STATISTICAL LABORATORY

Lisbon School of Economics & Management
Universidade de Lisboa

Applied Mathematics for Economics and Management
1st Year/1st Semester
2025/2026

CONTACT

Professor: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt


https://doity.com.br/estatistica-aplicada-a-nutricao


https://basiccode.com.br/produto/informatica-basica/

PROGRAM

1. Fundamental Concepts of Statistics
2. Exploratory Data Analysis
3. Organizing and Summarizing Data
4. Association and Relationships Between Variables
5. Index Numbers
6. Time Series Analysis

LECTURE 4: FREQUENCY DISTRIBUTIONS (REVIEW)

FREQUENCY TABLE FOR DISCRETE VARIABLE: EXAMPLE

1. Discrete Variable Example

Suppose we collected $n = 15$ observations of a discrete variable x:

$$x = (2, 3, 3, 4, 2, 5, 3, 4, 2, 5, 4, 3, 2, 4, 3)$$

Step 1: Identify distinct values

$$x'_1 = 2,\ x'_2 = 3,\ x'_3 = 4,\ x'_4 = 5$$

Step 2: Count frequencies

Absolute Frequencies

$$n_1 = 4,\ n_2 = 5,\ n_3 = 4,\ n_4 = 2$$

Step 3: Compute relative frequencies

Relative Frequencies

$$f_j = \frac{n_j}{n} \Rightarrow f_1 = 0.267, f_2 = 0.333, f_3 = 0.267, f_4 = 0.133$$

Step 4: Determine range

$$x_{\min} = 2, \quad x_{\max} = 5, \quad \text{Range} \quad = 3$$

FREQUENCY TABLE FOR DISCRETE VARIABLE: EXAMPLE

Step 5: Table (Frequency

Absolute Frequencies

Relative Frequencies

x'_j	n_j	f_j
2	4	0.267
3	5	0.333
4	4	0.267
5	2	0.133
Total	**15**	**1.0**

Sample Size
n = 15

$\sum_{j=1}^{m} n_j = n$ $\sum_{j=1}^{m} f_j = 1$

FREQUENCY TABLE FOR CONTINUOUS VARIABLE: EXAMPLE

Suppose we collected $n = 12$ observations of a continuous variable x:

$$x = (1.2, 2.3, 1.8, 2.0, 3.1, 2.7, 3.5, 1.5, 2.8, 3.0, 1.9, 2.5)$$

Step 1: Determine range and class width

$$x_{\text{min}} = 1.2, \quad x_{\text{max}} = 3.5, \quad \text{Range} \qquad = 3.5 - 1.2 = 2.3$$

Using **Sturges' rule**, $m = 4$ classes:

$$\text{Class width} = \frac{\text{Range}}{m} = \frac{2.3}{4} \approx 0.575 \approx 0.6$$

Step 2: Construct classes

Classes

Start at 1.2:

$$I_j = \left]l_{j-1},\ l_j\right] \quad j = 1,2,\dots,m$$

m = number of classes

1. $[1.2, 1.8]$
2. $]1.8, 2.4]$
3. $]2.4, 3.0]$
4. $]3.0, 3.6]$

$$I_j \cap I_k = \emptyset \quad \text{and}$$

$$D \subset \cup_{j=1}^{m} I_j$$

$$D = [x_{min}, x_{max}]$$

Step 3: Find midpoints

$$x_1' = \frac{1.2 + 1.8}{2} = 1.5, \quad x_2' = \frac{1.8 + 2.4}{2} = 2.1, \quad x_3' = \frac{2.4 + 3.0}{2} = 2.7, \quad x_4' = \frac{3.0 + 3.6}{2} = 3.3$$

FREQUENCY TABLE FOR CONTINUOUS VARIABLE: EXAMPLE

Step 4: Count frequencies

Absolute Frequencies

- Class [1.2, 1.8]: 1.2, 1.5, 1.8 → $n_1 = 3$
- Class]1.8, 2.4]: 1.9, 2.0, 2.3 → $n_2 = 3$
- Class]2.4, 3.0]: 2.5, 2.7, 3.0 → $n_3 = 3$
- Class]3.0, 3.6]: 3.1, 3.5 → $n_4 = 2$

Step 5: Compute relative frequencies

Relative Frequencies

$$f_1 = \frac{3}{12} = 0.25, \quad f_2 = 0.25, \quad f_3 = 0.25, \quad f_4 = \frac{2}{12} \approx 0.167$$

Step 6: Table (Frequency Distribution with Totals)

Class I_j	Midpoint x'_j	n_j	f_j
[1.2, 1.8]	1.5	3	0.25
]1.8, 2.4]	2.1	3	0.25
]2.4, 3.0]	2.7	3	0.25
]3.0, 3.6]	3.3	2	0.167
Total	—	**12**	**1.0**

Sample Size
n = 12

$\sum_{j=1}^{m} n_j = n$ $\sum_{j=1}^{m} f_j = 1$

LECTURE 4: CUMULATIVE FREQUENCY FUNCTIONS

CUMULATIVE FREQUENCY FUNCTION FOR DISCRETE VARIABLE

Cumulative Absolute Frequency Function

$$N(x) = \begin{cases} 0 & x < x'_1 \\ n_1 & x'_1 \le x < x'_2 \\ n_1 + n_2 & x'_2 \le x < x'_3 \\ \dots & \dots \\ n_1 + n_2 + \dots + n_{m-1} & x'_{m-1} \le x < x'_m \\ n_1 + n_2 + \dots + n_m = n & x \ge x'_m \end{cases}$$

- N(x) is a function whose domain is the real line, with $N(-\infty) = 0, N(+\infty) = \text{n}$ and $0 \le N(x) \le n$.
- It is a **non-decreasing function** and **right-continuous**.
- It is a **step function** where each jump corresponds to an absolute frequency. For exemple: $N(x'_3) - N(x'_2)$ $= (n_1 + n_2 + n_3) - (n_1 + n_2) = n_3$

Cumulative Relative Frequency Function

$$F^*(x) = \begin{cases} 0 & x < x'_1 \\ f_1 & x'_1 \le x < x'_2 \\ f_1 + f_2 & x'_2 \le x < x'_3 \\ \dots & \dots \\ f_1 + f_2 + \dots + f_{m-1} & x'_{m-1} \le x < x'_m \\ f_1 + f_2 + \dots + f_m = 1 & x \ge x'_m \end{cases}$$

- F*(x) has properties similar to $N(x)$,except with respect to the codomain: $0 \le F^*_x(\text{x}) \le 1$.

