Statistics Pre-course
Descriptive Statistics

Ilaria Mozzetta

Tor Vergata University of Rome

September 2026

General Information

- **Instructor:** Ilaria Mozzetta
- **Email:** ilaria.mozzetta@uniroma1.it
- **Office Hours:** By appointment.
- **Programme:**
  - Part I: Descriptive Statistics
  - Part II: Probability theory
  - Part III: Inference
- **References:**
  - Cicchitelli, G., D'Urso, P., & Minozzo, M. Statistics: principles and methods. Pearson, 2021
  - Newbold, P., Carlson, W. L., & Thorne, B. Statistics for Business and Economics. Pearson, 2012.

Part I Syllabus

1. Introduction
2. Types of Data
3. Frequency Distributions
4. Graphical Representation
5. Measures of Centrality
6. Measures of Variability
7. Measures of Association

What is Descriptive Statistics?

- Descriptive statistics is concerned with the collection, presentation, and characterization of data gathered from **statistical variables**.
- Its primary objective is to summarize and describe key features of the data clearly and concisely without drawing broader population-wide generalizations.
- It typically relies on two main tools:
  - **Graphical methods:** Visual summaries such as histograms, boxplots, and bar charts.
  - **Numerical measures:** Quantitative values such as averages, percentages, and measures of variability.

Basic Statistical Concepts

- **Statistical Unit:** An individual member of the population we are interested in studying (e.g., a person, a family, or a firm), usually indexed by i.
- **Population:** The complete collection of all statistical units we want to find out about.
- **Sample:** A subset of the population that we actually observed.
- **Statistical Variable:** Any characteristic observed or measured on the units in a study (e.g. age, level of education, income, etc).
- **Modality:** The specific value that a variable takes for a given statistical unit (e.g. if the variable is level of education the modalities are: primary school, middle school, high school diploma, bachelor degree etc). Denoted as $x_1, \ldots x_K$
- **Observation:** The recorded value measured from a variable.

Example

- **Example:** Grades in Tor Vergata Statistics' Course
- **Observations:** 18, 24, 22, 28, 30, 30, 25, 25, 23, 24, ...
- **Unit:** single student
- **Population:** all students of Tor Vergata
- **Sample:** one class
- **Modality:** 18 : 30

Types of Variables

- Qualitative (Categorical)
  - Nominal
  - Ordinal
- Quantitative (Numerical)
  - Discrete
  - Continuous

Categorical Variables: variables that are not numbers

A variable is categorical if its observations belong to one of a set of distinct categories

- **Nominal:** categories are disconnected
  - Eyes colour, McDonald's Sandwiches, Citizenships, etc
- **Ordinal:** categories are ranked
  - Level of Education (PhD, Master, Bachelor, ..), Energy classes (A++, A+, B, ..), Rating of products (5 stars, 4 stars, 3 stars, ...), etc

Quantitative variables: variables that are numbers

A variable is quantitative if its observations take numerical values that reports different magnitudes of a certain measurement

- **Discrete:** the variable assumes values in a countable set. $\rightarrow$ integer numbers
  - Number of TV Show episodes, Number of rooms in a house
- **Continuous:** the variable takes values in a continuous set. $\rightarrow$ decimal numbers
  - Time, length, weight, acceleration, most physical measures

Ordinal Qualitative Variables: The Numerical Trap

- **Definition:** Qualitative variables whose categories have a natural order or ranking (e.g., satisfaction levels, education degrees).
- **The Numerical Illusion:** They are frequently recorded using numbers (e.g., 1 to 5) for entry convenience and data processing.
- **Why they remain qualitative:**
  - The numbers are merely labels for positions, not true quantities.
  - Mathematical operations lack literal meaning (e.g., a rating of 4 is not "twice as good" as a rating of 2, and the gap between 1 and 2 may not equal the gap between 4 and 5).
- **Practical Treatment:** While software or researchers sometimes treat them numerically to calculate metrics like averages, **conceptually and mathematically** they remain ordinal qualitative data.

Wrap Up: Alternative Exercises

For the following variables, determine the type, modalities, and the statistical unit of measurement:

1. Daily commuting time of university students to campus (in minutes).
2. Number of goals scored by football players in a specific tournament.
3. Monthly rent prices of apartments in downtown Rome (in euros).
4. Classification of smartphone operating systems (iOS, Android, etc.).
5. Restaurant customer satisfaction ratings (measured on a scale from 1 to 5).

Summarizing Data: What are we looking for?

Raw data is just a long list of numbers or words. To understand it, we look for different features depending on the variable type:

- **Categorical Data:** We look at **frequencies and proportions** (how observations are distributed across categories).
  - *Example:* How many days in the year were "Sunny" versus "Rainy"?
- **Quantitative Data:** We look at the **center and variability** (how high the values are and how much they spread out).
  - *Example:* What is the average annual precipitation, and how much does it vary from year to year?

Frequency Distributions

Consider a sample of N observations denoted as $y_1, y_2, \ldots, y_N$.

- **Absolute Frequency (n_i):** The number of times a specific value or category appears in the sample.
- **Relative Frequency (f_i):** The proportion of observations taking a certain value, calculated as:

$$f_i = \frac{n_i}{N}$$

- **Key Properties:**
  - The sum of all absolute frequencies equals the total sample size (N).
  - The sum of all relative frequencies always equals 1.

Frequencies Table

A frequency distribution is a table that lists the values of a variable X along with the corresponding frequencies.

Modalities	Absolute frequency n_i	Relative frequency f_i
x_1	n_1	f_1
.	.	.
.	.	.
x_k	n_k	f_k
$\sum$	N	1

Example: Global Social Media Active Users (Q4 2025)

Data:

- **Variable type:** Categorical
- **Population:** Global social media users
- **Sample:** 5,000.0 million users (World)
- **Unit:** Single active user

Platform (x_i)	n_i **(M)**	f_i
Facebook	2000.0	0.400
YouTube	1250.0	0.250
WhatsApp	800.0	0.160
Instagram	500.0	0.100
TikTok	250.0	0.050
WeChat	100.0	0.020
Telegram	50.0	0.010
Snapchat	30.0	0.006
X	15.0	0.003
Pinterest	5.0	0.001
Σ	**5000.0**	**1.000**

Cumulative relative frequency: adding a new column to the table

- We can define the **Cumulative Relative Frequency** as the proportion of observations taking a value less than or equal to x_i.
- Take the modalities in increasing order $x_1, \ldots, x_k$ and define

$$F_i = \sum_{j \leq i} f_j$$

Example

Consider the following sample, consisting of the grades achieved in last year's Statistics exam.

