Chapter 1: Statistical Series of 1 Variable

LAIDI Mohamed

2025/2026

Introduction

What is Statistics?

- Data collection
- Analysis, treatment and interpretation of results
- Presentation to make data understandable to everyone

Basic Statistical Concepts

Statistical Population

Set on which observations are carried out
Examples: People in a survey, Countries with economic data

Statistical Unit (Individual)

Elements of the statistical population
Examples: Each interviewed person, Each country studied

Characteristic and Variable

Characteristic (Caractre)

A **characteristic** is a property or feature that we observe or measure on each individual of the population.

Examples:

- Age of students
- Height of individuals
- Profession of employees

Variable

A **variable** is the actual measurement or observed value that the characteristic takes for specific individuals.

Examples:

- Age: 20, 22, 19, ...
- Height: 175 cm, 162 cm, 181 cm, ...

Relationship: Characteristic vs Variable

Characteristic is to **Variable** as
Question is to **Answer**

Analogy

- **Characteristic (The Question):** "What is your eye color?"
- **Variable (The Answers):** "Blue", "Brown", "Green", etc.

Key Distinction

- **Characteristic:** Abstract concept we are interested in
- **Variable:** Concrete data we collect for that concept

Comparison: Characteristic vs Variable

Characteristic (Caractre)	**Variable**
Abstract concept or property	Concrete measurement or value
Defines *what* is being studied	Represents the *data* collected
Is not a number	Is often a number or a category
Example: *Height* of students	Example: *175 cm*, *162 cm*, *181 cm*

Note

In practice, once the study begins, we work directly with the **variable** (the data).

Statistical Variables

We reminder the definition of statistical variable:

Definition

What is observed or measured on individuals in a population

Types of Variables:

- **Qualitative:** Values expressed literally
- **Quantitative:** Numerical values

Examples:

- Gender, profession
- Height, weight, temperature

Types of Statistical Variables

Qualitative Variables

- **Nominal:** Cannot be ordered
  (gender, family situation)
- **Ordinal:** Can be ordered
  (satisfaction levels, education level)

Quantitative Variables

- **Discrete:** Finite number of values
  (number of children, rooms)
- **Continuous:** Infinite values in interval
  (height, weight, temperature)

Frequency Distribution

Key Concepts

- **Frequency (n_i):** Number of times a modality is observed
- **Total frequency (N):** $N = \sum n_i$
- **Relative frequency (f_i):** $f_i = \frac{n_i}{N}$
- **Cumulative frequency:**
  1. **Increasing** relative cumulative frequencies $F_i\,(f_{ic})$: is

$$F_1 = f_1, F_2 = f_1 + f_2 \text{ and}$$

$$F_i = f_{ic} = f_1 + f_2 + \cdots + f_i = \sum_{p=1}^{i} f_p$$

  2. **Decreasing** relative cumulative frequencies $F_i'\,(f_{id})$: is

$$F_r' = f_r, F_{r-1}' = f_r + f_{r-1} \quad \text{and}$$

$$F_i' = f_{id} = f_r + f_{r-1} + \cdots + f_i = \sum_{p=i}^{r} f_p$$

Remark 1:
In the same way we define the increasing cumulative frequencies $N_i\,(n_{ic})$ and the decreasing cumulative frequencies $N'_i\,(\boldsymbol{n}_{\text{id}}\,)$.

$$N_i = n_{ic} = \sum_{p=1}^{i} n_p \quad \text{et} \quad N'_i = n_{id} = \sum_{p=i}^{r} n_p$$

Remark 2:

$$\sum_{i=1}^{r} f_i = 1 \quad \text{and} \quad \sum_{i=1}^{r} n_i = N$$

Example: Frequency Table

Example

Out of 200 families, 50 have 2 children:

$$f_i = \frac{n_i}{N} = \frac{50}{200} = 0.25 = 25\%$$

Modality	Frequency n_i	Relative frequency f_i
0 children	30	0.15
1 child	40	0.20
2 children	50	0.25
3 children	45	0.225
4+ children	35	0.175
Total	200	1.00

Graphical Representations

Qualitative case

The best graphics used to represent a statistical series in the case of Qualitative variable are :

- Bar Chart
- Pie Chart

Bar Chart

A bar chart consists of bars that represent the modalities of a character.

- The height of each bar is determined by either:
  - Absolute frequency
  - Relative frequency
- Bar charts are useful for comparing different categories or modalities visually.