CUMULATIVE FREQUENCY FUNCTION FOR DISCRETE VARIABLE: EXAMPLE

Sample information:

- Observations: $x = (2, 3, 3, 4, 2, 5, 3, 4, 2, 5, 4, 3, 2, 4, 3)$
- Number of observations: $n = 15$
- Sample range: $\text{Sample range} = x_{\max} - x_{\min} = 5 - 2 = 3$

Frequency Distribution Table (with cumulative frequencies)

x'_j	n_j	f_j	$N(x'_j)$	$F^*(x'_j)$
2	4	0.267	4	0.267
3	5	0.333	9	0.600
4	4	0.267	13	0.867
5	2	0.133	15	1.0
Total	**15**	**1.0**	—	—

Cumulative Absolute Frequency Function

$$N(x) = \begin{cases} 0, & x < 2 \\ 4, & 2 \leq x < 3 \\ 9, & 3 \leq x < 4 \\ 13, & 4 \leq x < 5 \\ 15, & x \geq 5 \end{cases}$$

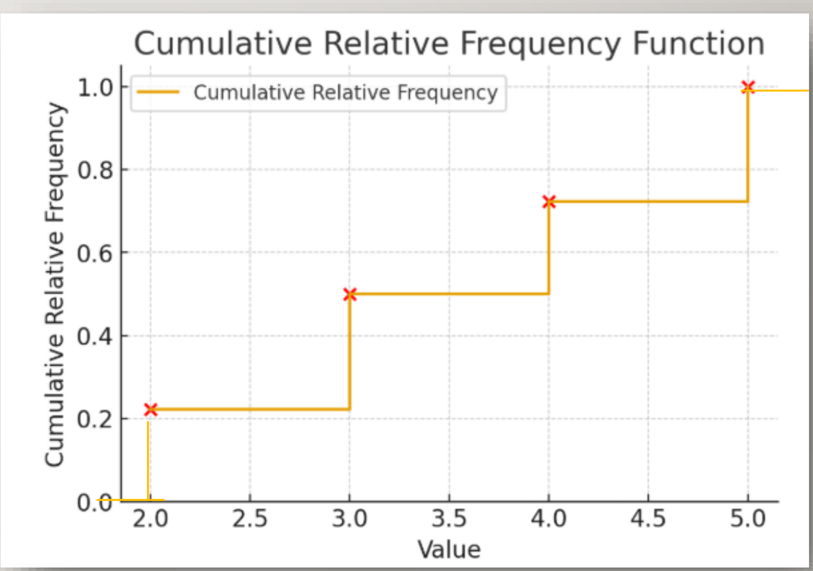
Cumulative Relative Frequency Function

$$F^*(x) = \begin{cases} 0, & x < 2 \\ 0.267, & 2 \leq x < 3 \\ 0.600, & 3 \leq x < 4 \\ 0.867, & 4 \leq x < 5 \\ 1, & x \geq 5 \end{cases}$$

CUMULATIVE FREQUENCY GRAPH FOR A DISCRETE VARIABLE: EXAMPLE

Cumulative Relative Frequency Function

$$F^*(x) = \begin{cases} 0, & x < 2 \\ 0.267, & 2 \le x < 3 \\ 0.600, & 3 \le x < 4 \\ 0.867, & 4 \le x < 5 \\ 1, & x \ge 5 \end{cases}$$



- Here is the graph of the **cumulative relative frequency function** (ogive) for your sample:
  Values: 2, 3, 4, 5
  Frequencies: 4, 5, 4, 5
- The red dots mark the cumulative proportions.
- This shows how the cumulative relative frequency increases step by step until it reaches **1** (100%).

CUMULATIVE FREQUENCY FUNCTION FOR CONTINUOUS VARIABLE

Cumulative Relative Frequency Function

$$
\begin{aligned}
&F^*(l_0) = 0 \\
&F^*(l_1) = f_1 \\
&F^*(l_2) = f_1 + f_2 \\
&\ldots \\
&F^*(l_{m-1}) = f_1 + f_2 + \ldots + f_{m-1} \\
&F^*(l_m) = f_1 + f_2 + \ldots + f_m = 1
\end{aligned}
$$

It is assumed that the frequencies are **uniformly distributed within each class** and that the cumulative frequencies refer to the **class boundaries**.

FREQUENCY CUMULATIVE FUNCTION FOR CONTINUOUS VARIABLE: EXAMPLE

Sample information:

- Observations: $x = (1.2, 2.3, 1.8, 2.0, 3.1, 2.7, 3.5, 1.5, 2.8, 3.0, 1.9, 2.5)$
- Number of observations: $n = 12$
- Sample range: $\text{Sample range} = x_{\max} - x_{\min} = 3.5 - 1.2 = 2.3$
- Classes (Sturges, width ≈ 0.6): [1.2,1.8],]1.8,2.4],]2.4,3.0],]3.0,3.6]
- Midpoints: 1.5, 2.1, 2.7, 3.3

Frequency Distribution Table (with cumulative frequencies)

Class	l_j	n_j	f_j	$N(x_j)$	$F^*(x_j)$
[1.2, 1.8]	1.2	3	0.25	3	0.25
]1.8, 2.4]	1.8	3	0.25	6	0.50
]2.4, 3.0]	2.4	3	0.25	9	0.75
]3.0, 3.6]	3.0	3	0.25	12	1.00
Total	—	12	1.00	—	—

Cumulative Absolute Frequency Function

$$N(x) = \begin{cases} 0, & x < 1.2 \\ 3, & 1.2 \le x < 1.8 \\ 6, & 1.8 \le x < 2.4 \\ 9, & 2.4 \le x < 3.0 \\ 12, & x \ge 3.0 \end{cases}$$

Cumulative Relative Frequency Function

$$F^*(x) = \begin{cases} 0, & x < 1.2 \\ 0.25, & 1.2 \le x < 1.8 \\ 0.50, & 1.8 \le x < 2.4 \\ 0.75, & 2.4 \le x < 3.0 \\ 1.0, & x \ge 3.0 \end{cases}$$

CUMULATIVE FREQUENCY GRAPH FOR A CONTINUOUS VARIABLE: EXAMPLE

Class	l_j	n_j	f_j	$N(x_j)$	$F^*(x_j)$
[1.2, 1.8]	1.2	3	0.25	3	0.25
]1.8, 2.4]	1.8	3	0.25	6	0.50
]2.4, 3.0]	2.4	3	0.25	9	0.75
]3.0, 3.6]	3.0	3	0.25	12	1.00
Total	—	12	1.00	—	—

Interval Endpoints	F*(x)
1,2	0
1,8	0,25
2,4	0,5
3	0,75
3,6	1

Ogive (Cumulative Frequency Graph)

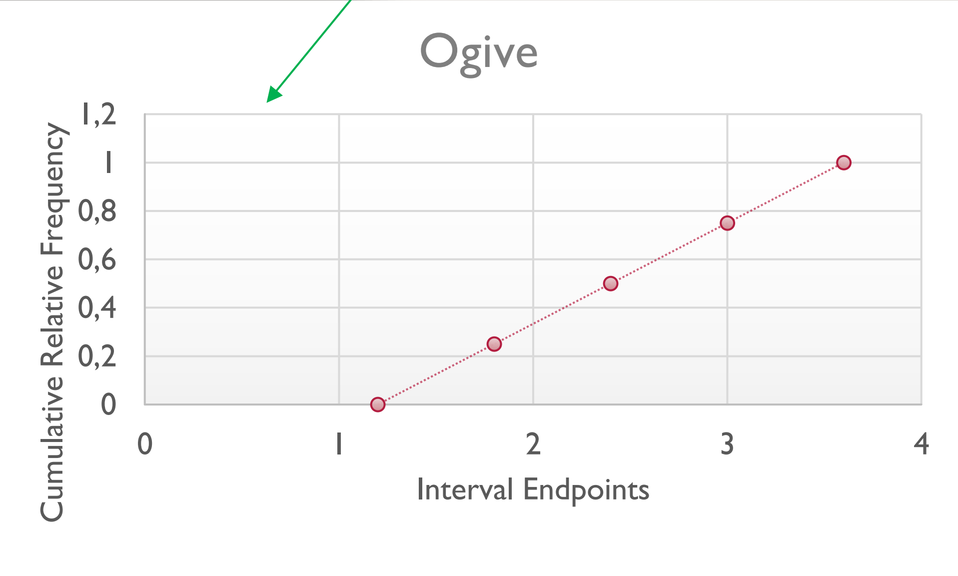
Where does it start?