19, 24, 26, 30, 24, 30, 29, 24, 28, 18, 29, 29, 21, 30,
25, 19, 20, 28, 23, 26, 23, 22, 30, 30, 18, 23, 28, 30, 22

- **Sample:** 29 students from last year
- **Population:** Statistics' students at Tor Vergata
- **Variable:** Grades

Example

x_k	n_k	f_k	F_k
18	2	0.069	0.069
19	2	0.069	0.138
20	1	0.034	0.172
21	1	0.034	0.207
22	2	0.069	0.276
23	3	0.103	0.379
24	3	0.103	0.483
25	1	0.034	0.517
26	2	0.069	0.586
28	3	0.103	0.690
29	3	0.103	0.793
30	6	0.207	1.000
Total	**29**	**1.000**	—

Exercise

- The following table reports the number of books Tiziano read from January to September 2020.

Month	x
Jan	2
Feb	1
Mar	3
Apr	0
May	?
Jun	4
Jul	1
Aug	3
Sep	2
Σ	19

- Describe the variable: type, unit, etc
- Fill in the missing value ?
- Complete the table adding columns for relative and cumulative frequencies

Frequency tables for continuous data

- When variables are continuous, frequency tables are built based on intervals rather than on single modalities
- Divide all the modalities in a finite number of classes $[\ell_1, u_1), \ldots, [\ell_k, u_k]$
- Absolute (Raw) frequency n_i: how many times the i-th modality appears in the sample (in a given class)

$$n_i = \sum_{j=1}^{N} (y_j \in [\ell_i, u_i)) \quad \text{for } i = 1, \ldots, k$$

Frequency tables for continuous data

Classes $[\ell, u)$	Absolute frequency n_k	Relative Frequency f_k
$[\ell_1, u_1)$	n_1	f_1
.	.	.
.	.	.
$[\ell_K, u_K]$	n_K	f_K
Σ	N	1

Classes can have different sizes!

Example: Quantitative Continuous (Class Intervals)

Data:

- **Context:** Adult Height Survey
- **Variable:** Height (Continuous)
- **Observation (y_i):** Exact height of i-th person (cm)
- **Sample Size:** $N = 100$ individuals
- **Classes:** 10 cm intervals $[l, u)$

Class $[\ell, u)$	n_k	f_k
$[150, 160)$	5	0.05
$[160, 170)$	20	0.20
$[170, 180)$	45	0.45
$[180, 190)$	25	0.25
$[190, 200]$	5	0.05
Σ	**100**	**1.00**

Unequal Class Sizes and Densities

When classes have different widths $w_k = u_k - \ell_k$, raw frequencies (n_k or f_k) become misleading for comparison.

To adjust for varying class sizes, we define frequency density:

- **Absolute Density:**

$$d_k = \frac{n_k}{w_k} = \frac{n_k}{u_k - \ell_k}$$

- **Relative Density:**

$$h_k = \frac{f_k}{w_k} = \frac{f_k}{u_k - \ell_k}$$

Example: Continuous Data with Unequal Classes

Data:

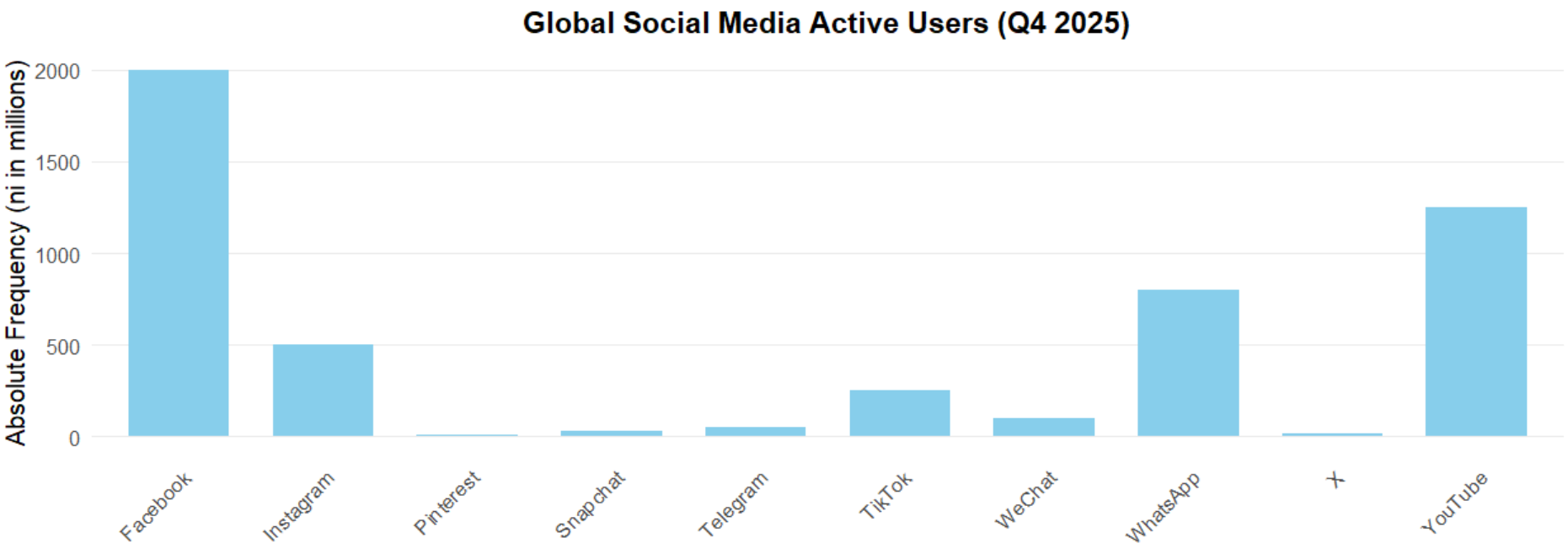
- **Context:** Annual Household Income
- **Variable:** Income (Continuous, in thousands of €)
- **Observation (y_i):** Income of i-th household
- **Sample Size:** $N = 100$ households
- **Classes:** Unequal widths ($w_k = u_k - \ell_k$)

Class $[\ell, u)$	w_k	n_k	f_k	$d_k = \frac{n_k}{w_k}$
$[0, 20)$	20	15	0.15	0.75
$[20, 50)$	30	45	0.45	1.50
$[50, 100)$	50	30	0.30	0.60
$[100, 200]$	100	10	0.10	0.10
Σ	–	**100**	**1.00**	–

