

STATISTICAL LABORATORY

Lecture 4: Frequency Distribution and Measures of Location

Yiu Lim LUI, Allen

Lisbon School of Economics & Management
Universidade de Lisboa

`allenlui@iseg.ulisboa.pt`

Applied Mathematics for Economics and Management
1st Year / 1st Semester – 2026/2027

- 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

- ① Frequency Distributions for Discrete Variables
- ② Frequency Distributions for Continuous Variables
- ③ Cumulative Frequency Functions and Ogives
- ④ Measures of Location (Mode, Median, Mean)
- ⑤ Quantiles for Grouped and Ungrouped Data
- ⑥ Measures of Dispersion and Concentration (Gini Index)

Measures of location and 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.”

Core Purpose

Measures of location describe the center or typical value of the dataset.

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.”

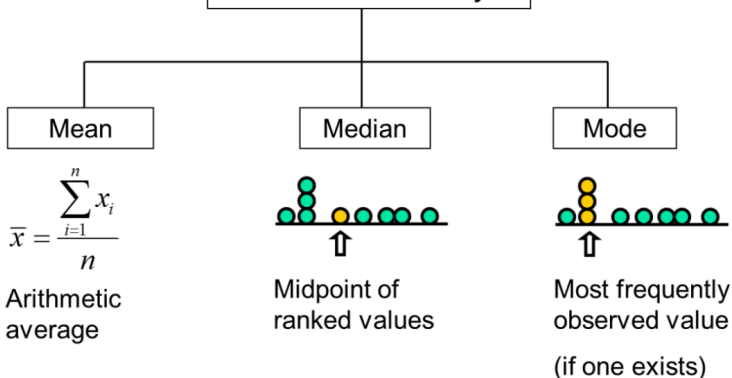
Key Difference:

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

Overview

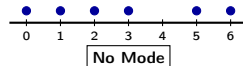
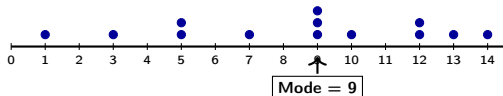
Overview

Central Tendency



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)

① Amodal Distribution (No Mode)

- Definition: No value repeats $\implies$ 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 $\implies$ Mode = 2

③ Bimodal Distribution (Two Modes)

- Definition: Two values appear with the same highest frequency
- Example: 1, 2, 2, 3, 3, 4 $\implies$ 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 $\implies$ 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):

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

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)

Intuition behind King's Formula

- $\frac{f^{**}}{f^* + f^{**}}$ is an adjustment factor.
- This formula estimates the true mode by adding an adjustment factor.
- If we let the mode $= l + \frac{h}{2}$, what happens?
- The adjustment factor uses the neighbor relative frequency to determine the location shift: if f^{**} is relatively large compared to f^* , it moves towards $l + h$.
- Alternatively, it moves towards l when f^{**} is relatively small.

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.

Step 1: Identify modal class $\rightarrow 20 - 30$

Step 2: Extract values:

- $l = 20$

- $h = 10$

- $f^* = 0.16$

- $f^{**} = 0.14$

Step 3: Apply King's formula:

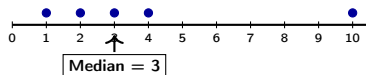
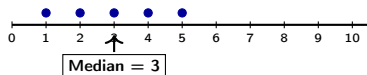
$$Mo = 20 + \frac{0.14}{0.16 + 0.14} \cdot 10 = 20 + 4.67$$

Step 4: Result

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

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:

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

- 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

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

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

Core Principle

The median separates the lower half from the upper half of a data sample arranged in ascending order.

Calculating the Median: Examples

Example 1 – Position is integer

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

$$\text{Position} = \frac{5 + 1}{2} = 3 \implies \text{Median} = 3\text{rd value} = 6$$

Example 2 – Position is not integer

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

$$\text{Position} = \frac{6 + 1}{2} = 3.5 \implies \text{Median} = \frac{8 + 12}{2} = 10$$

Calculating the median: formula

Formula to Find Median Position

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

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

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)

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

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

- What happens when $F^*(I)$ is close to 0.5?
- What happens when $F^*(L)$ is close to 0.5?
- What is the similarity to the previous mode formula with grouped data?