Example Bar Chart

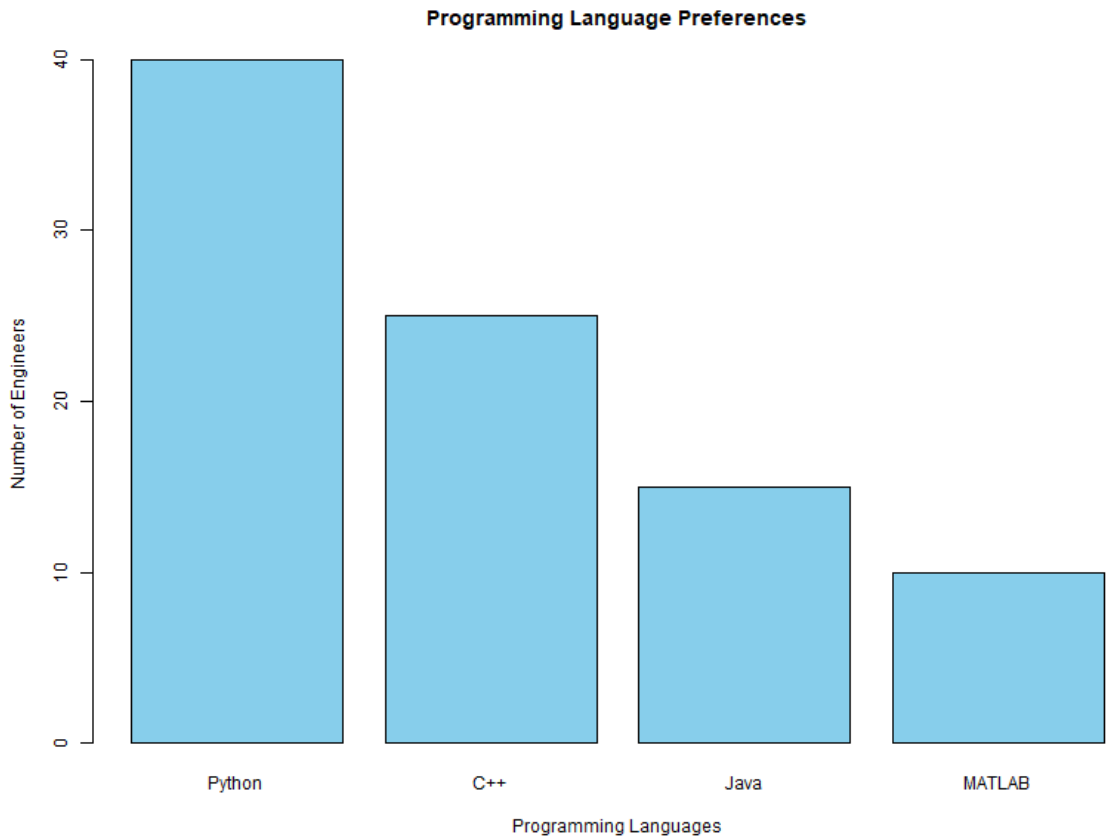


Figure: Example of a Bar Chart Representing Modalities

Pie Chart

A pie chart is a circle partitioned into segments.

- Each segment represents a modality.
- The size of each segment depends on the relative frequency:

$$\theta_i = f_i \times 360^\circ$$

where:

- θ_i is the angle for the i-th segment
- f_i is the relative frequency of the modality

