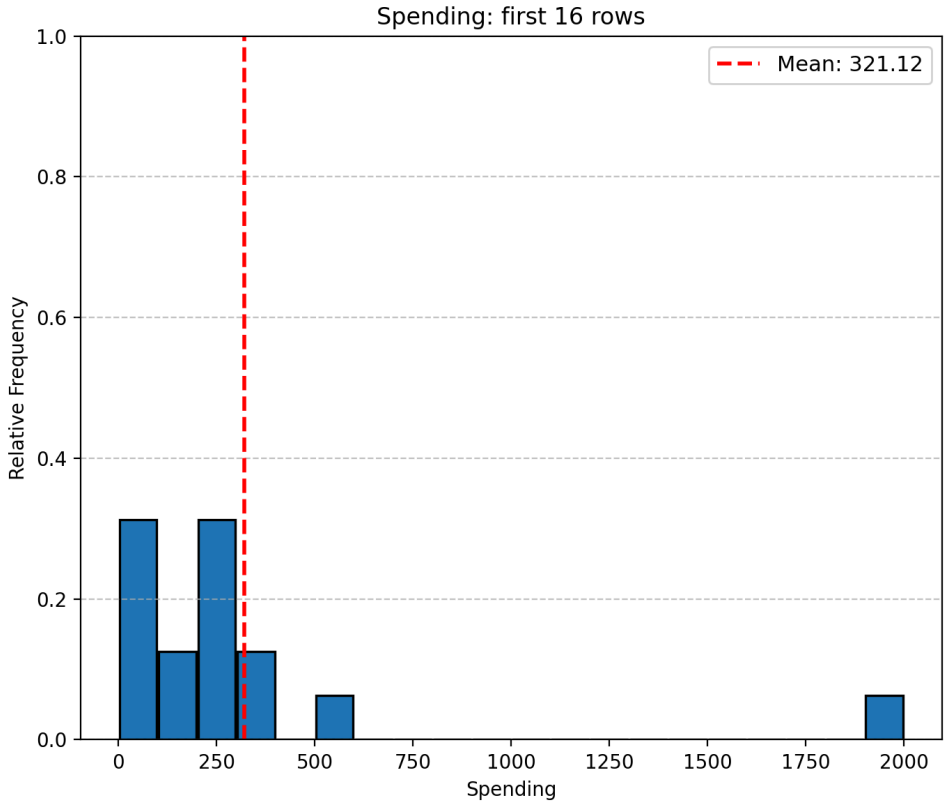
Now suppose a new customer spends 2000 euros. Including this value, the mean becomes:

$$\overline{x} = \frac{294 + 380 + 45 + \cdots + 245 + 2000}{15}$$
$$\approx 321.12$$