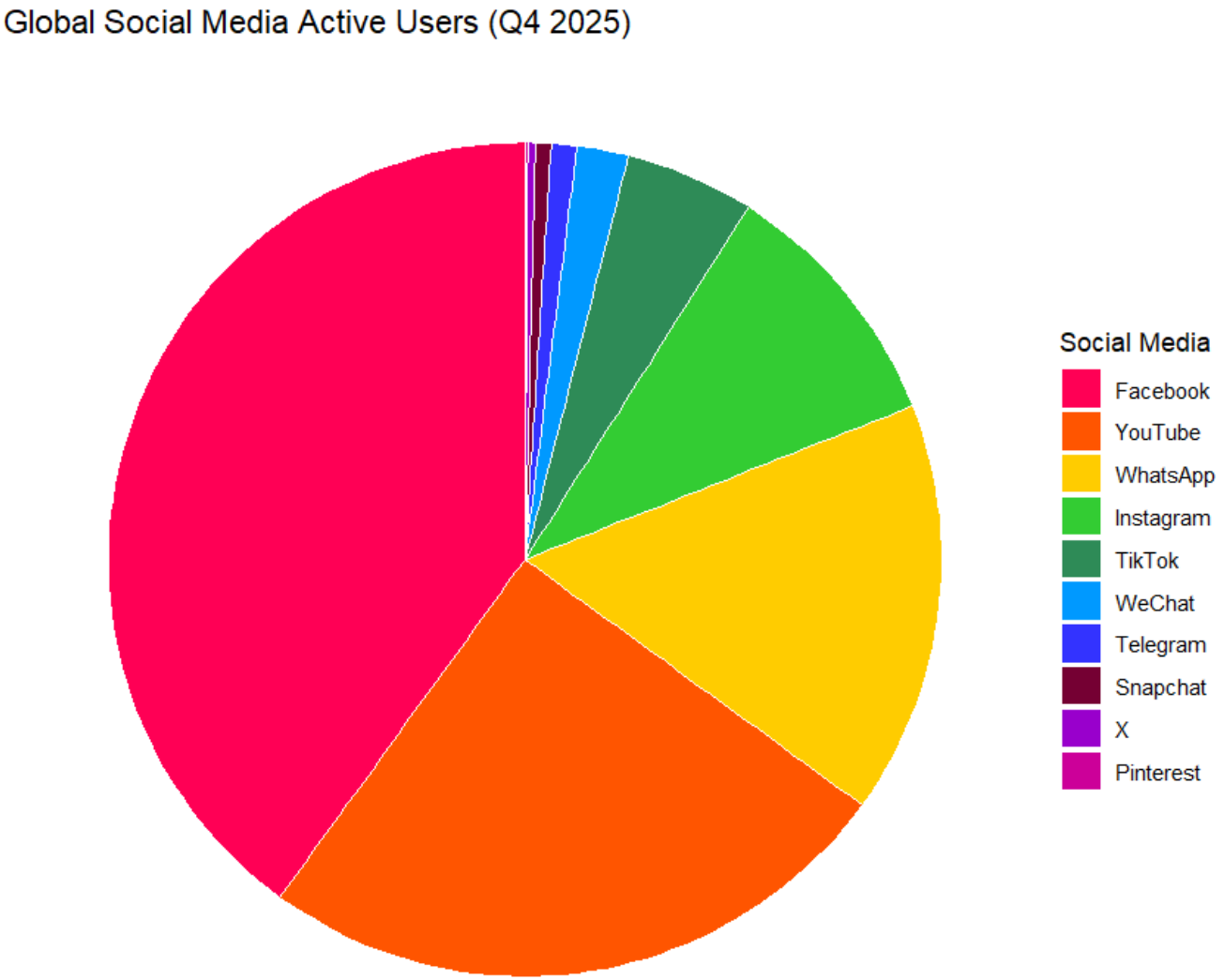
Graphical Representation

- Frequency tables allow us to describe and summarise the data
- Graphical representations makes it even easier to analyze
- **Barplot:** for categorical and discrete data
  - each modality is associated to a bar whose height corresponds to the absolute frequency
- **Pie chart:** for categorical and discrete data
  - each modality is associated to slice of a pie whose width corresponds to the absolute frequency
- **Histogram:** for continuous data
  - each class is associated to a bar whose height corresponds to its frequency in the case of equal size classes
  - each class is associated to a bar whose height corresponds to its density in the case of unequal size classes