Example Pie Chart

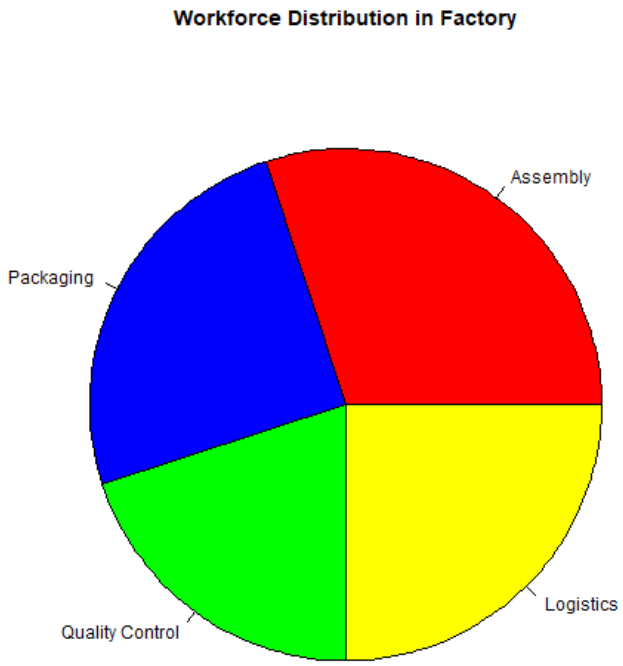


Figure: Example of a Pie Chart Representing Modalities

Graphical Representations

Quantitative Case

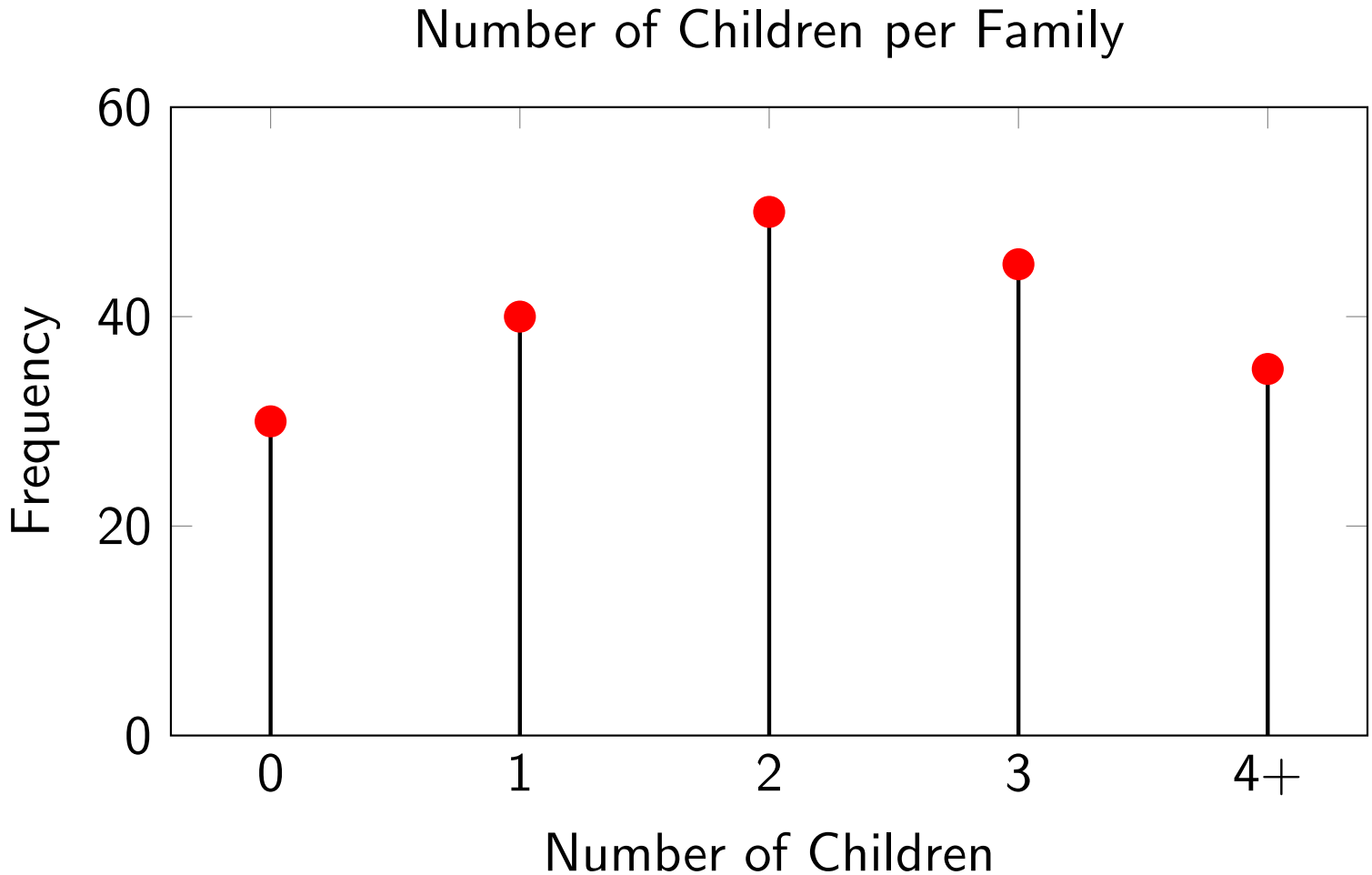
Discrete Variables

- **Line Diagram:** Vertical segments at each x_i
- Height = frequency
- Shows distribution pattern
- **Pie Chart**
- Modalities = disk segments
- Angle: $\alpha_i = f_i \times 360°$
- Shows proportion of whole