Median for Grouped Data: Table & Definition

Relative & Cumulative Frequencies

Class	Rel. Freq.	Cum. 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

Definition

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

Median for Grouped Data: Calculation Steps

Step 1: Identify the median class $\rightarrow 20 - 30$

Step 2: Extract values for the formula:

- $l = 20, L = 30, h = 10$
- $F^*(l) = 0.26$ (cum. rel. freq. before median class)
- $F^*(L) = 0.50$ (cum. rel. freq. at upper boundary)

Step 3 & 4: Apply formula and find result:

$$\text{Me} = 20 + \frac{0.5 - 0.26}{0.50 - 0.26} \cdot 10 = 20 + \frac{0.24}{0.24} \cdot 10 = 20 + 10 \approx 30$$

Arithmetic mean

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

- For a population of N values:

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

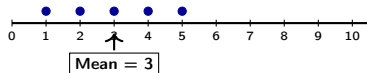
- For a sample of size n :

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

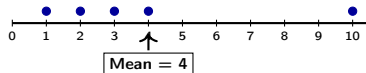
Newbold et al (2013)

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 $x_1, x_2, \dots, x_n$:

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

- It gives more weight to smaller values.
- Typical use: averages of rates or speeds. (H measures average speed when $H = \frac{\text{total.distance}}{\text{total.time}}$.)

Geometric Mean

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

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

- Often used in population growth, compound interest, or growth rates.

General comparison between means:

$$H \leq G \leq A$$

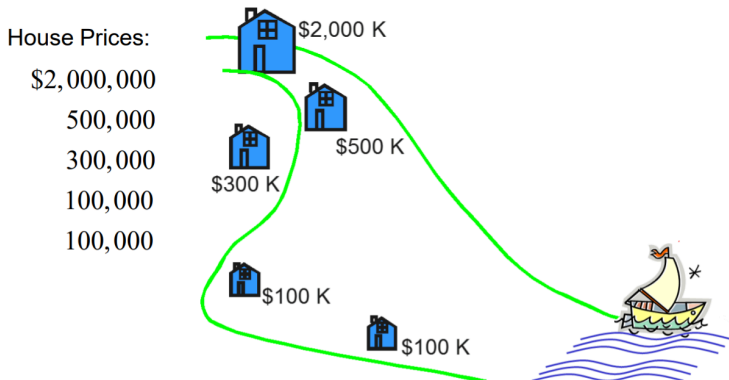
Which measure of location is the “best”?

- **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



Newbold et al (2013)

Review example: summary statistics

House Prices:

\$2,000,000
500,000
300,000
100,000
100,000
Sum 3,000,000

- **Mean:**

$$\left(\frac{\$3,000,000}{5} \right) = \$600,000$$

- **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

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

where w_i = weight of the i th observation and $n = \sum_{i=1}^n w_i$.

Newbold et al (2013)

Weighted Mean: Example

Example 2.17: Stock Recommendation

Zack's Investment Research makes stock recommendations with values: Strong Buy (1), Moderate Buy (2), Hold (3), Moderate Sell (4), Strong Sell (5).

Given data for a specific stock:

- 10 analysts recommend Strong Buy
- 3 analysts recommend Moderate Buy
- 6 analysts recommend Hold

Find the mean recommendation based on Zack's weights.

Conversion Scale: 1 = Strong Buy; 1.1-2.0 = Moderate Buy; 2.1-3.0 = Hold; 3.1-4.0 = Moderate Sell; 4.1-5 = Strong Sell.

Weighted Mean: Example

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

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

The weighted mean of 1.79 yielded a **Moderate Buy** recommendation.

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):

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

Legend / Notation:

- $j = 1, 2, \dots, m \rightarrow$ 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
- $n = \sum_{j=1}^m n_j \rightarrow$ total number of observations

Mean for grouped data: example

Example Calculation:

Class Interval	x'_j (Midpoint)	n_j (Absolute Freq.)	f_j (Relative Freq.)
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:

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

Step 2: Using relative frequencies:

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

Result:

$$\text{Mean} \approx 20.5$$

Proportions: Concepts and Interpretation

Key Ideas

- A proportion represents the part of the total corresponding to a specific category.
- **As a probability:** Interpreted as the likelihood of selecting an observation from that category, ranging from 0 to 1.
- **As a percentage:** Expressed as a share of the total, ranging from 0% to 100%.

Formulas:

- **Proportion / Probability:** $f_j = \frac{n_j}{n}$ where n_j is category frequency and n is total sample size.
- **Percentage:** $\text{Percentage} = f_j \cdot 100$

Proportions: Quantities and Frequencies Example

Example Calculation The table below demonstrates how raw category quantities are converted into relative frequencies, probabilities, and percentages.

Category	Quantity (n_j)	Proportion / Probability (f_j)	Percentage
A	10	0.25	25%
B	20	0.50	50%
C	10	0.25	25%
Total	40	1.00	100%

FOUR PROPERTIES OF THE MEAN

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:

$$m(x + c) = m(x) + c$$

- Subtracting a constant c from all values:

$$m(x - c) = m(x) - c$$

2. Multiplication / Division by a Constant:

- Multiplying all values by a constant c :

$$m(c \cdot x) = c \cdot m(x)$$

- Dividing all values by a constant c :

$$m(x/c) = m(x)/c$$

Four properties of the mean

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

3. Mean of Deviations is Zero:

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

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

4. Mean of Grouped Values:

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

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

Silvestre (2007)

QUANTILES: DEFINITION (REVIEW)

What are Quantiles?