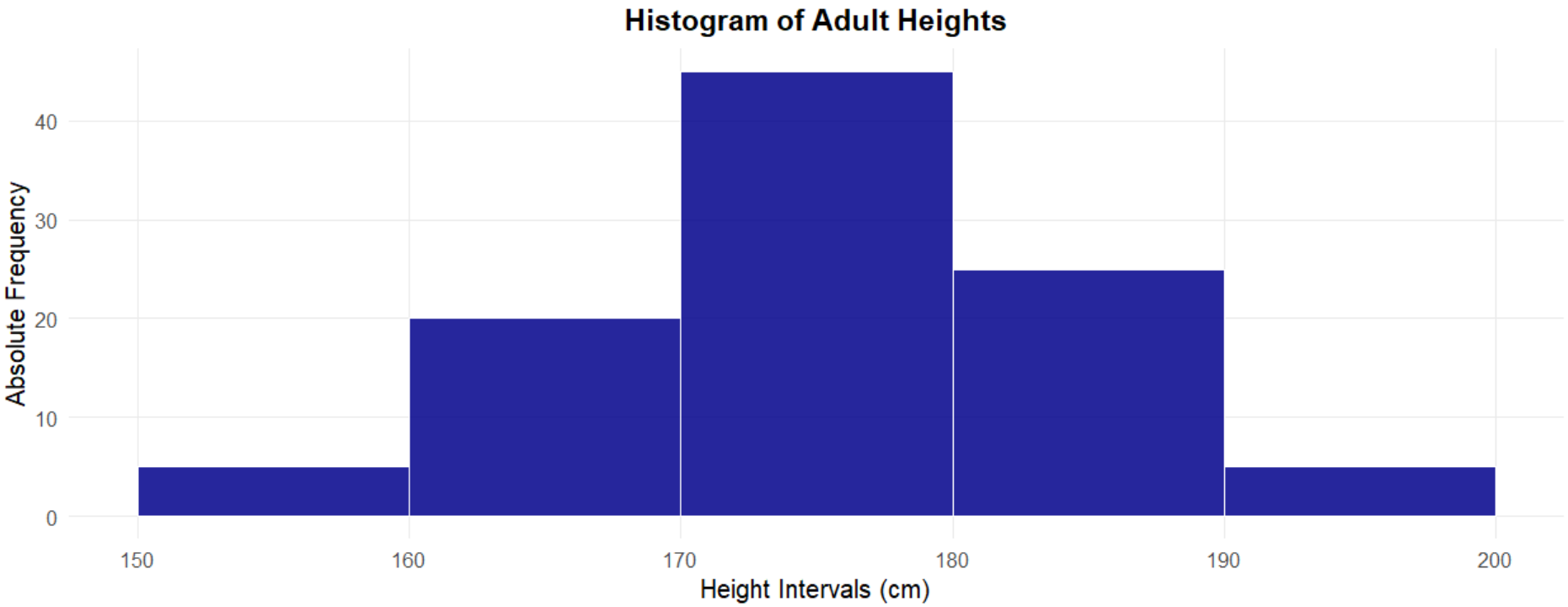
Barplot: Social Media Users



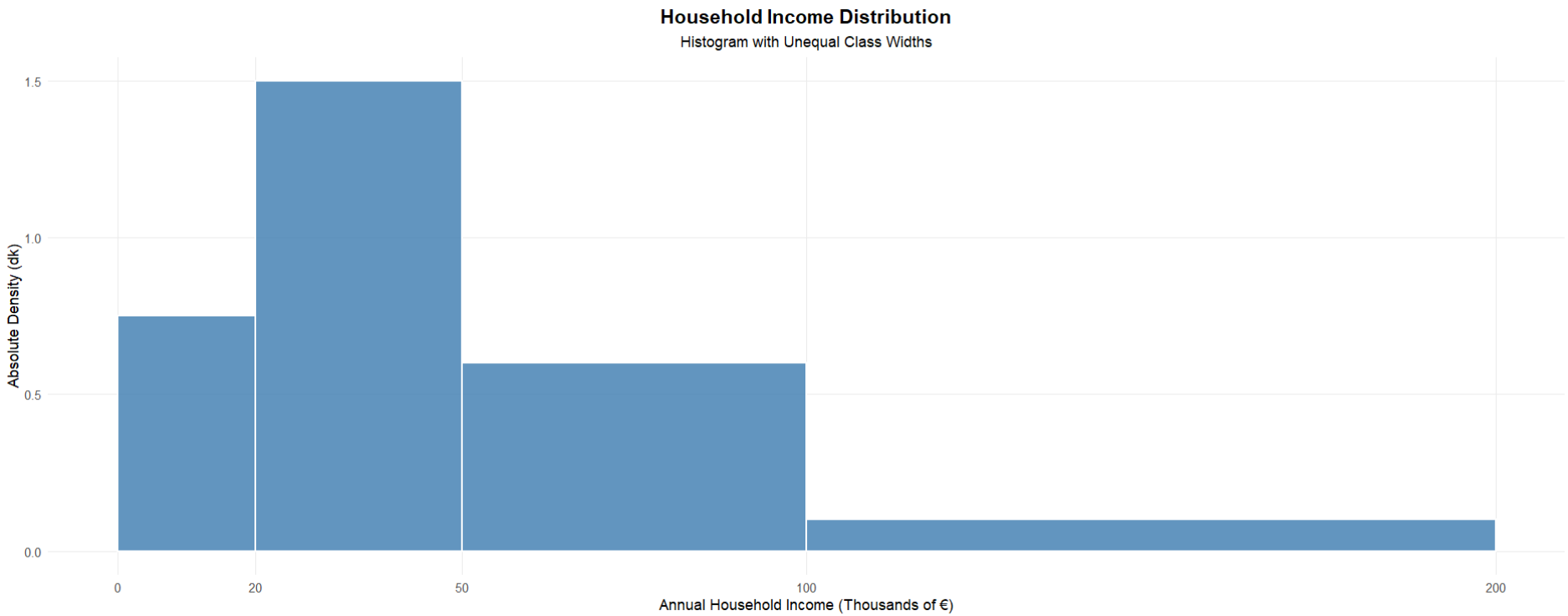
Pie chart: Social Media Users



Histogram: adult heights



Histogram: income



Exercise

- The following table reports the number of books read last year by a sample of students:

Books (x_i)	**Students (n_i)**
0	2
1	3
2	4
3	2
4	1
Σ	**12**

- Represent this distribution graphically and justify your choice of chart type.

Exercise

- The weights (in kg) of a sample of $N = 30$ adults $(y_1, \ldots, y_N)$ are recorded as follows:

  68.5, 72.0, 55.4, 81.2, 63.0, 75.6, 69.1, 88.4, 70.5, 62.1,
  77.3, 65.8, 80.0, 71.2, 59.6, 74.3, 67.9, 83.1, 66.5, 78.0,
  73.4, 61.2, 85.0, 69.8, 76.2, 64.0, 79.5, 72.1, 68.3, 77.0

- **Tasks:**
  1. Construct the frequency distribution table.
  2. Draw a suitable graphical representation and justify your choice.

Numerical Summaries

- Sometimes we want to describe a whole dataset using just a few simple numbers.
- We focus on two main questions:
  - **Centrality:** What is the "typical" or middle value?
  - **Variability:** How spread out are the data? (Are they close together or very different?)

Measures of central tendency
location of the distribution

What is the "typical" value for the observations?

- **Mode**
  - the value that appears more often (it can be more than one!)
- **Median**
  - the value that splits the distribution in half
- **Mean**
  - the balance point of the distribution

These can all be computed from a frequency table!

The Mode

- The mode is the modality with the highest observed frequency

$$y_{Mode} = \{x_j : f_j \geq f_i \ \forall \ j \neq i\}$$

- This is the most general notion of centrality and applies to all types of data (numerical and categorical)
- If data are grouped (e.g. divided into classes) then the notion of mode becomes the **modal class**

Note: there can be more than one mode!

Example: Mode vs. Modal Class

Mode

x_i	n_i
Red	4
Green	7
Blue	12
Yellow	3

Task: Identify the **Mode**.

Modal Class

x_i	n_i
$[10, 20)$	5
$[20, 30)$	16
$[30, 40)$	9
$[40, 50)$	4

Task: Identify the **Modal Class**.

Examples: Mode vs. Modal Class

Mode

x_i	n_i
Red	4
Green	7
Blue	**12**
Yellow	3

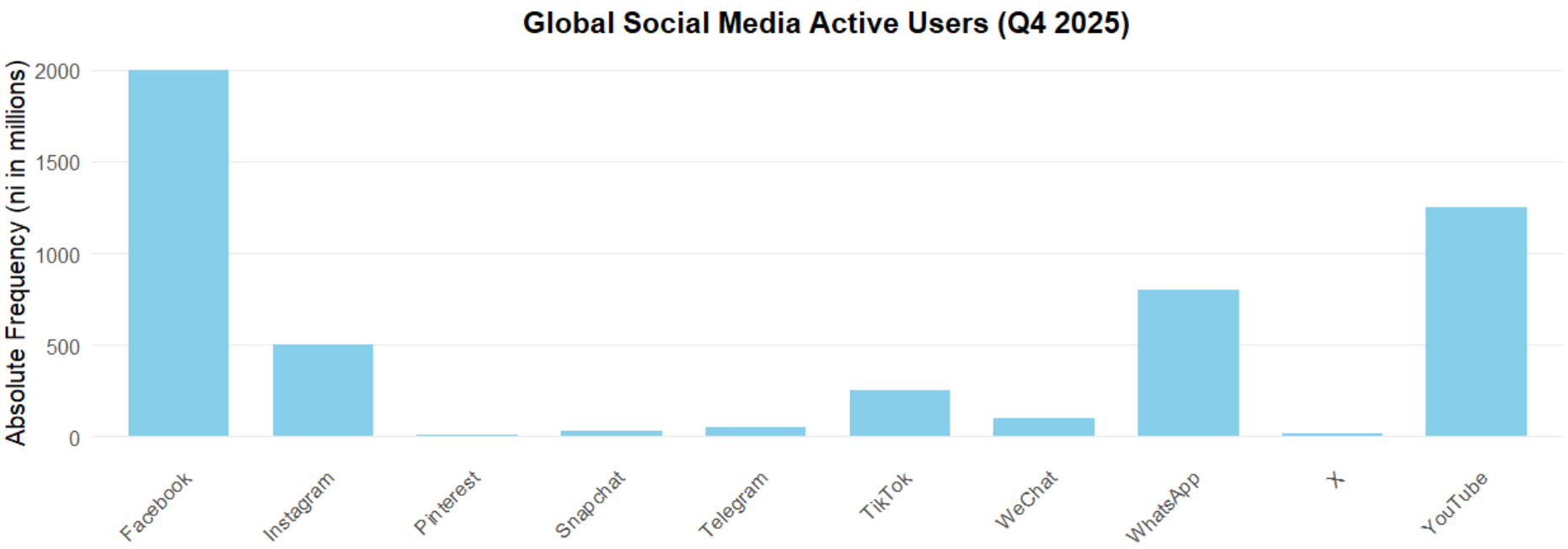
Mode: $y_{\text{Mode}} =$ **Blue**
(Highest frequency $n = 12$)