- It starts at the **lower limit of the first class**, with cumulative frequency = 0.
- The graph is built using the upper class limits on the x-axis and the cumulative frequencies (or percentages) on the y-axis.

What is it used for?

- To show how data accumulate across classes.
- To identify the median, quartiles, deciles, and percentiles.
- To compare cumulative distributions.

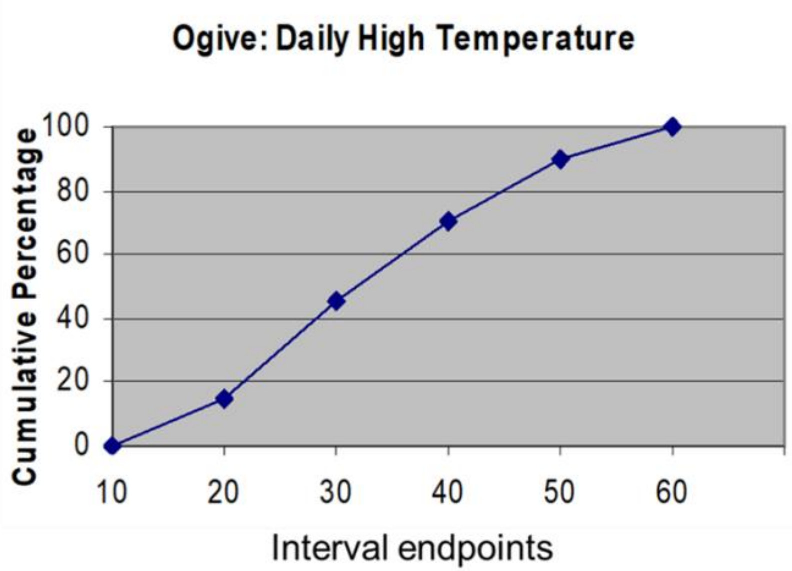
In this case, the ogive appears as a straight line rather than a curve, because the absolute and relative frequencies have the same values for all classes.



THE OGIVE GRAPHING CUMULATIVE FREQUENCIES: OTHER EXAMPLE

Cumulative Relative Frequency Function

Interval	Upper interval endpoint	Cumulative Percentage
Less than 10	10	0
10 but less than 20	20	15
20 but less than 30	30	45
30 but less than 40	40	70
40 but less than 50	50	90
50 but less than 60	60	100



Newbold et al (2013)

LECTURE 4: MEASURES OF LOCATION AND POSITION

MEASURES OF LOCATION VS. MEASURES OF POSITION

Measures of Location

- Describe **where the data are concentrated**.
- Give an idea of the "center" or typical value of the data.
- Examples:
  - **Mean** – sum of values divided by the number of observations.
  - **Median** – value that separates the data into two equal halves.
  - **Mode** – most frequently occurring value.
- **Summary:** indicate "on average, where the data are."

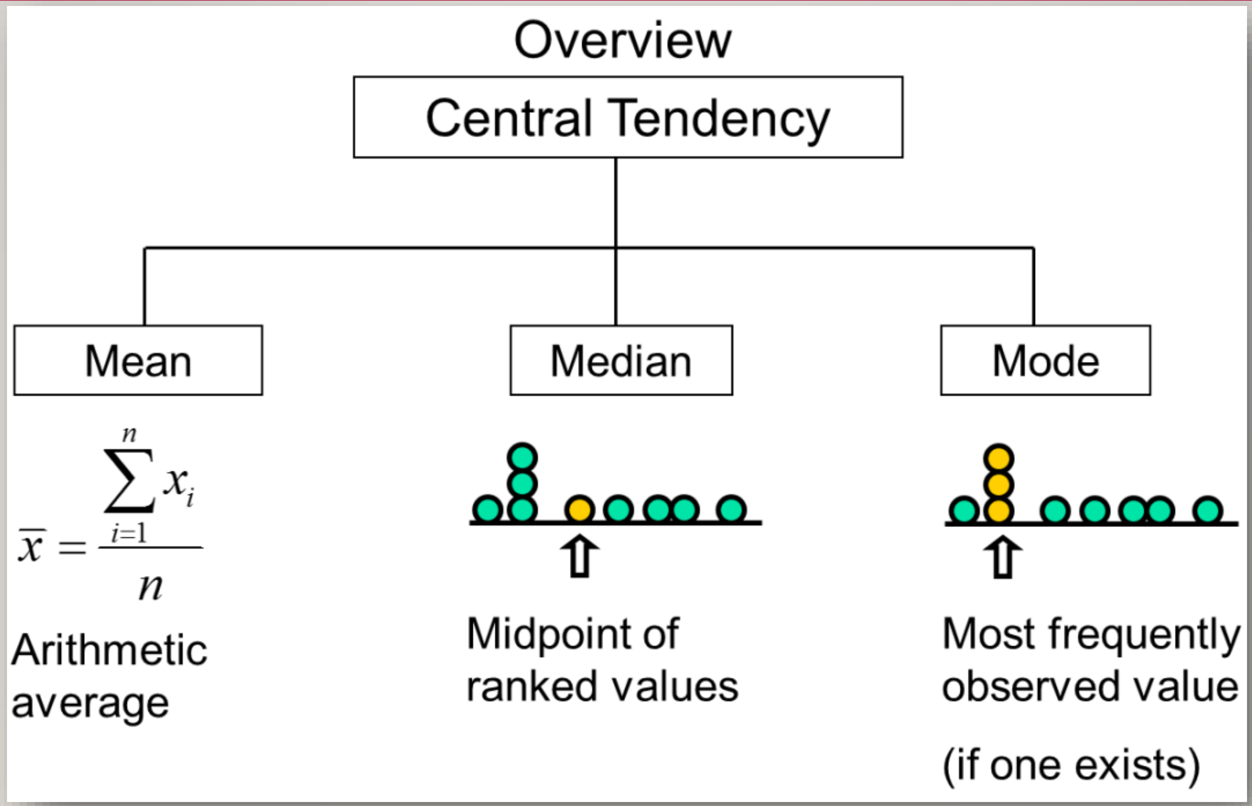
Measures of Position

- Indicate **the relative position of a value within the dataset** or **the position of reference values**.
- Useful for comparing values and identifying quantiles.
- Examples:
  - **Quartiles, Deciles, Percentiles** – divide the data into equal parts and show relative positions.
  - **Z-score** – shows how many standard deviations a value is above or below the mean.
- **Summary:** indicate "where a value stands in relation to the whole dataset."

Key Difference:

- **Location:** describes the center or typical value of the dataset.
- **Position:** describes the relative position of a value within the dataset.