- Quantiles are values that divide a dataset into equal parts.
- Special cases:
 - **Quartiles** $\rightarrow Q_1, Q_2, Q_3, Q_4$ (divide data into 4 equal parts)
 - Median = Q_2
 - **Deciles** $\rightarrow D_1, D_2, \dots, D_{10}$ (divide data into 10 equal parts)
 - Median = D_5
 - **Percentiles** $\rightarrow P_1, P_2, \dots, P_{100}$ (divide data into 100 equal parts)
 - Median = P_{50}

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 Q_1)
- $h = L - l$ = 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 (Q_1).

Step 1: Identify the class containing Q_1

- Q_1 corresponds to the 25th percentile, so $p = 0.25$.
- Look at the cumulative relative frequencies:
 - (10, 20]: 0.35 ✓
- $\rightarrow Q_1$ lies in the (10, 20] class.

Quantile for grouped data: example (cont.)

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)
- $F^*(l) = 0.10$
- $F^*(L) = 0.35$
- $p = 0.25$
- $h = L - l = 10$

$$Q_1 = 10 + \frac{0.25 - 0.10}{0.35 - 0.10} \cdot 10$$

Step 3: Calculate

$$Q_1 = 10 + \frac{0.15}{0.25} \cdot 10 = 10 + 0.6 \cdot 10 = 10 + 6 = 16$$

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
- Interquartile Range (IQR)
- Mean Absolute Deviation (MAD)
- 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: $(Q_3 - Q_1)/\text{Median}$

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

- Gini Index

Measures of Absolute Dispersion (Part 1)

1. Range

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

2. Interquartile Range (IQR)

$$\text{IQR} = Q_3 - Q_1$$

3. Mean Absolute Deviation (MAD) (Ungrouped data only)

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

4. Variance and Standard Deviation

- **Ungrouped (raw) data:**

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

- **Grouped (classified) data:**

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

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

Measures of Absolute Dispersion (1/2)

1. Range

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

2. Interquartile Range (IQR)

$$\text{IQR} = Q_3 - Q_1$$

3. Mean Absolute Deviation (MAD) (only for ungrouped data)

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

4. Variance & Standard Deviation

- Ungrouped (raw) data:

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

- Grouped (classified) data:

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

Absolute dispersion measures (ungrouped data): example

Data (scores):

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

1. Range

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

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

• $n = 10$

• Q_1 position: $(n + 1) \cdot 0.25 = 2.75 \implies Q_1 = 0.25 \cdot 18 + 0.75 \cdot 20 = 19.5$

• Q_3 position: $(n + 1) \cdot 0.75 = 8.25 \implies Q_3 = 0.75 \cdot 35 + 0.25 \cdot 38 = 35.75$

$$\text{IQR} = Q_3 - Q_1 = 16.25$$

3. Mean Absolute Deviation (MAD)

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

4. Variance

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

5. Standard Deviation

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

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

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

Absolute dispersion measures (grouped data): example

Data: Monthly Expenses of 20 Families (\$)

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

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

Step 4: Standard Deviation

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

Properties of Variance (Part 1: Shifts and Scaling)

1. Variance of constants and shifts

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

- Adding or subtracting a constant does not change dispersion.

2. Variance of scaled variables

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

- Scaling by a factor c scales the variance by c^2 .

Properties of Variance (Part 2: Decomposition)

3. Variance decomposition for grouped data

- Suppose n observations are divided into m groups with sizes $n_1, \dots, n_m$, group means $\bar{x}_1, \dots, \bar{x}_m$, and group variances $s_1^2, \dots, s_m^2$.

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

Note: Total variability is the sum of within-group and between-group variability.

Silvestre (2007)

Relative dispersion measures

1. Coefficient of Variation (CV)

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

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

2. Interquartile Range Ratio (IQR / Median)

$$\text{Relative IQR} = \frac{Q_3 - Q_1}{\text{Median}}$$

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

Notes:

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

Relative Dispersion Measures: Example (Part 1)

Dataset:

2000, 2200, 2500, 2700, 2800, 3000, 3200, 3500 ($n = 8$)

Step 1: Mean and Standard Deviation

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

Step 2: Coefficient of Variation (CV)

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{467.3}{2712.5} \times 100\% \approx 17.2\%$$

Note: CV standardizes dispersion relative to the mean, allowing comparison across different units.