Modal Class

x_i	n_i
$[10, 20)$	5
$[20, 30)$	**16**
$[30, 40)$	9
$[40, 50)$	4

Modal Class: $[\mathbf{20}, \mathbf{30})$
(Highest frequency $n = 16$)

Back to the social media users example



The Median

- The **median** (y_{MED}) is the central value that splits an ordered distribution into two equal halves.
- It applies to both **quantitative data** (discrete or continuous) and **ordinal qualitative data**.

Let $y_{(1)}, y_{(2)}, \ldots, y_{(N)}$ be the **ordered sample**. The median can be computed:

1. From Raw Data

- **If N is odd:**

$$y_{\text{MED}} = y_{\left(\frac{N+1}{2}\right)}$$

- **If N is even:**

$$y_{\text{MED}} = \frac{y_{\left(\frac{N}{2}\right)} + y_{\left(\frac{N}{2}+1\right)}}{2}$$

2. From Frequency Table

- Compute cumulative relative frequencies (F_k).
- The median is the **first modality** x_k where:

$$F_k \geq 0.50$$

The Median

- The median can be computed for (almost) all types of data: quantitative (discrete and continous) and ordinal qualitative.

Examples:

- Grades from Statistics' exam
  - $18, 18, 19, 20, 21, 22, 22, 22, 26, 26, 26, 27, 27, 28, 28 \implies y_{MED} = ?$
  - $18, 18, 19, 20, 21, 22, 22, 22, 26, 26, 26, 27, 27, 28, 28, 29 \implies y_{MED} = ?$
- Customers' survey data
  - Bad, Bad, Average, Average, Good, Great, Great

Example: Median of Grades

Dataset

Grades from Statistics students ($N = 26$):
24, 22, 23, 23, 23, 24, 22, 30, 28, 28, 27, 26, 27, 26, 28, 19, 18, 18, 18, 19, 20, 20, 21, 21, 22, 23

- **Task:** Compute the **Median** both directly from the raw data and using a frequency distribution table.

An example with grades

- Grade from Statistics' students:
  24, 22, 23, 23, 23, 24, 22, 30, 28, 28, 27, 26, 27, 26, 28, 19, 18, 18, 18, 19, 20, 20, 21, 21, 22, 23
- Ordered sample:
  18, 18, 18, 19, 19, 20, 20, 21, 21, 22, 22, 22, 23, 23, 23, 23, 24, 24, 26, 26, 27, 27, 28, 28, 28, 30
- $N = 26 \implies y_{MED} = \frac{1}{2}(23 + 23) = 23$

x_k	n_k	f_k	F_k
18	3	0.115	0.115
19	2	0.077	0.192
20	2	0.077	0.269
21	2	0.077	0.346
22	3	0.115	0.461
23	4	0.154	0.615
24	2	0.077	0.692
26	2	0.077	0.769
27	2	0.077	0.846
28	3	0.115	0.961
30	1	0.039	1
N	26	1	

The median is the first modality x_k for which $F_k \geq 0.50 \rightarrow Med = 23$

An Introduction to Quantiles

- The **median** tells us what level is reached by at least 50% of the population ($p = 0.50$).
- **Quantiles** are extensions that tell us the level reached by $p \cdot 100\%$ of the population. They are defined by the first modality x_k for which:

$$F_k \geq p$$

- Quantiles of order $p = 0.25$ and $p = 0.75$ have special roles and are called the **1st Quartile (Q_1)** and **3rd Quartile (Q_3)**, respectively.

Example: Finding Quartiles (Q_1, Q_2, Q_3)

- Using the cumulative relative frequencies (F_k) from the Statistics grades example, find:
  1. **1st Quartile (Q_1)** ($p = 0.25$)
  2. **2nd Quartile (Q_2 / Median)** ($p = 0.50$)
  3. **3rd Quartile (Q_3)** ($p = 0.75$)

x_k	n_k	f_k	F_k
18	3	0.115	0.115
19	2	0.077	0.192
20	2	0.077	0.269
21	2	0.077	0.346
22	3	0.115	0.461
23	4	0.154	0.615
24	2	0.077	0.692
26	2	0.077	0.769
27	2	0.077	0.846
28	3	0.115	0.961
30	1	0.039	1.000
N	26	1	

Finding Quartiles (Q_1 and Q_3): Example

- **1st Quartile (Q_1):**
  First x_k where $F_k \geq 0.25$

  $F_{20} = 0.269 \geq 0.25 \implies \mathbf{Q_1 = 20}$

- **2nd Quartile (Q_2):**
  First x_k where $F_k \geq 0.50$

  $F_{23} = 0.615 \geq 0.50 \implies \mathbf{Q_2 = 23}$

- **3rd Quartile (Q_3):**
  First x_k where $F_k \geq 0.75$

  $F_{26} = 0.769 \geq 0.75 \implies \mathbf{Q_3 = 26}$

x_k	n_k	f_k	F_k
18	3	0.115	0.115
19	2	0.077	0.192
20	2	0.077	0.269
21	2	0.077	0.346
22	3	0.115	0.461
23	4	0.154	0.615
24	2	0.077	0.692
26	2	0.077	0.769
27	2	0.077	0.846
28	3	0.115	0.961
30	1	0.039	1.000
N	26	1	

Exercise

- Let X be the number of items returned per day at a retail store over a sample of 21 days:

$$3, 4, 4, 5, 5, 5, 6, 7, 7, 8, 9, 9, 10, 10, 11, 12, 12, 13, 14, 15, 16$$

- Compute the following summary measures:
  1. The Median: Median(X)
  2. The 1st Quartile: $q_{0.25}(X)$
  3. The 3rd Quartile: $q_{0.75}(X)$

The Mean

- The (arithmetic) mean is the sum of the observations divided by the sample size.

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

- The mean can be computed directly from a frequency distribution using the alternative definition

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{K} x_j n_j = \sum_{j=1}^{K} x_j \frac{n_j}{N} = \sum_{j=1}^{K} x_j f_j$$

where x_j are the modalities of the variable

N.B.

It can be computed only for quantitative data!