Continuous Variables

- **Histogram:** Rectangles with areas proportional to frequencies
- Uses density: $d_i = \frac{n_i}{a_i}$
- Handles different class amplitudes

Line Diagram example for Discrete Variables



- Vertical lines show frequency at each discrete value
- Red points mark exact frequency values
- Used for discrete quantitative variables

Pie Chart for Discrete Variables

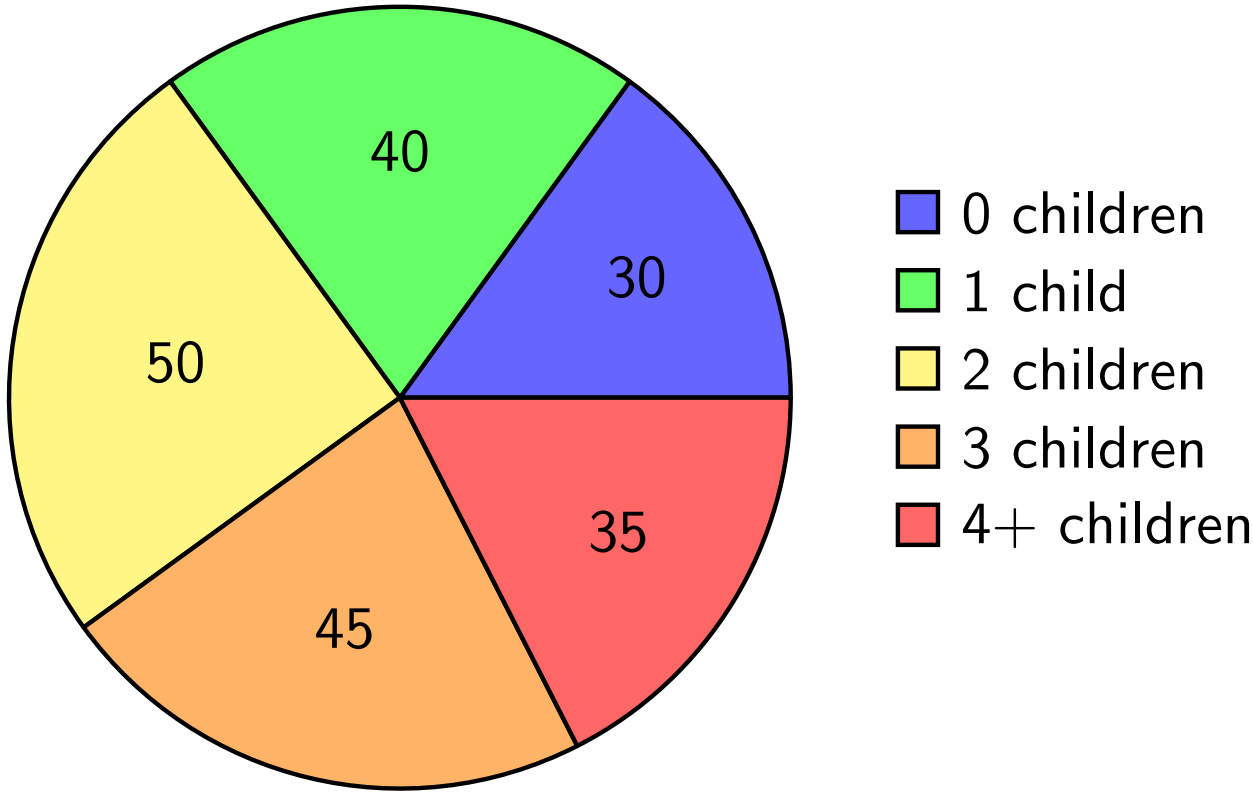


Figure: Distribution of Number of Children per Family

Histogram

- A histogram consists of bars that correspond to class intervals.
- For all $i \in \{1, 2, \ldots, k\}$, the height of the i-th bar is given by:

$$h_i = \frac{f_i}{a_i}$$

  where $a_i = l_i - l_{i-1}$ denotes the magnitude of the i-th class interval $[l_{i-1}, l_i[$.
- The area of each bar is proportional to the corresponding relative frequency.

Remark on Histogram Heights

- If all classes have the same magnitude, we can simplify the height calculation:

$$h_i = f_i \quad \text{for all } i \in \{1, 2, \ldots, k\}$$

- This results in uniform bar widths, making the visualization straightforward.