While 321.12 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

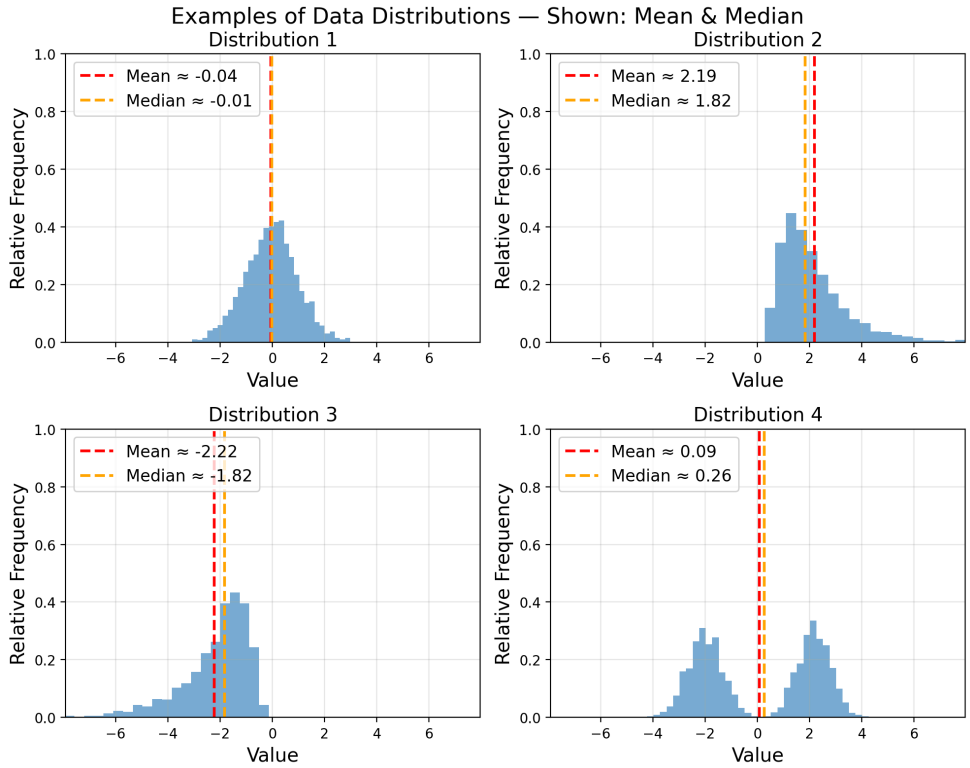
- Definition

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.



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:

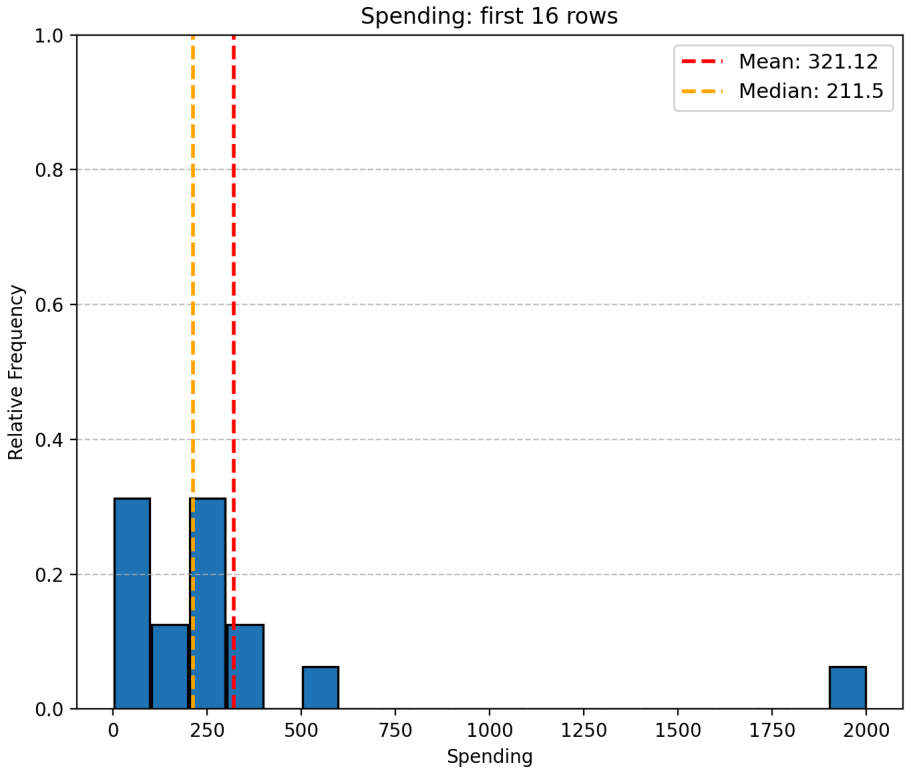
$$\underbrace{38, 45, 55, 88, 90, 125, 165,}_{\text{lower half}} \ \underbrace{207, 216}_{\text{middle}} \ \underbrace{224, 245, 294, 371, 380, 595, 2000}_{\text{upper half}}$$

The median is therefore:

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

The Median

- Example: Online Customer Dataset (Continued)