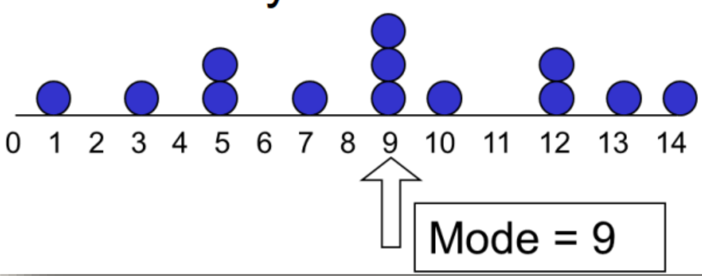
MEASURES OF LOCATION

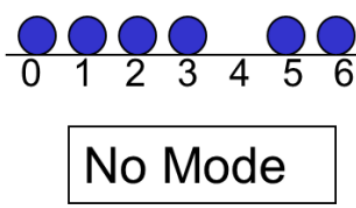


Newbold et al (2013)

MODE

- A measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes





Newbold et al (2013)

DISTRIBUTIONS BY MODE

1 Amodal Distribution (No Mode)

- **Definition:** No value repeats → no mode
- **Example:** 2, 3, 5, 7, 11

2 Unimodal Distribution (One Mode)

- **Definition:** One value appears most frequently
- **Example:** 1, 2, 2, 3, 4 → **Mode = 2**

3 Bimodal Distribution (Two Modes)

- **Definition:** Two values appear with the same highest frequency
- **Example:** 1, 2, 2, 3, 3, 4 → **Modes = 2, 3**

4 Multimodal Distribution (More than Two Modes)

- **Definition:** More than two values appear with the same highest frequency
- **Example:** 1, 1, 2, 2, 3, 3, 4 → **Modes = 1, 2, 3**

MODE FOR GROUPED DATA

Definition:

The **modal class** is the class interval with the **highest frequency** in a grouped frequency distribution

King's Formula (Grouped Data):

$$\boxed{\mathrm{Mo} = l + \frac{f^{**}}{f^{*} + f^{**}} \cdot h}$$

Legend / Notation:

- l = lower boundary of the modal class
- h = class width
- f^{*} = **relative frequency of the class before the modal class**
- f^{**} = **relative frequency of the class after the modal class**

Silvestre (2007)

MODE FOR GROUPED DATA: EXAMPLE

Relative Frequencies

Class	Rel. Freq.
0–10	0.10
10–20	0.16
20–30	0.24
30–40	0.14
40–50	0.06

Modal class is the class interval with the highest absolute or relative frequency.

Modal Class: 20 - 30

Step 1: Identify the modal class → 20–30

Step 2: Extract values for the formula:

- $l = 20$
- $h = 10$
- $f^* = 0.16$
- $f^{**} = 0.14$

Step 3: Apply King's formula:

$$\boxed{\mathrm{Mo} = 20 + \frac{0.14}{0.16 + 0.14} \cdot 10 = 20 + \frac{0.14}{0.30} \cdot 10 = 20 + 4.67}$$

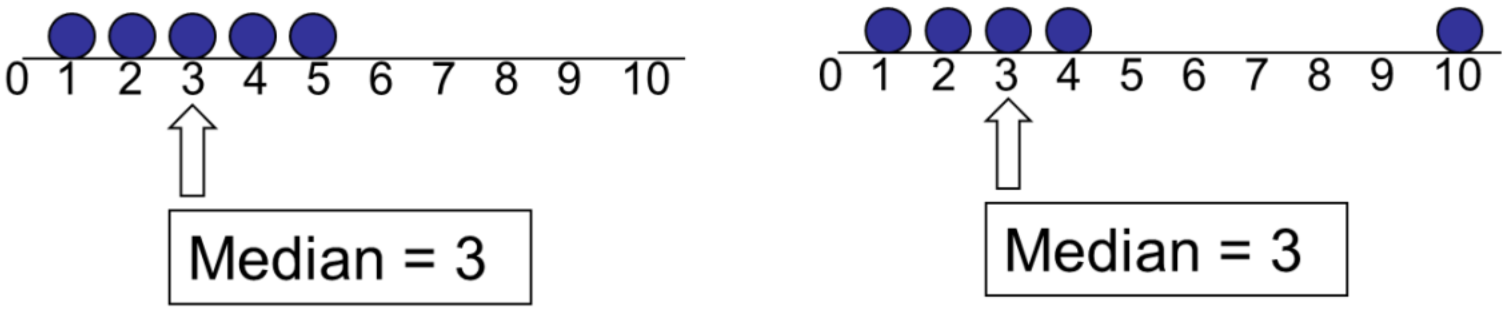
Step 4: Result

$$\text{Mode} \approx 24.67$$

✅ **Mode ≈ 24.67**

MEDIAN

- In an ordered list, the median is the “middle” number (50% above, 50% below)



- Not affected by extreme values

Newbold et al (2013)

FINDING THE MEDIAN

- The location of the median:

$$\boxed{\text{Median position} = \left(\frac{n+1}{2}\right)^{\text{th}}} \text{ position in the ordered data}$$

  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- Note that $\frac{n+1}{2}$ is not the value of the median, only the position of the median in the ranked data

Newbold et al (2013)

CALCULATING THE MEDIAN: EXAMPLES

Formula to Find Median Position

$$\boxed{\text{Position} = \frac{n+1}{2}}$$

- n = number of observations
- If **Position is integer** → Median = value at that position
- If **Position is not integer** → Median = average of values at floor and ceil(Position)

Examples

Example 1 – Position is integer

- Data: 2, 4, 6, 8, 10 ($n = 5$)

$$\text{Position} = \frac{5+1}{2} = 3 \quad \text{(integer)}$$

$$\text{Median} = \text{3rd value} = 6$$

Example 2 – Position is not integer

- Data: 3, 5, 8, 12, 15, 18 ($n = 6$)

$$\text{Position} = \frac{6+1}{2} = 3.5 \quad \text{(not integer)}$$

$$\text{Median} = \frac{\text{3rd value} + \text{4th value}}{2} = \frac{8+12}{2} = 10$$

MEDIAN FOR GROUPED DATA

Definition:

The **median class** is the class interval that contains the **middle value** (0.5 in cumulative relative frequency) of the distribution.

Formula (Grouped Data

$$\boxed{\mathrm{Me} = l + \frac{0.5 - F^*(l)}{F^*(L) - F^*(l)} \cdot h}$$

Legend / Notation:

- l = lower boundary of the median class
- L = upper boundary of the median class
- $h = L - l$ = class width
- $F^*(l) =$ cumulative relative frequency **before the median class**
- $F^*(L) =$ cumulative relative frequency **at the upper boundary of the median class**

Silvestre (2007)

MEDIAN FOR GROUPED DATA: EXAMPLE

Relative Frequencies

Cumulative Relative Frequencies

Median Class: 20 - 30

Class	Relative Frequency	Cumulative Rel. Freq.
0–10	0.10	0.10
10–20	0.16	0.26
20–30	0.24	0.50
30–40	0.14	0.64
40–50	0.06	0.70

The median class is the first class for which the cumulative relative frequency is equal to or greater than 0.5.