Example: Raw Data

Sample Data

Daily study hours for $N = 10$ students:

$$y = 0, 1, 1, 1, 2, 2, 2, 3, 3, 5$$

Formula:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Computation:

$$\bar{y} = \frac{0 + 1 + 1 + 1 + 2 + 2 + 2 + 3 + 3 + 5}{10}$$

$$= \frac{20}{10} = \mathbf{2} \text{ hours}$$

Example: Frequency distribution

Sample Data

Consider the frequency distribution of study hours for $N = 10$ students:

x_j	n_j	f_j
0	1	0.10
1	3	0.30
2	3	0.30
3	2	0.20
5	1	0.10
Total	**10**	**1.00**

- **Task:** Compute the mean $\bar{y}$ using both absolute frequencies (n_j) and relative frequencies (f_j).

Example: Frequency distribution

x_j	n_j	f_j
0	1	0.10
1	3	0.30
2	3	0.30
3	2	0.20
5	1	0.10
Total	**10**	**1.00**

Method A: Using absolute frequencies (n_j)

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{K} x_j n_j$$

$$= \frac{(0 \cdot 1) + (1 \cdot 3) + (2 \cdot 3) + (3 \cdot 2) + (5 \cdot 1)}{10}$$

$$= \frac{0 + 3 + 6 + 6 + 5}{10} = \frac{20}{10} = \mathbf{2}$$

Method B: Using relative frequencies (f_j)

$$\bar{y} = \sum_{j=1}^{K} x_j f_j$$

$$= (0 \cdot 0.10) + (1 \cdot 0.30) + (2 \cdot 0.30) + (3 \cdot 0.20) + (5 \cdot 0.10)$$

$$= 0 + 0.30 + 0.60 + 0.60 + 0.50 = \mathbf{2}$$

Properties of the Mean

- **Internality** $y_{(1)} \leq \bar{y} \leq y_{(N)}$
- **Linearity**
  - If $z_i = ay_i + b$ then
$$\bar{z} = a\bar{y} + b$$
- **Zero-deviation**
$$\sum_{i=1}^{N}(y_i - \bar{y}) = 0$$
- **Associativity**
  - Let $\bar{y}$ and $\bar{x}$ be the arithmetic means of two samples of sizes N and M, respectively. The mean of the combined sample is then
$$\bar{z} = \frac{N * \bar{y} + M * \bar{x}}{N + M}$$

Mean of grouped data

- Consider the formula of the mean for frequency distribution:

$$\bar{y} = \sum_{j=1}^{K} x_j f_j$$

- When observations are grouped in intervals (ℓ_j, u_j): it is enough to replace x_j with the center of the interval

$$\bar{y} = \sum_{j=1}^{K} c_j f_j \quad \text{where } c_j = \frac{1}{2}(u_j + \ell_j)$$

Example: Mean of Grouped Data

Dataset

Monthly mobile data usage for a sample of 100 subscribers (in GB):

Class Interval (x)	n	f	F
$[0, 10)$	15	0.15	
$[10, 30)$	35	0.35	
$[30, 70)$	30	0.30	
$[70, 150)$	15	0.15	
$[150, 300)$	5	0.05	
Total (Σ)	**100**	**1.00**	

- **Task:** Calculate the approximate mean monthly data usage $\bar{y}$.

Example

- Monthly mobile data usage for a sample of 100 subscribers (in GB):

Class Interval (x)	**Midpoint (c)**	n	f	F	$c \cdot f$
$[0, 10)$	5	15	0.15	0.15	0.75
$[10, 30)$	20	35	0.35	0.50	7.00
$[30, 70)$	50	30	0.30	0.80	15.00
$[70, 150)$	110	15	0.15	0.95	16.50
$[150, 300)$	225	5	0.05	1.00	11.25
Total	—	**100**	**1.00**	—	**50.50**

Mean Calculation

$$\bar{y} = \sum_{i=1}^{K} c_i f_i = 0.75 + 7.00 + 15.00 + 16.50 + 11.25 = \mathbf{50.50} \text{ GB}$$

(Or using absolute frequencies: $\bar{y} = \frac{1}{N} \sum c_i n_i = \frac{5050}{100} = \mathbf{50.50}$ GB)

Exercise

Dataset

Number of customer support tickets resolved daily by a team of $N = 15$ employees:

$$12, 14, 15, 15, 16, 18, 18, 20, 20, 20, 22, 24, 25, 27, 28$$

- For the dataset above, compute:
  1. **Mean** ($\bar{x}$)
  2. **Median** (2^{nd} Quartile, Q_2)
  3. **1^{st} Quartile** (Q_1)
  4. **3^{rd} Quartile** (Q_3)

Exercise

Dataset

Weekly video streaming time (in hours) for a sample of $N = 120$ users:

Class Interval (Hours)	Number of Users (n_i)
$[0, 5)$	20
$[5, 15)$	45
$[15, 25)$	35
$[25, 40)$	20
Total	**120**

- Based on the frequency distribution above:
  1. Identify the **Modal Class**
  2. Compute the **Mean** ($\bar{y}$).

Mean vs Median

- In the "Mean vs Median" fight we have no winners, they simply measure different things.
- The mean takes into account all the observations, including outliers $\rightarrow$ sensitive
  - An **outlier** is an extreme observation that lies abnormally far from the rest of the values in a dataset.
- The median takes into account only the order of the observations, regardless of their values $\rightarrow$ robust

The Mean and the Median coincide only when the distribution of the variable is symmetric

Mean vs. Median: Sensitivity to Outliers

1. Standard Dataset (No Outliers)

Sample values: $y = (20, 22, 23, 24, 26)$

- **Mean:** $\bar{y} = \frac{115}{5} = \mathbf{23}$
- **Median:** $y_{\text{MED}} = \mathbf{23}$

2. Introduce One Extreme Outlier

Replace 26 with an extreme value (200): $y' = (20, 22, 23, 24, \mathbf{200})$

- **Mean:** $\bar{y} = \frac{289}{5} = \mathbf{57.8}$ $\leftarrow$ *Shifted drastically!*
- **Median:** $y_{\text{MED}} = \mathbf{23}$ $\leftarrow$ *Completely unaffected!*

Takeaway

- The **Mean is sensitive** to outliers (pulled towards extreme values).
- The **Median is robust** (resistant to extreme values).

Numerical Summaries

- Sometimes we want to describe a whole dataset using just a few simple numbers.
- We focus on two main questions:
  - **Centrality:** What is the "typical" or middle value? ✓
  - **Variability:** How spread out are the data? (Are they close together or very different?)

Variability: how unpredictable is our variable?

When do we consider observations sufficiently "similar" among each other?