The Mode

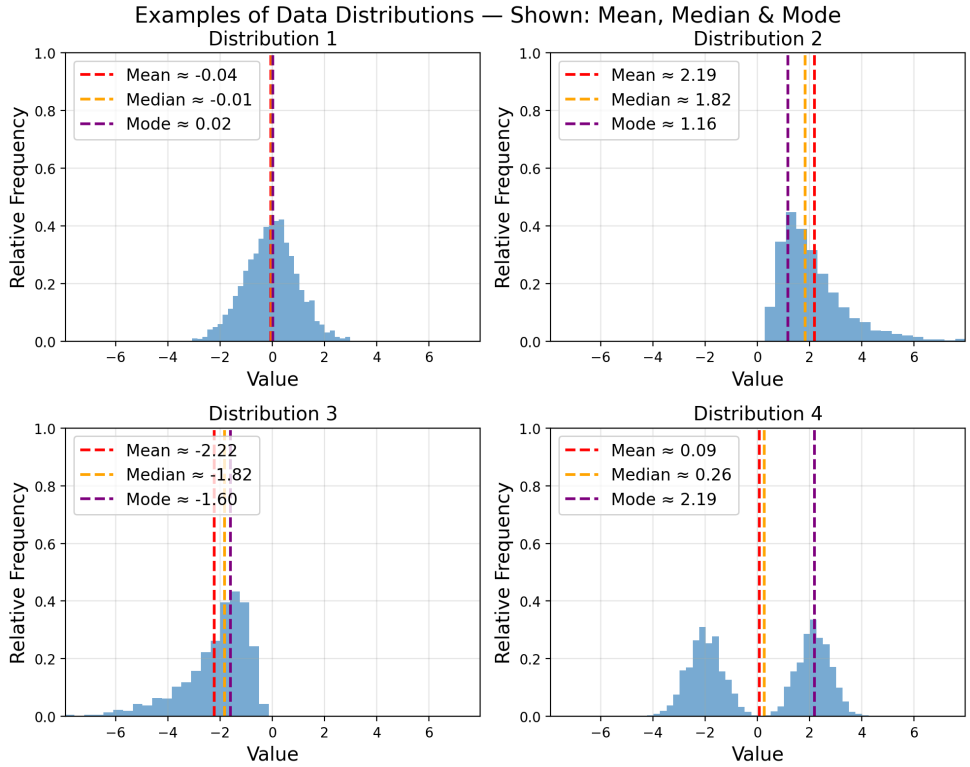
- Definition

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 mode is found by identifying the bin or interval containing the largest number of values.



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. Uniform distribution

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:

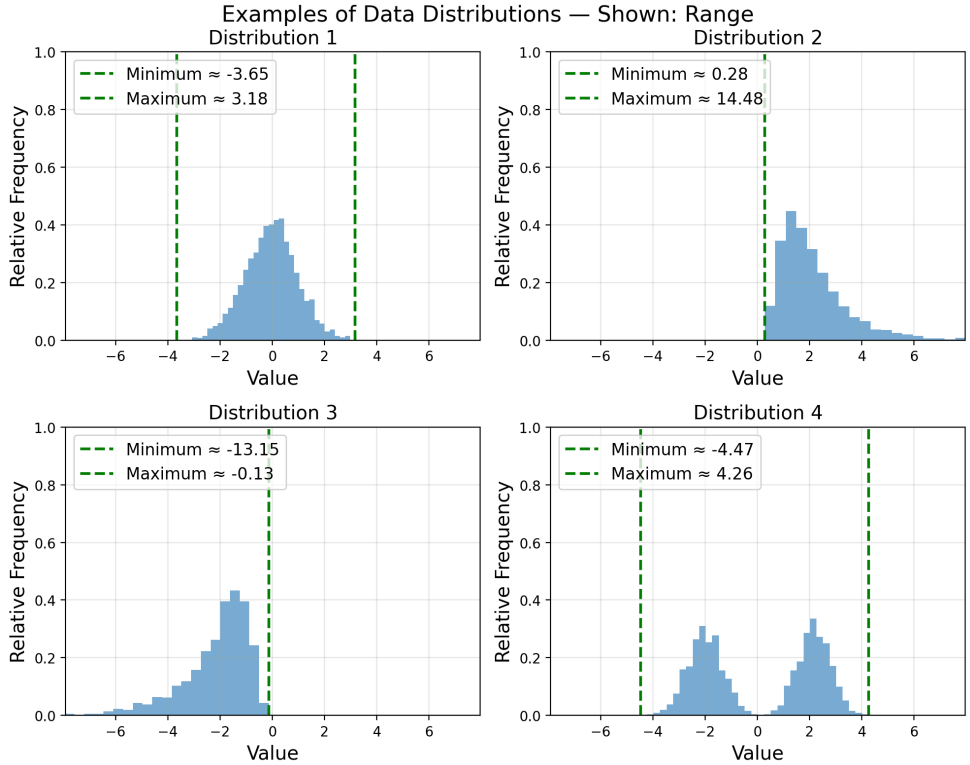
- The largest value (maximum)
- The smallest value (minimum)