Relative Dispersion Measures: Example (Part 2)

Step 3: Interquartile Range (IQR) via Interpolation

- Q_1 position: $9 \times 0.25 = 2.25 \implies Q_1 = 0.75(2200) + 0.25(2500) = 2275$
 - Q_3 position: $9 \times 0.75 = 6.75 \implies Q_3 = 0.25(3000) + 0.75(3200) = 3150$
- $$\text{IQR} = Q_3 - Q_1 = 3150 - 2275 = 875$$

Step 4: Relative IQR

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

K-th Moment about the Origin: Definition

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

- For ungrouped data (simple data):

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

- For grouped data:

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

Notation:

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

K-th Moment about the Origin: Special Cases ($k = 1$ and $k = 2$)

2. First Moment ($k = 1$) – Mean

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

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

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

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

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

K-th Central Moment about the Mean: Definition

k-th Central Moment ($k = 1, 2, \dots$)

- **Ungrouped data:**

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

- **Grouped data:**

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

Notation:

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

K-th Central Moment about the Mean: Special Cases

Special Cases ($k = 1$ and $k = 2$)

- $k = 1$ (First central moment):

$$m_1 = 0$$

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

$$m_2 = s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{or (for grouped data): } \frac{1}{n} \sum_{j=1}^m n_j (x'_j - \bar{x})^2$$

Silvestre (2007)

Measure of Concentration: Gini Index (Formula & Notation)

1. Gini Index (G) – Discrete version

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

Notation:

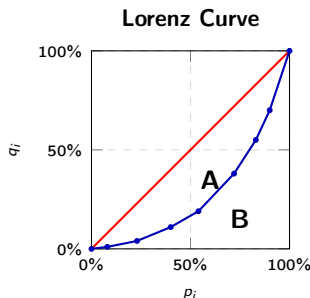
- $p_i = \frac{\sum_{j=1}^i n_j}{\sum_{j=1}^m n_j} \rightarrow$ cumulative proportion of observations up to class i (e.g., People)
- $q_i = \frac{\sum_{j=1}^i t_j}{\sum_{j=1}^m t_k} \rightarrow$ cumulative proportion of the variable up to class i (e.g., Income)
- $t_j = n_j \cdot x'_j \rightarrow$ total of the variable in class j (x'_j = class midpoint)
- m = total number of classes

Conditions:

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

Measure of concentration: Lorenz curve

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



Gini Index Formulation:

$$\text{Gini index} = \frac{A}{A + B}$$

$$\text{Area A} = \sum (p_i - q_i), \text{ Area B} = \sum q_i.$$

Silvestre (2007)

Gini Index for Grouped Data: Example (Part 1)

Initial Data and Class Totals

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

Notation & Totals:

- n_j : class frequency
- x'_j : class midpoint
- $t_j = n_j x'_j$ (class total)
- $N = \sum_j n_j = 50$ (total population)
- $T = \sum_j t_j = 1900$ (total variable sum)

Gini Index for Grouped Data: Example (Part 2)

Shares and Cumulative Shares

Class	n_j/N	t_j/T	P_i (cum.)	Q_i (cum.)
0–20	0.20000	0.05263	0.20000	0.05263
20–40	0.30000	0.23684	0.50000	0.28947
40–60	0.40000	0.52632	0.90000	0.81579
60–80	0.10000	0.18421	1.00000	1.00000

Notation Definitions:

- n_j/N : class population share | t_j/T : class variable share
- P_i : cumulative population share up to class i
- Q_i : cumulative variable share up to class i

(Values shown to 5 decimal places for clarity)

Gini index for data grouped: example (cont.) - Part 1

Apply Silvestre's formula

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

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

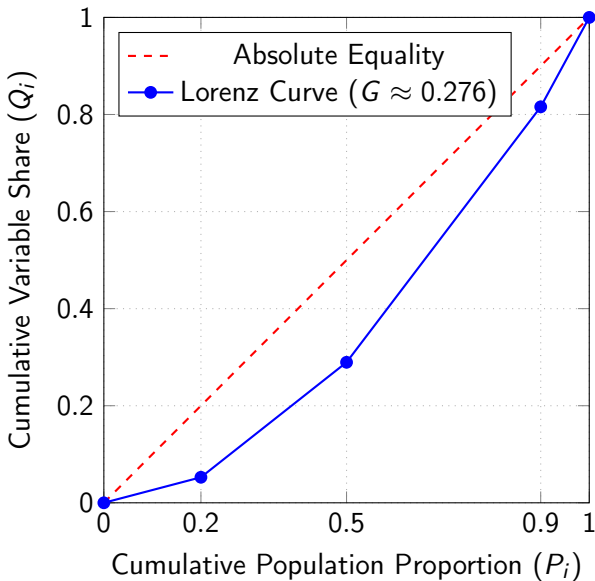
Compute the sums (exclude last class):

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

Now

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

Lorenz Curve Chart



Gini index for data grouped: example (cont.) - Part 2

In theory:

- $G = 0 \rightarrow$ perfect equality (everyone has exactly the same income)
- $G = 1 \rightarrow$ 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)
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?