Step 1: Identify the median class → 20–30

Step 2: Extract values for the formula:

- $l = 20$, $L = 30$, $h = 10$
- $F^*(l) = 0.26$ (cumulative relative frequency before the median class)
- $F^*(L) = 0.50$ (cumulative relative frequency at the upper boundary of the median class)

Step 3: Apply formula:

$$\boxed{\mathrm{Me} = 20 + \frac{0.5 - 0.26}{0.50 - 0.26} \cdot 10 = 20 + \frac{0.24}{0.24} \cdot 10 = 20 + 10}$$

Step 4: Result

$$\text{Median} \approx 30$$

✅ **Median ≈ 30**

ARITHMETIC MEAN

The arithmetic mean (mean) is the most common measure of central tendency

– For a population of *N* values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Population values

Population size

– For a sample of size *n*:

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

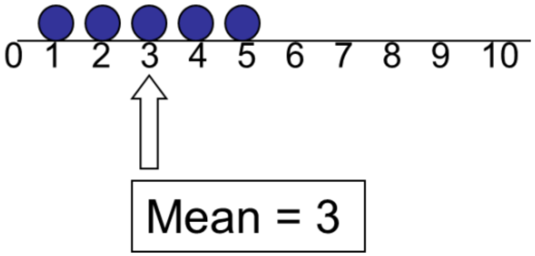
Observed values

Sample size

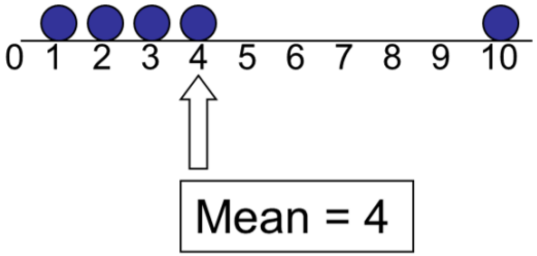
Newbold et al (2013)

ARITHMETIC MEAN: EXAMPLES

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Newbold et al (2013)

TYPES OF MEANS

◆ **Harmonic Mean**

For a sample of n positive values $x_1, x_2, \ldots, x_n$:

$$\boxed{H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}}$$

👉 It gives more weight to **smaller values.**

Typical use: averages of **rates** or **speeds**.

◆ **Geometric Mean**

For a sample of n positive values $x_1, x_2, \ldots, x_n$:

$$\boxed{G = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}}$$

👉 Often used in **population growth**, **compound interest**, or **growth rates**.

⚖️ **General comparison between means:**

$$\boxed{H \leq G \leq A}$$

where A is the **arithmetic mean**.

WHICH MEASURE OF LOCATION IS THE "BEST"?

- **Mean** is generally used, unless extreme values (outliers) exist …
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

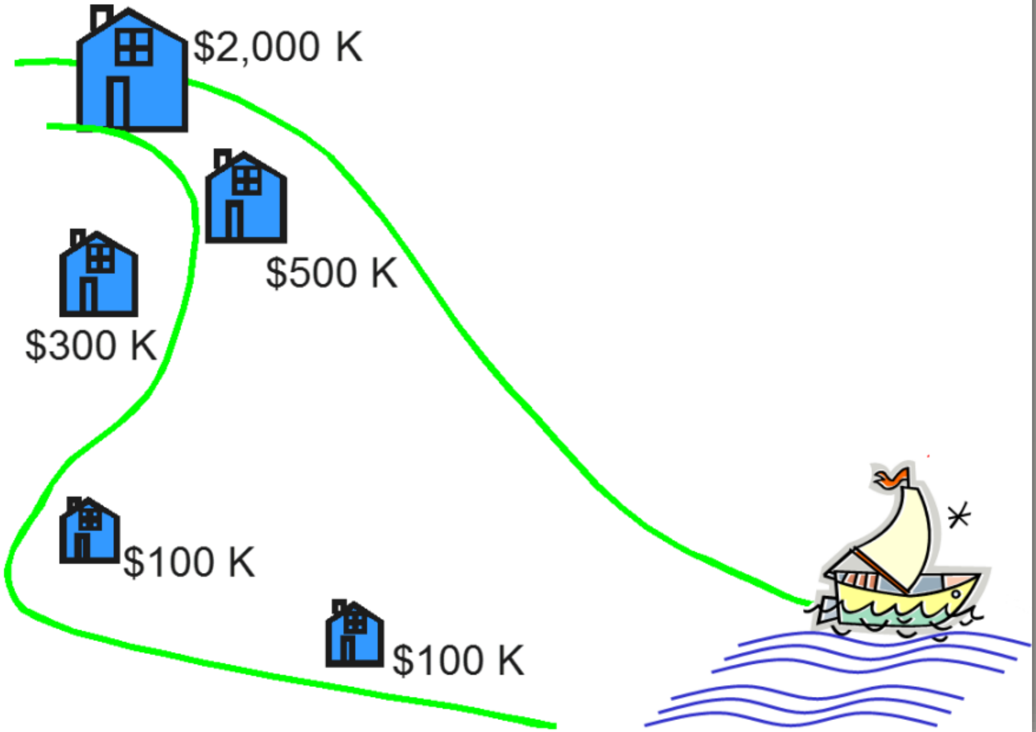
Newbold et al (2013)

REVIEW EXAMPLE

- Five houses on a hill by the beach

House Prices:

$2,000,000
500,000
300,000
100,000
100,000



Newbold et al (2013)

REVIEW EXAMPLE: SUMMARY STATISTICS

House Prices:

$$
\begin{array}{r}
\$2{,}000{,}000 \\
500{,}000 \\
300{,}000 \\
100{,}000 \\
100{,}000 \\
\hline
\text{Sum } 3{,}000{,}000
\end{array}
$$

- **Mean:** $\left(\dfrac{\$3{,}000{,}000}{5}\right)$

  = **$600,000**

- **Median:** middle value of ranked data

  = **$300,000**

- **Mode:** most frequent value

  = **$100,000**

Newbold et al (2013)

WEIGHTED MEAN

Weighted Mean

The **weighted mean** of a set of data is

$$\boxed{\bar{x} = \frac{\sum w_i x_i}{n}}$$

where w_i = weight of the ith observation and $n = \sum w_i$.

Newbold et al (2013)

WEIGHTED MEAN: EXAMPLE

Example 2.17 Stock Recommendation (Weighted Mean)

Zack's Investment Research is a leading investment research firm. Zack's will make one of the following recommendations with corresponding weights for a given stock: Strong Buy (1), Moderate Buy (2), Hold (3), Moderate Sell (4), or Strong Sell (5). Suppose that on a particular day, 10 analysts recommend Strong Buy, 3 analysts recommend Moderate Buy, and 6 analysts recommend Hold for a particular stock. Based on Zack's weights, find the mean recommendation.

Solution Table 2.8 shows the weights for each recommendation and the computation leading to a recommendation based on the following weighted mean recommendation conversion values: if the weighted mean is 1, Strong Buy; 1.1 through 2.0, Moderate Buy; 2.1 through 3.0, Hold; 3.1 through 4.0, Moderate Sell; 4.1 through 5, Strong Sell.

Table 2.8 Computation of Zack's Investment Research's Average Brokerage Recommendation

Action	Number of Analysts, w_i	Value, x_i	$w_i x_i$
Strong Buy	10	1	10
Moderate Buy	3	2	6
Hold	6	3	18
Moderate Sell	0	4	0
Strong Sell	0	5	0

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{n} = \frac{10 + 6 + 18 + 0 + 0}{19} = 1.79}$$

The weighted mean of 1.79 yielded a Moderate Buy recommendation.