- **Ranges**
  - observations vary in a certain interval
- **Variance**
  - the spread of the observations around a defined value (typically the mean)

Ranges

Based on the concept of variability meant as the size of the interval in which the observations fall, we can define two measures of spread:

- **(Global) Range of Variation:** the difference between the maximum and the minimum value observed in the sample

$$RV = y_{(N)} - y_{(1)}$$

- **The Interquartile Range:** the difference between the 3rd and 1st quartile, which gives the smallest interval in which the 50% of the observations falls

$$IQ = y_{Q3} - y_{Q1}$$

Variance

The variance is based on the idea that the larger the deviations from the mean $(y_i - \bar{y})$, regardless of their signs, the larger the variability
NOTE: To exploit the idea of deviation from the center we cannot use simply the average of the deviations

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})$$

because of the zero-deviation property of the mean
Solution: square them!

- **From Raw Data:**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

- **From Absolute Frequencies (n_j):**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{K}(x_j - \bar{y})^2 n_j$$

Standard Deviation

The Standard Deviation is the square root of the variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

The Standard Deviation is on the same scale of the data, making it more interpretable than the Variance

Example: Variance from Raw Data

Sample Data

Consider a sample of $N = 5$ observations: $y = (1, 3, 5, 7, 9)$

- **Task:** Compute the mean $\bar{y}$ and the sample variance σ^2.

Example

Sample Data

Consider a sample of $N = 5$ observations: $y = (1, 3, 5, 7, 9)$

$$\bar{y} = \mathbf{5}$$

Observation (y_i)	**Deviation ($y_i - \bar{y}$)**	**Squared Deviation ($(y_i - \bar{y})^2$)**
1	$1 - 5 = -4$	$(-4)^2 = 16$
3	$3 - 5 = -2$	$(-2)^2 = 4$
5	$5 - 5 = 0$	$0^2 = 0$
7	$7 - 5 = 2$	$2^2 = 4$
9	$9 - 5 = 4$	$4^2 = 16$
Total (Σ)	**0**	**40**

Variance Calculation

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{40}{5} = \mathbf{8}$$

Example: Variance from Frequency Distribution

Frequency Table

Given the following frequency distribution:

Value (x_j)	**Frequency (n_j)**
1	1
2	2
3	4
4	2
5	1
Total	**10**

- **Task:** Compute the mean $\bar{y}$ and the sample variance σ^2.

Example

$$\bar{y} = \mathbf{3}$$

x_k	n_k	$x_k \cdot n_k$	$(x_k - \bar{y})^2$	$(x_k - \bar{y})^2 \cdot n_k$
1	1	1	$(1-3)^2 = 4$	$4 \cdot 1 = 4$
2	2	4	$(2-3)^2 = 1$	$1 \cdot 2 = 2$
3	4	12	$(3-3)^2 = 0$	$0 \cdot 4 = 0$
4	2	8	$(4-3)^2 = 1$	$1 \cdot 2 = 2$
5	1	5	$(5-3)^2 = 4$	$4 \cdot 1 = 4$
Total (Σ)	**10**	**30**		**12**

Variance Calculation

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{K}(x_j - \bar{y})^2 n_j = \frac{12}{10} = \mathbf{1.2}$$

Exercise

Customer Wait Times at a Bank Branch (in minutes).

- $N = 30$:

$$5, 5, 6, 6, 6, 7, 8, 8, 9, 10, 10, 10, 11, 12, 12, 12,$$
$$12, 14, 15, 15, 16, 18, 20, 20, 22, 25, 25, 28, 30, 35$$

- Compute the Mean, Median, and Mode. Comment on the findings.
- Plot the data, justifying your graphical choice.
- Compute the Interquartile Range, Variance, and Standard Deviation. Comment on the findings.

Two Variables: A more interesting case

- So far we have only dealt with describing, representing and summarizing a single variable
- However, in reality, we usually have to analyze more than one variable that may or may not be related
- Examples: weight and height of a subject, length and budget of a movie, age and hair color etc

We will focus only on the case of two numerical variables, but association measures also for categorical and mixed-type variables as well

Association: formalizing dependency

When we have more than one variable at hand, we want to know whether there is a relationship among them

- **Positive:** when x goes up, y tends to go up
- **Negative:** when x goes up, y tends to go down

The Covariance is an indicator that measures the association between two variables:

$$cov_{x,y} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

The Covariance: Some limitations

- The covariance is a measure of linear association; i.e. the case where the relationship between two variables x and y is of the form

$$y_i = ax_i + b$$

- The value of the covariance depends on the scale of the data. We have no general reference to assess if the sample covariance is large or small

The Correlation: A More Interpretable Measure

- The Correlation is a rescaled version of the Covariance:

$$r_{x,y} = \frac{\text{cov}_{x,y}}{\sigma_x \sigma_y}$$

- We have that $r_{x,y} \in [-1, 1]$. The closer $|r_{x,y}|$ is to 1, the stronger the linear association.
- **Scale-invariant**: Correlation does not depend on the variables' units; i.e., it is not affected by the scale of observations.
- **Symmetric**: $r_{x,y} = r_{y,x}$.

Example: Correlation

Sample Data ($N = 5$)

Study Hours (X) vs. Test Score (Y):

Study Hours (X)	**Test Score (Y)**
1	1
2	2
3	4
4	3
5	5

- **Task:** Calculate the sample correlation coefficient $r_{x,y}$ and interpret the result.

Example

Sample Data ($N = 5$)

Study Hours (X) vs. Test Score (Y): $\bar{x} = 3, \quad \bar{y} = 3$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	1	-2	-2	4	4	4
2	2	-1	-1	1	1	1
3	4	0	1	0	1	0
4	3	1	0	1	0	0
5	5	2	2	4	4	4
Total (Σ)		**0**	**0**	**10**	**10**	**9**