Histogram Construction Cases

Case 1: Equal Class Widths

- $a_i = a_j$ for all i, j
- Height = frequency n_i
- Simple construction

Case 2: Different Class Widths

- $a_i \neq a_j$
- Use density: $d_i = \frac{n_i}{a_i}$
- Height = corrected frequency

Corrected Frequency

$$n_i^c = d_i \times a^* = \frac{n_i}{a_i} \times a^*$$

where a^* is reference amplitude

Example: Different Class Widths

The distribution of 100 individuals by age classes is given by the next table

Classes $[b_{i-1}, b_i[$	n_i	Amplitude a_i	Density d_i	Corrected freq. n_i^c	f_i	Rel- Freq - cor f_i^c
[5,10[	11	5	2.2	22	0.11	0.22
[10,15[	10	5	2	20	0.10	0.20
[15,20[	15	5	3	30	0.15	0.30
[20,30[	20	10	2	20	0.20	0.20
[30,40[	18	10	1.8	18	0.18	0.18
[40,60[	16	20	0.8	08	0.16	0.08
[60,80[	10	20	0.5	05	0.10	0.05
Total	100				1.00	

Remark: In this example the reference amplitude $a^* = 10$.

The Empirical Cumulative Distribution Function

Discrete case

The empirical cumulative distribution function (ECDF) of X is the function $F : \mathbb{R} \longrightarrow [0, 1]$ defined for all $x \in \mathbb{R}$ by

$$F(x) = \begin{cases} 0 & \text{if } x < x_1 \\ F_i & \text{if } x \in [x_i, x_{i+1}[, \text{ for } 1 \leq i \leq k-1 \end{cases}$$

The **cumulative frequency curve** is the graph of the ECDF.

Remark 2: Properties of the ECDF

The ECDF F satisfies the following properties:

Remark 2: Properties of the ECDF

The ECDF F satisfies the following properties:

- $\forall x \in \mathbb{R}, 0 \leq F(x) \leq 1$

Remark 2: Properties of the ECDF

The ECDF F satisfies the following properties:

- $\forall x \in \mathbb{R}, 0 \leq F(x) \leq 1$
- F is a non-decreasing right continuous function

Remark 2: Properties of the ECDF

The ECDF F satisfies the following properties:

- $\forall x \in \mathbb{R}, 0 \leq F(x) \leq 1$
- F is a non-decreasing right continuous function
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$