Newbold et al (2013)

MEAN FOR GROUPED DATA

Definition:

The **mean** for grouped data is the weighted average of the class midpoints, weighted by the frequencies.

Formula (Grouped Data – Silvestre, 2007):

$$\boxed{\bar{x} = \frac{\sum_{j=1}^{m} n_j \cdot x'_j}{n} = \sum_{j=1}^{m} f_j \cdot x'_j}$$

Legend / Notation:

- $j = 1, 2, ..., m$ → class index
- n_j = absolute frequency of class j
- f_j = relative frequency of class j
- x'_j = midpoint of class j
- m = number of classes
- $n = \sum_{j=1}^{m} n_j$ → total number of observations

Silvestre (2007)

MEAN FOR GROUPED DATA: EXAMPLE

Absolute Frequencies

Relative Frequencies

Example Calculation:

Class Interval	x'_j (Midpoint)	n_j	f_j
0–10	5	4	0.20
10–20	15	6	0.30
20–30	25	5	0.25
30–40	35	5	0.25

Step 1: Using **absolute frequencies**:

$$\boxed{\bar{x} = \frac{(4 \cdot 5) + (6 \cdot 15) + (5 \cdot 25) + (5 \cdot 35)}{4 + 6 + 5 + 5} = \frac{410}{20} = 20.5}$$

Step 2: Using **relative frequencies**:

$$\boxed{\bar{x} = (0.20 \cdot 5) + (0.30 \cdot 15) + (0.25 \cdot 25) + (0.25 \cdot 35) = 1 + 4.5 + 6.25 + 8.75 = 20.5}$$

✅ **Mean ≈ 20.5**

PROPORTIONS: QUANTITIES AND FREQUENCIES

Key Idea:

- A **proportion** represents the **part of the total** corresponding to a category.
- **As a probability:** A proportion can be interpreted as the **likelihood of selecting an observation from that category**, ranging from 0 to 1.
- **As a percentage:** A proportion can be expressed as a **percentage of the total**, ranging from 0% to 100%

Formulas:

1. **Proportion / Probability:**

$$\boxed{f_j = \frac{n_j}{n}}$$

2. **Percentage:**

$$\boxed{\text{Percentage} = f_j \cdot 100}$$

Key Points:
Proportions **sum to 1** → total probability = 1
Percentages **sum to 100%**

Example:

Category	Quantity (n_j)	Proportion / Probability (f_j)	Percentage
A	10	0.25	25%
B	20	0.50	50%
C	10	0.25	25%

FOUR PROPERTIES OF THE MEAN

Note: The notation used here follows **Silvestre (2007)**.

Let $m(x)$ or $\bar{x}$ be the mean of variable x.

1. Addition / Subtraction of a Constant:

- Adding a constant c to all values:

$$\boxed{m(x+c) = m(x) + c}$$

- Subtracting a constant c from all values:

$$\boxed{m(x-c) = m(x) - c}$$

2. Multiplication / Division by a Constant:

- Multiplying all values by a constant c:

$$\boxed{m(c \cdot x) = c \cdot m(x)}$$

- Dividing all values by a constant c:

$$\boxed{m(x/c) = m(x)/c}$$

Silvestre (2007)

FOUR PROPERTIES OF THE MEAN

Note: The notation used here follows **Silvestre (2007)**.

3. Mean of Deviations is Zero:

$$\boxed{m(x - \bar{x}) = m(x) - \bar{x} = 0}$$

- The mean is the **balance point** of the data.

4. Mean of Grouped Values:

- If the data are divided into groups $G_1, G_2, \ldots, G_k$ with group means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ and sizes $n_1, n_2, \ldots, n_k$:

$$\boxed{\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k}}$$

Silvestre (2007)

QUANTILES: DEFINITION (REVIEW)

1️⃣ What are Quantiles?

- Quantiles are **values that divide a dataset into equal parts**.
- Special cases:
  - **Quartiles** → Q1, Q2, Q3, Q4 (divide data into 4 equal parts)
    - **Median = Q2**
  - **Deciles** → D1, D2, ..., D10 (divide data into 10 equal parts)
    - **Median = D5**
  - **Percentiles** → P1, P2, ..., P100 (divide data into 100 equal parts)
    - **Median = P50**

QUANTILES FOR GROUPED DATA

$$q_p = l + \frac{p - F^*(l)}{F^*(L) - F^*(l)} \cdot h$$

Silvestre (2007)

Legend / Notes:

- $q_p =$ p-th quantile of grouped data
- $l =$ lower class boundary of the class containing q_p
- $L =$ upper class boundary of that class
- $F^*(l) =$ cumulative relative frequency at l
- $F^*(L) =$ cumulative relative frequency at L
- $p =$ quantile proportion (e.g., 0.25 for Q1)
- $h = L - l =$ class width of the class containing the quantile

QUANTILE FOR GROUPED DATA: EXAMPLE

Suppose we have the following **frequency distribution** of exam scores:

Score Interval	Frequency	Relative Frequencies	Cumulative Relative Frequencies
0,10 ↗	2	0.10	0.10
(10,20]	5	0.25	0.35
(20,30]	8	0.40	0.75
(30,40]	4	0.20	0.95
(40,50]	1	0.05	1.00

Question: Find the **1st quartile (Q1)**.

Step 1: Identify the class containing Q1

- Q1 corresponds to the **25th percentile**, so $p = 0.25$.
- Look at the cumulative relative frequencies:
  -
  - (10,20]: 0.35 ✅

→ Q1 lies in the **(10,20]** class.

QUANTILE FOR GROUPED DATA: EXAMPLE

Suppose we have the following **frequency distribution** of exam scores:

Score Interval	Frequency	Relative Frequency	Cumulative Relative Frequen
0,10 ↗	2	0.10	0.10
(10,20]	5	0.25	0.35
(20,30]	8	0.40	0.75
(30,40]	4	0.20	0.95
(40,50]	1	0.05	1.00

Question: Find the **1st quartile (Q1)**.

Step 2: Apply the formula

$$q_p = l + \frac{p - F^*(l)}{F^*(L) - F^*(l)} \cdot h$$

Where:

- $l = 10$ (lower boundary of the class)
- $L = 20$ (upper boundary)
- $F^*(l) = 0.10$
- $F^*(L) = 0.35$
- $p = 0.25$
- $h = L - l = 10$

$$Q1 = 10 + \frac{0.25 - 0.10}{0.35 - 0.10} \cdot 10$$

Step 3: Calculate

$$Q1 = 10 + \frac{0.15}{0.25} \cdot 10 = 10 + 0.6 \cdot 10 = 10 + 6 = 16$$

LECTURE 4: MEASURES OF DISPERSION AND CONCENTRATION

MEASURES OF DISPERSION AND CONCENTRATION

1. Measures of Absolute Dispersion – describe the **spread of the data in absolute terms:**

- Range (*Amplitude de variação*)
- Interquartile Range (IQR) (*Amplitude inter-quartis*)
- Mean Absolute Deviation (MAD) (*Desvio médio absoluto*)
- Variance and Standard Deviation
  - For **ungrouped (raw) data**
  - For **grouped (classified) data**