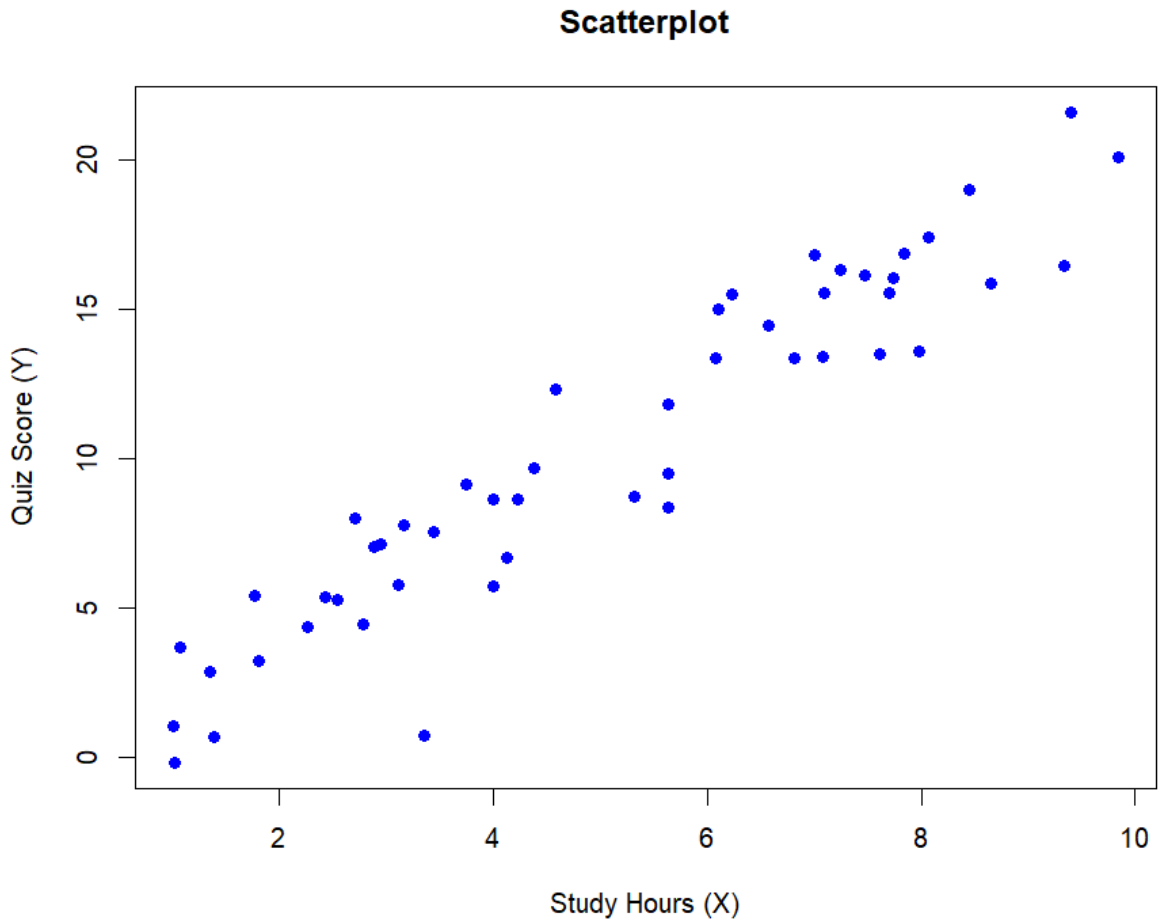
$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{9}{\sqrt{10 \cdot 10}} = \frac{9}{10} = \mathbf{0.90}$$

Interpretation: $r_{x,y} = 0.90$ indicates a **strong positive linear relationship** between study hours and test scores.

Scatterplot: graphical representation of two variables

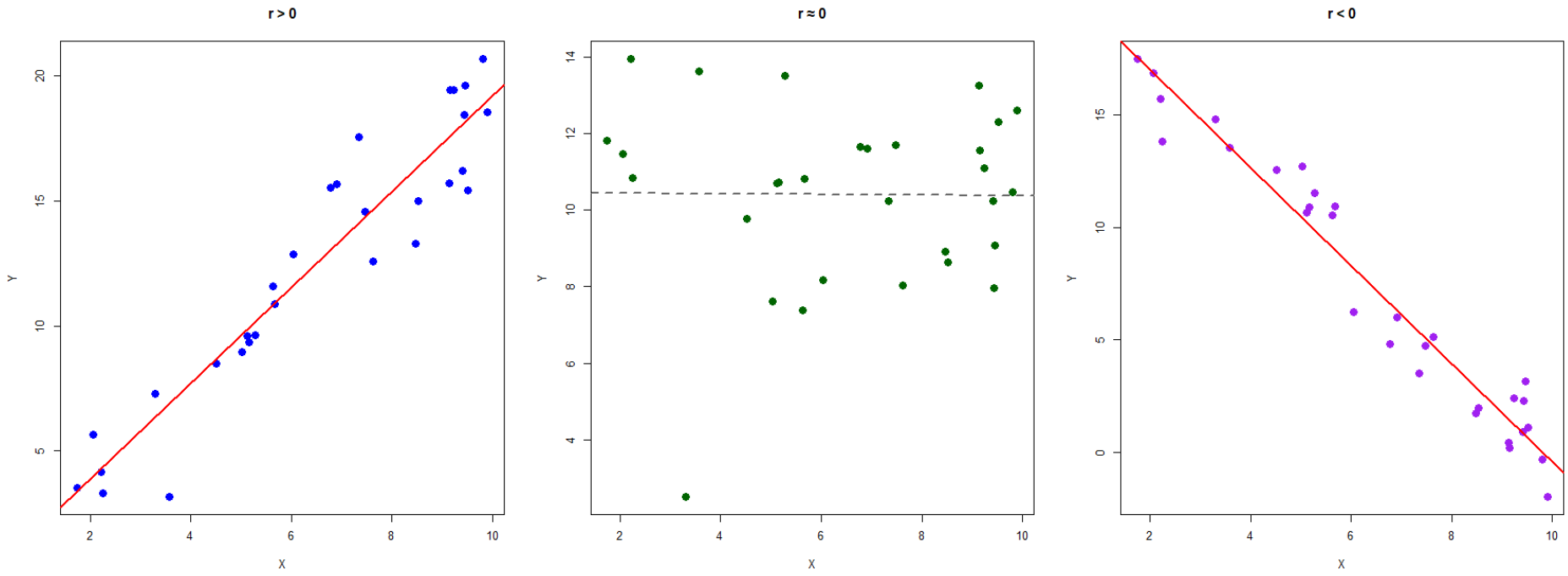
A two dimensional variable (X, Y) is usually represented through a scatterplot

- Each axis is related to one of the variables of interest
- Each point represents a unit and its coordinates correspond to the values of the variables observed on it



Visualizing Correlation

- By inspecting a **scatter plot**, we can immediately get an intuitive visual idea about the direction and strength of the correlation between two variables.



Correlation and causality

- **Correlation only measures LINEAR dependence**:
  - Complex or non-linear relationships (e.g., $y_i = ax_i^2 + b$) are not captured by $r_{x,y}$.
- **CORRELATION $\neq$ CAUSALITY!**
  - A strong statistical correlation between two variables does **not** imply that changes in one variable cause changes in the other.
- High correlation without a direct causal link is called a **spurious correlation**, often driven by an unobserved third variable (*confounding factor*).

Do Storks Deliver Babies?

The Observation

Historical data in rural European villages showed a strong **positive correlation** between the number of stork nests on rooftops and the local birth rate.

So is it true that storks deliver babies?

- **Apparent Link:** Storks $\longrightarrow$ Babies? **(False!)**
- **The Confounding Variable (Heat / House Size):**
  - Houses with large families required more heating $\longrightarrow$ more **fireplace usage**.
    $\Rightarrow$ Warm chimneys attracted storks to nest on top.

Takeaway

The correlation between storks and birth rates was **spurious**—entirely created by a third underlying factor (fireplace heat / home size).

Exercise Does LSD help with Math?

- X = Tissue concentration of Lysergic Acid Diethylamide (LSD)
- Y = Math Test score
- For the two variables X and Y compute:
  1. Mean
  2. Variance
  3. Standard Dev.
  4. Covariance
  5. Correlation

X	Y
1.17	78.93
2.97	58.20
3.26	67.47
4.69	37.47
5.83	45.65
6.00	32.92
6.41	29.97