The Empirical Cumulative Distribution Function

Continuous case

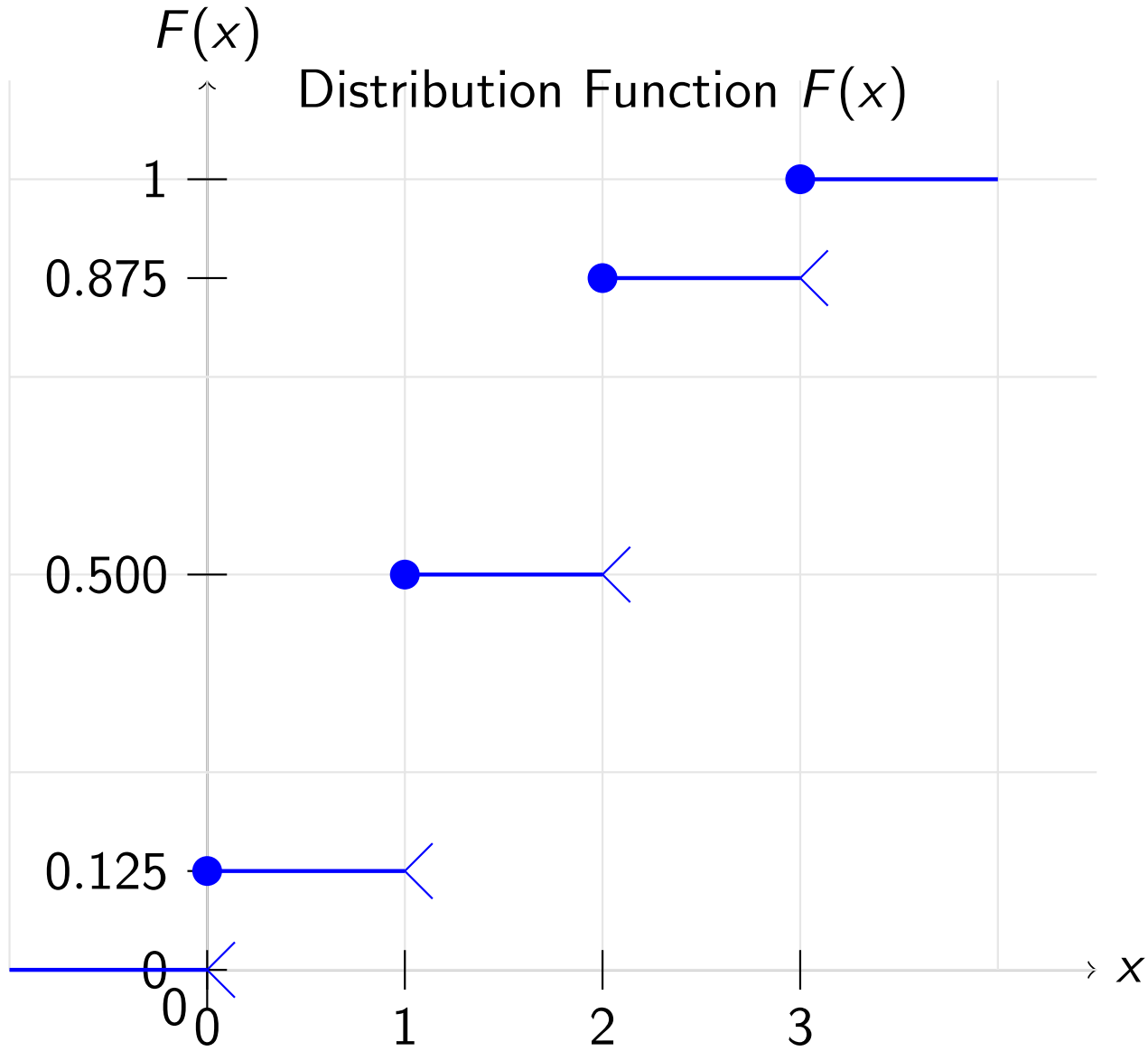
- Cumulative absolute and relative frequencies can be defined similarly to the discrete case.
- The cumulative frequency curve represents the Empirical Cumulative Distribution Function (ECDF).
- In the continuous case, the ECDF is defined as follows: The empirical cumulative distribution function (ECDF) of the continuous variable X is the function $F : \mathbb{R} \longrightarrow [0, 1]$ defined for all $x \in \mathbb{R}$ by

$$F(x) = \begin{cases} 0 & \text{if } x < l_0 \\ F_{i-1} + \frac{f_i}{l_i - l_{i-1}} (x - l_{i-1}) & \text{if } x \in [l_{i-1}, l_i[, \text{ for } 1 \leq i \leq k-1 \\ 1 & \text{if } x \geq l_k, \end{cases}$$

with $F_0 = 0$.