---

2. Measures of Relative Dispersion – describe **spread relative to a central value**:

- Coefficient of Variation (CV)
- Quartile-based measure: (Q3 – Q1) / Median

---

3. Measures of Concentration – describe **how data are focused or unequal**:

- Gini Index

MEASURES OF ABSOLUTE DISPERSION

1. Range (Amplitude de variação)

$$\boxed{\text{Range} = \max(x_i) - \min(x_i)}$$

2. Interquartile Range (IQR) – Amplitude inter-quartis

$$\boxed{\text{IQR} = Q3 - Q1}$$

3. Mean Absolute Deviation (MAD) – Desvio médio absoluto (only for ungrouped data)

$$\boxed{d = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}}$$

4. Variance – Variância

- Ungrouped (raw) data:

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

$$\boxed{s = \sqrt{s^2}} \text{ (standard deviation)}$$

- Grouped (classified) data:

$$\boxed{s^2 = \frac{\sum_{j=1}^{m} n_j(x'_j - \bar{x})^2}{n} = \sum_{j=1}^{m} f_j(x'_j - \bar{x})^2 = \frac{\sum_{j=1}^{m} n_j(x'_j)^2}{n} - \bar{x}^2}$$

$$\boxed{s = \sqrt{s^2}} \text{ (standard deviation)}$$

💡 Notes:

- x_i = individual data points (ungrouped)
- x'_j = class midpoint (grouped)
- $\bar{x}$ = mean of the data
- n_j = absolute frequency of the j-th class
- f_j = relative frequency of the j-th class
- n = total number of observations

ABSOLUTE DISPERSION MEASURES (UNGROUPED DATA): EXAMPLE

Data (scores):

$$12, 18, 20, 22, 25, 28, 30, 35, 38, 40$$

1. Range

$$\boxed{\text{Range} = \max(x_i) - \min(x_i) = 40 - 12 = 28}$$

2. Interquartile Range (IQR) – using position formula and interpolation

- $n = 10$
- Q1 position: $(n+1) \cdot 0.25 = 2.75 \Rightarrow Q1 = 0.25 \cdot 18 + 0.75 \cdot 20 = 19.5$
- Q3 position: $(n+1) \cdot 0.75 = 8.25 \Rightarrow Q3 = 0.75 \cdot 35 + 0.25 \cdot 38 = 35.75$

$$\boxed{\text{IQR} = Q3 - Q1 = 16.25}$$

3. Mean Absolute Deviation (MAD)

$$\bar{x} = 26.8, \quad \boxed{d = \frac{\sum_{i=1}^{10} |x_i - \bar{x}|}{10} \approx 7.3}$$

4. Variance

$$\boxed{s^2 = \frac{(12 - 26.8)^2 + \cdots + (40 - 26.8)^2}{10} \approx 74.68}$$

5. Standard Deviation

$$\boxed{s = \sqrt{s^2} \approx 8.64} \quad \text{(standard deviation)}$$

ABSOLUTE DISPERSION MEASURES (GROUPED DATA): EXAMPLE

Data: Monthly Expenses of 20 Families ($)

Class Interval	Midpoint x'_j	Frequency n_j
[0, 100]	50	2
]100, 200]	150	5
]200, 300]	250	8
]300, 400]	350	4
]400, 500]	450	1

Step 1: Total observations

$$n = 2 + 5 + 8 + 4 + 1 = 20$$

Step 2: Range

$$\boxed{\text{Range} = \max(\text{upper limit}) - \min(\text{lower limit}) = 500 - 0 = 500}$$

ABSOLUTE DISPERSION MEASURES (GROUPED DATA): EXAMPLE

Data: Monthly Expenses of 20 Families ($)

Absolute Frequencies

Class Interval	Midpoint x'_j	Frequency n_j
[0, 100]	50	2
]100, 200]	150	5
]200, 300]	250	8
]300, 400]	350	4
]400, 500]	450	1

Step 1: Total observations

$$n = 2 + 5 + 8 + 4 + 1 = 20$$

Step 2: Mean

$$\bar{x} = \frac{\sum n_j x'_j}{n} = 235$$

Step 3: Variance

$$\boxed{s^2 = \frac{(50 - 235)^2 \cdot 2 + \cdots + (450 - 235)^2 \cdot 1}{20} \approx 10750}$$

Step 4: Standard Deviation

$$\boxed{s = \sqrt{s^2} \approx 103.7}$$

PROPERTIES OF VARIANCE

1. Variance of constants and shifts

$$\boxed{v(c) = 0, \quad v(x+c) = v(x), \quad v(x-c) = v(x)}$$

- Adding or subtracting a constant does **not change variance**.

2. Variance of scaled variables

$$\boxed{v(cx) = c^2 v(x), \quad v\left(\frac{x}{c}\right) = \frac{v(x)}{c^2}, \quad c \neq 0}$$

- Scaling by a factor c multiplies variance by c^2.

3. Variance decomposition for grouped data

- Suppose n observations divided into k groups $g_1, g_2, \ldots, g_k$ with $n_1, n_2, \ldots, n_k$ elements, such that $n_1 + n_2 + \cdots + n_k = n$.
- Let $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ be the group means and $s_1^2, s_2^2, \ldots, s_k^2$ the group variances.

$$\boxed{v_{\text{total}} = \underbrace{\sum_{j=1}^{k} \frac{n_j}{n} s_j^2}_{\text{within-group}} + \underbrace{\sum_{j=1}^{k} \frac{n_j}{n} (\bar{x}_j - \bar{x})^2}_{\text{between-group}}}$$

💡 **Note:** This decomposition shows how total variability is explained by **variability within groups** and **variability between groups**.

Silvestre (2007)

Note: The notation used here follows **Silvestre (2007)**.

RELATIVE DISPERSION MEASURES

1. Coefficient of Variation (CV)

$$\boxed{CV = \frac{s}{\bar{x}} \times 100\%}$$

- Measures **relative variability** compared to the mean.
- Useful to compare variability across datasets with different units or scales.

2. Interquartile Range Ratio (IQR / Median)

$$\boxed{\text{Relative IQR} = \frac{Q3 - Q1}{\text{Median}}}$$

- Measures dispersion relative to the **central value** (median).
- Robust to extreme values (outliers).

💡 **Notes:**

- Both measures are **dimensionless**, allowing comparison between different datasets.
- CV is more sensitive to changes in the mean; Relative IQR is more robust.

Silvestrel (2007)

RELATIVE DISPERSION MEASURES: EXAMPLE

$$2000, 2200, 2500, 2700, 2800, 3000, 3200, 3500$$

Step 1: Mean and Standard Deviation

$$\bar{x} = 2712.5, \quad s \approx 467.3$$

Step 2: Coefficient of Variation (CV)

$$\boxed{CV = \frac{s}{\bar{x}} \times 100\% \approx 17.2\%}$$

Step 3: Interquartile Range (IQR) – using interpolation formula

- **Q1 position:** $(n+1) \cdot 0.25 = 9 \cdot 0.25 = 2.25$
  - $r = 2, \alpha = 0.25$

$$Q1 = (1 - 0.25) \cdot x_2 + 0.25 \cdot x_3 = 0.75 \cdot 2200 + 0.25 \cdot 2500 = 2275$$

- **Q3 position:** $(n+1) \cdot 0.75 = 6.75$
  - $r = 6, \alpha = 0.75$

$$Q3 = (1 - 0.75) \cdot x_6 + 0.75 \cdot x_7 = 0.25 \cdot 3000 + 0.75 \cdot 3200 = 3150$$

$$\text{IQR} = Q3 - Q1 = 3150 - 2275 = 875$$

Step 4: Relative IQR

$$\boxed{\text{Relative IQR} = \frac{\text{IQR}}{\text{Median}} = \frac{875}{2750} \approx 0.318}$$

K-TH MOMENT ABOUT THE ORIGIN

1. k-th Moment about the Origin (Positive Integer k = 1, 2, ...)