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

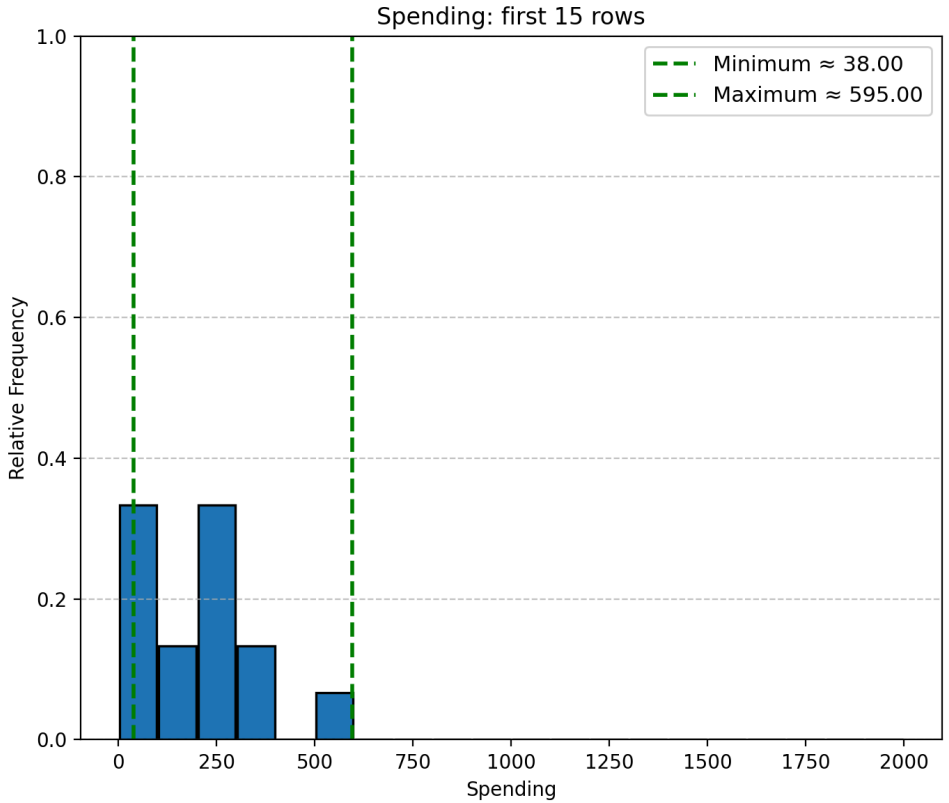
Consider the spending data for 15 customers:

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

- 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

- Example: Online Customer Dataset

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 (€)	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

- Example: Online Customer Dataset

Deviations alone are not enough because positive and negative values cancel, so to measure the average distance from the mean, we square each deviation:

- Squaring makes all deviations positive
- Larger deviations have a bigger effect

The sample variance is the mean of the squared deviations:

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

For this dataset:

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

Spending (€)	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

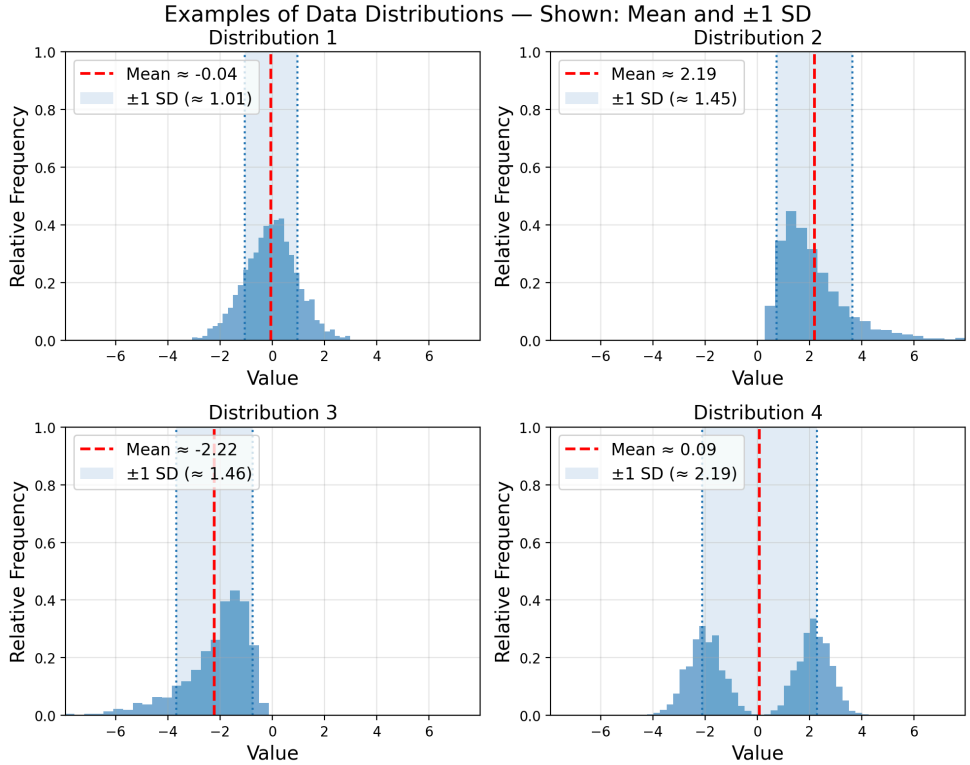
Variance is in squared units, which makes it less intuitive. On the other hand, the standard deviation expresses spread in the same units as the data:

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

- s is the average distance from the mean
- 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

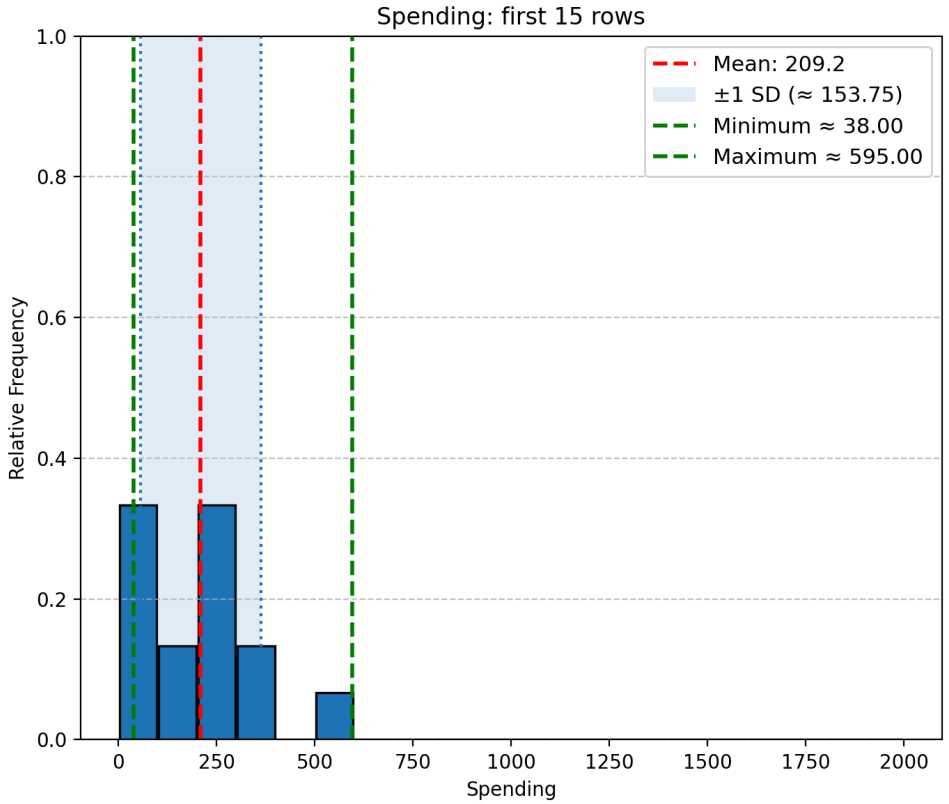


Standard Deviation

- Example: Online Customer Dataset

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

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



Percentiles

- Definition

A percentile is a value in the dataset below which a given percentage of data values fall:

- The k-th percentile is the value at or below which $k\%$ of data values fall
- Percentiles are a measure of position

Quartiles are common percentiles that split the data into four equal parts (each $\approx 25\%$):

- 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 values of the five-number summary split the data into four equal quarters:

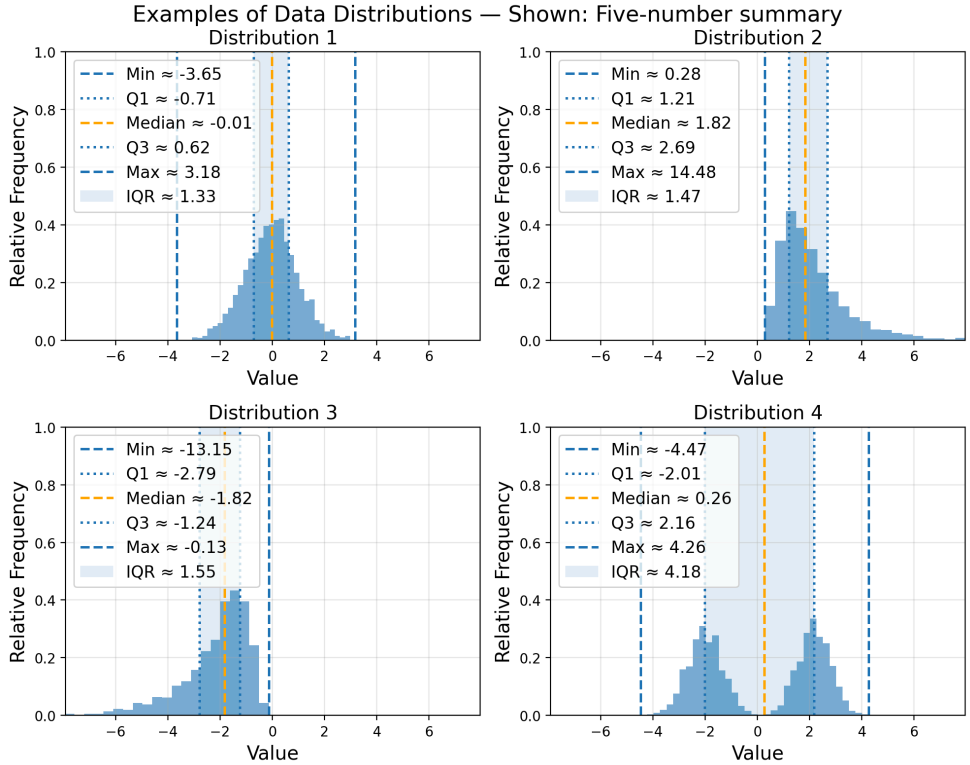
- 25% between Minimum and Q_1
- 25% between Q_1 and Median
- 25% between Median and Q_3
- 25% between Q_3 and Maximum