Visualizing the Concepts

Discrete case



Visualizing the Concepts

Continuous case

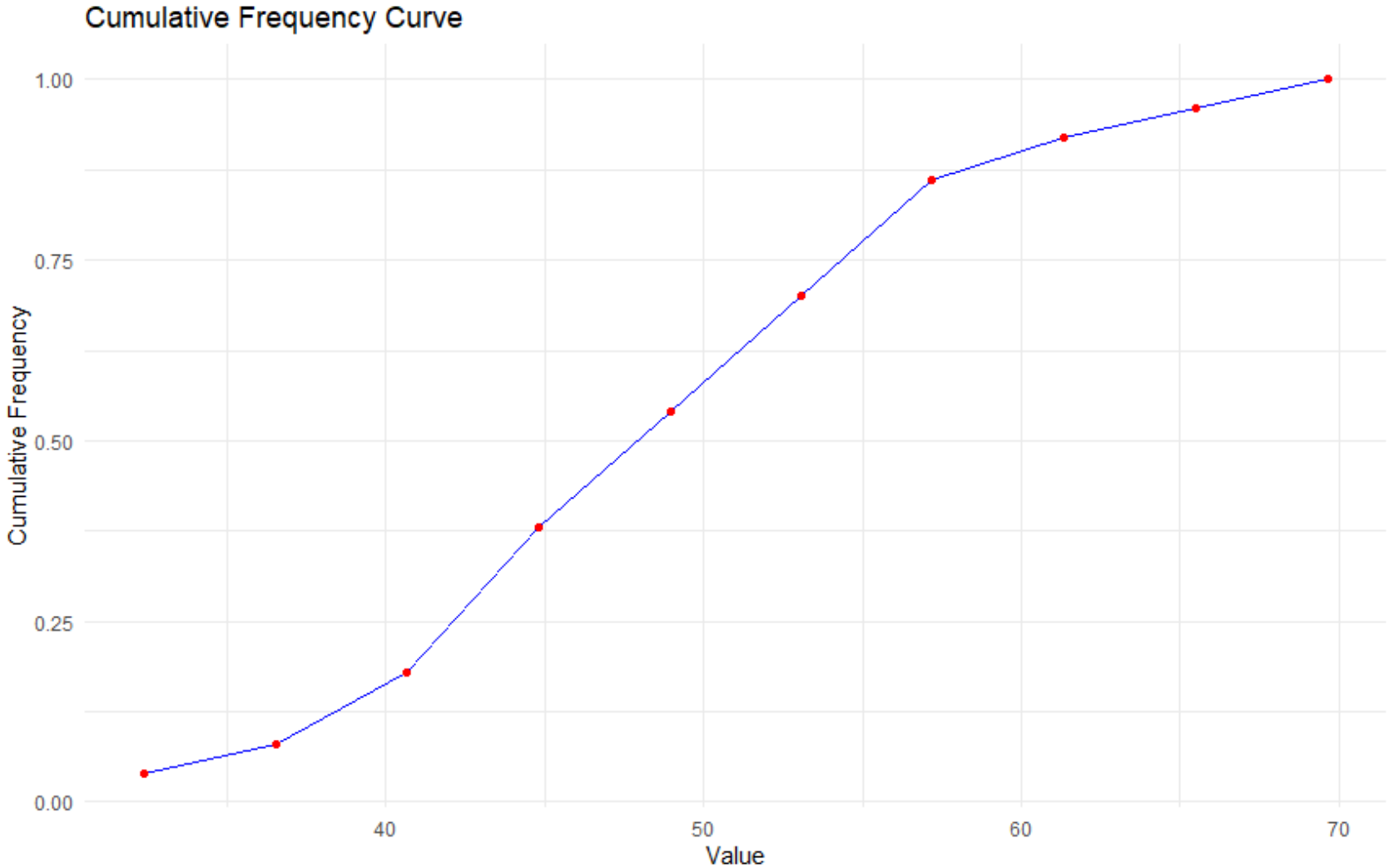


Figure: Cumulative Frequency Curve

Construction of Bins (Classes)

For Continuous Data

Why Construct Bins?

- Continuous data has infinite possible values
- Binning groups data into manageable intervals
- Essential for creating histograms and frequency distributions
- Helps reveal patterns (models) in continuous variables

Construction of Bins (Classes)

For Continuous Data

Why Construct Bins?

- Continuous data has infinite possible values
- Binning groups data into manageable intervals
- Essential for creating histograms and frequency distributions
- Helps reveal patterns (models) in continuous variables

Key Question

How do we determine the number and boundaries of bins?

Step-by-Step Bin Construction

The 5-Step Process

1. **Sort** the data in ascending order
2. Calculate the **Range** = Maximum - Minimum
3. Determine number of bins (**k**)
4. Calculate **Bin Width** = Range / k
5. Define **Bin Boundaries**

Universal Formula

$$\text{Bin Width} = \frac{\text{Range}}{\text{Number of Bins}} = \frac{\text{Max} - \text{Min}}{k}$$

Step 1: Sort the Data

Example Dataset: Heights (meters)

1.74, 1.64, 1.73, 1.76, 1.80, 1.81, 1.79, 1.67, 1.80, 1.70,
1.76, 1.75, 1.78, 1.75, 1.65, 1.79, 1.77, 1.73, 1.76, 1.76,
1.76, 1.85, 1.74, 1.76, 1.82, 1.81, 1.82, 1.71, 1.87, 1.75

Sorted Data

1.64, 1.65, 1.67, 1.71, 1.72, 1.73, 1.73, 1.74, 1.74, 1.75,
1.75, 1.75, 1.76, 1.76, 1.76, 1.76, 1.76, 1.77, 1.78, 1.79,
1.79, 1.80, 1.80, 1.81, 1.81, 1.82, 1.82, 1.85, 1.87

- Minimum: 1.64
- Maximum: 1.87
- Range: 1.87 - 1.64 = 0.23

Step 2: Determine Number of Bins (k)