- For ungrouped data (simple data):

$$\boxed{m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k}$$

- For grouped data:

$$\boxed{m'_k = \frac{1}{n}\sum_{j=1}^{m} n_j (x'_j)^k}$$

where x'_j = class midpoint, n_j = frequency, m = number of classes.

Note:

- **Raw moments** are computed about the origin.
- They are used to calculate **variance, skewness, and kurtosis**

2. First Moment (k=1) – Mean

$$\text{Ungrouped: } m'_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\text{Grouped: } m'_1 = \frac{1}{n}\sum_{j=1}^{m} n_j x'_j = \bar{x}_{grouped}$$

3. Second Moment (k=2) – Raw Second Moment

$$\text{Ungrouped: } m'_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$\text{Grouped: } m'_2 = \frac{1}{n}\sum_{j=1}^{m} n_j (x'_j)^2$$

Silvestre (2007)

K-TH CENTRAL MOMENT ABOUT THE MEAN

k-th Central Moment (k = 1, 2, ...)

- Ungrouped data:

$$\boxed{m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k}$$

- Grouped data:

$$\boxed{m_k = \frac{1}{n}\sum_{j=1}^{m}n_j(x'_j - \bar{x})^k}$$

where x'_j = class midpoint, n_j = frequency, m = number of classes.

Note:

- Central moments measure **variability around the mean**, with the second central moment equal to the variance.

Special cases:

- k = 1 (First central moment):

$$\boxed{m_1 = 0}$$

- k = 2 (Second central moment – Variance):

$$\boxed{m_2 = s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 \quad \text{or} \quad \frac{1}{n}\sum n_j(x'_j - \bar{x})^2}$$

Silvestre (2007)

MEASURE OF CONCENTRATION: GINI INDEX & LORENZ CURVE

1. Gini Index (G) – Discrete version

$$\boxed{G = \frac{\sum_{i=1}^{m-1}(p_i - q_i)}{\sum_{i=1}^{m-1} p_i} = 1 - \frac{\sum_{i=1}^{m-1} q_i}{\sum_{i=1}^{m-1} p_i}}$$

Notation:

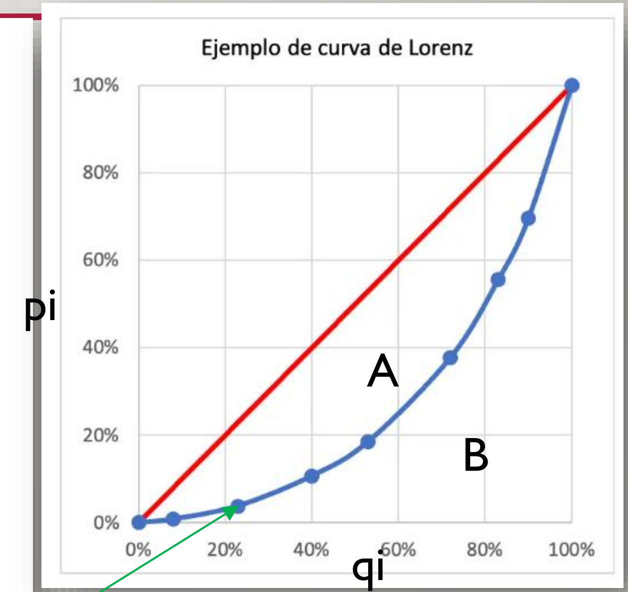
- $p_i = \frac{\sum_{j=1}^{i} n_j}{n}$ → cumulative proportion of **observations** up to class i
- $q_i = \frac{\sum_{j=1}^{i} t_j}{\sum_{k=1}^{m} t_k}$ → cumulative proportion of the **variable** up to class i
- $t_j = n_j \cdot x'_j$ → total of the variable in class j (x'_j = class midpoint)
- m = total number of classes

Conditions:

$$0 \leq p_i \leq 1, \quad 0 \leq q_i \leq 1, \quad p_i \geq q_i, \quad i = 1, \ldots, m$$

💡 **Interpretation:**

- $G = 0$ → perfect equality
- $G = 1$ → maximum inequality



$$Gini\ Index = \frac{A}{A+B}$$

2. Lorenz Curve

- Graph of q_i (cumulative share of variable) vs. p_i (cumulative population)
- **The area between the Lorenz curve and the line of equality** is used to calculate G

GINI INDEX FOR DATA GROUPED: EXAMPLE

Data

Class (k€)	n_j	x'_j	$t_j = n_j x'_j$
0–20	10	10	100
20–40	15	30	450
40–60	20	50	1000
60–80	5	70	350

Totals:

$N = 10 + 15 + 20 + 5 = 50$

$T = 100 + 450 + 1000 + 350 = 1900$

- n_j: class frequency
- x'_j: class midpoint
- $t_j = n_j x'_j$ (class total income)
- $N = \sum_j n_j$ (total population)
- $T = \sum_j t_j$ (total income)
- n_j/N (class population share)
- t_j/T (class income share)
- P_i = (cumulative population share up to class i)
- Q_i : (cumulative income share up to class i)

Shares and cumulative shares

Class	n_j/N	t_j/T	P_i (cum.)	Q_i (cum.)
0–20	0.20000	0.0526316	0.20000	0.0526316
20–40	0.30000	0.2368421	0.50000	0.2894737
40–60	0.40000	0.5263158	0.90000	0.8157895
60–80	0.10000	0.1842105	1.00000	1.0000000

(Values shown to 5–7 significant digits for clarity.)

GINI INDEX FOR DATA GROUPED: EXAMPLE

Apply Silvestre's formula

Silvestre's version sums the cumulative shares **up to the penultimate class** (i.e. exclude the last row where $P_m = Q_m = 1$):

$$\boxed{G = 1 - \frac{\sum_{i=1}^{m-1} Q_i}{\sum_{i=1}^{m-1} P_i}.}$$

Compute the sums (exclude last class):

- $\sum_{i=1}^{m-1} P_i = 0.20000 + 0.50000 + 0.90000 = 1.60000$
- $\sum_{i=1}^{m-1} Q_i = 0.0526316 + 0.2894737 + 0.8157895 = 1.1578948$

Now

$$G = 1 - \frac{1.1578948}{1.60000} = 1 - 0.72368425 = 0.27631575 \approx \mathbf{0.27632}.$$

📌 In theory:

- $G = 0$ → perfect equality (everyone has exactly the same income).
- $G = 1$ → maximum inequality (one person has all the income).

📌 In practice:

Researchers often classify Gini values into approximate ranges:

Gini Value	Typical Interpretation
0.00 – 0.20	Very low inequality (almost perfect equality, rarely observed in large societies).
0.20 – 0.30	Low inequality.
0.30 – 0.40	Moderate inequality.
0.40 – 0.50	High inequality.
> 0.50	Very high (extreme) inequality.

THANKS!

Questions?