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

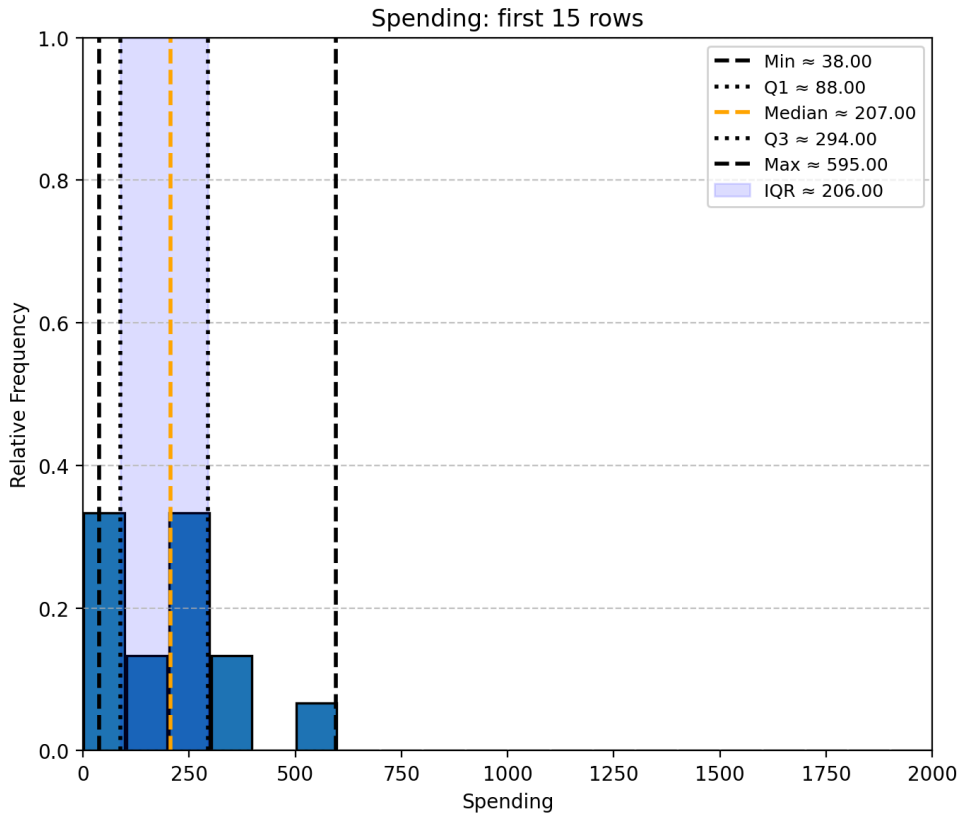
- Example: Online Customer Dataset

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

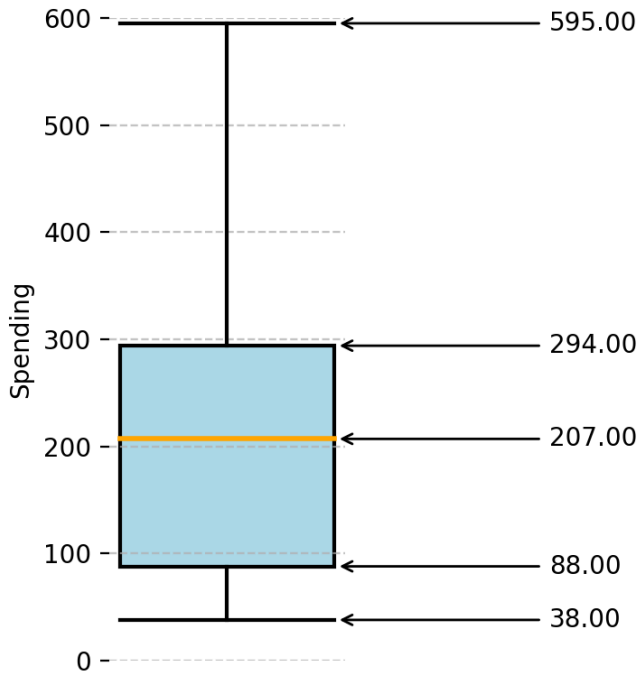
A box plot, or box-and-whisker plot, is a graphical representation of the five-number summary.

Key parts of a box plot:

- The box spans from Q_1 to Q_3
- A vertical line inside the box marks the median (Q_2)
- Whiskers extend to the minimum and maximum
- Symbols can be added to indicate outliers beyond the whiskers

What a box plot shows:

- The center of the data (median)
- The spread (range and IQR)
- Any skewness or outliers



Exercise Set

Given the following dataset:

$$4, 5, 5, 6, 7, 7, 8, 8, 10, 12, 13, 15$$

Complete the following tasks:

1. Calculate the mean
2. Calculate the sample variance
3. Calculate standard deviation
4. Find the five-number summary and IQR

Based on the mean, median, and five-number summary, try to answer:

6. Is the data symmetrically distributed or skewed?
7. Which measure of center (mean or median) best represents the data?
8. Does the spread suggest that most values are close to the center?