Sturges' Rule

$$k = 1 + 3.322 \log_{10}(n)$$

- Classic method
- Good for normal data
- Most commonly used

Square Root Rule

$$k = \sqrt{n}$$

- Simple approach
- Conservative bins
- Quick calculation

Yule's Rule

$$k = 2.5 \times \sqrt[4]{n}$$

- Less common but useful
- Good for skewed data
- Reveals more detail

For n = 30

- Sturges: $1 + 3.322 \times 1.477 \approx 5.91$
- Square Root: $\sqrt{30} \approx 5.48$
- Yule: $2.5 \times \sqrt[4]{30} = 2.5 \times 2.34 \approx 5.85$
- Use k = 6 (rounded from all methods)

Step 3: Calculate Bin Width

Width Formula

$$\text{Width} = \frac{\text{Range}}{k} = \frac{0.23}{6} \approx 0.0383$$

Practical Adjustment

- Exact calculation: 0.0383
- Rounded value: 0.04 (for interpretability)
- Why round? Easier boundaries, better communication

Starting Point

Choose starting point : the minimum of the series

$$\text{Start} = 1.64.$$

Step 4: Define Bin Boundaries

Right-Open Interval Convention [a, b)

- Includes a but excludes b
- Clear, unambiguous classification
- Standard statistical practice
- Prevents double-counting

Bin Boundaries

Table: Bin Boundaries using Sturges' method

Class	Interval
1	[1.64, 1.68[
2	[1.68, 1.72[
3	[1.72, 1.76[
4	[1.76, 1.80[
5	[1.80, 1.84[
6	[1.84, 1.88[

Step 5: Complete Frequency Distribution Table

Bin	Interval	n_i	$f_i = \frac{n_i}{30}$	Cumulative F_i
1	[1.64, 1.68[	3	0.100	0.100
2	[1.68, 1.72[	2	0.067	0.167
3	[1.72, 1.76[	7	0.233	0.400
4	[1.76, 1.80[	10	0.333	0.733
5	[1.80, 1.84[	6	0.200	0.933
6	[1.84, 1.88[	2	0.067	1.000

Table: Frequency distribution of the 30 observations

Observations

- Modal class: [1.755, 1.795) with 11 values (36.7%)
- Distribution: Approximately normal
- Most data clustered in middle bins

Practical Exercise: Bin Construction

New Dataset: Product Weights (grams)

45.2, 47.8, 46.3, 45.9, 48.1, 47.2, 46.8, 45.5, 47.9, 46.1,
46.7, 47.3, 45.8, 48.2, 46.9, 47.1, 46.4, 45.7, 47.5, 46.2,
47.0, 48.0, 46.5, 47.4, 47.6, 46.6, 45.6, 48.3, 47.7, 46.0

Instructions

1. Sort the data and find range
2. Calculate k using both Sturges and Square Root rules
3. Determine bin width
4. Define bin boundaries
5. Create frequency distribution table

Expected Results

Range: 3.1, k: 6, Width: 0.52, Modal class: 46.5-47.0

Parameters of a Series

In this section, we will study the main types of parameters used to describe a statistical distribution:

- **Central tendency** indicates where the data are centered.
- **Dispersion** measures how spread out the data are around the center.
- **Form (shape)** characterizes the symmetry and flatness of the distribution.

We will examine these parameters for both discrete and continuous statistical variables.

Location Parameters (Central tendency)

Consider the statistical series:

$$\{(x_1, n_1), (x_2, n_2), \ldots, (x_k, n_k)\} \quad (\text{with } x_1 < x_2 < \cdots < x_k)$$

This series corresponds to a discrete statistical variable X.
We will discuss the following location parameters:

- Mode
- Arithmetic Mean
- Median

Mode

The mode of X, denoted by Mo, is defined as:

- The value(s) having the largest absolute frequency.
- The mode may not be unique; there can be multiple modes in a dataset.

Example:

- For the data set: $\{2, 3, 4, 4, 5, 5, 5, 6\}$
- The mode is 5 (occurs 3 times).

Arithmetic Mean

The arithmetic mean (average) of a statistical series is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$$

where:

- x_i is the value of the i-th modality.
- n_i is the absolute frequency of the i-th modality.

Example:

- For the data set: $\{(2,1),(3,2),(4,4),(5,3)\}$
- The arithmetic mean is:

$$\bar{x} = \frac{(2 \times 1) + (3 \times 2) + (4 \times 4) + (5 \times 3)}{1 + 2 + 4 + 3} = \frac{2 + 6 + 16 + 15}{10} = \frac{39}{10}$$

Remarks:

1. If we use a transformation $Y = aX + b$ with $a, b \in \mathbb{R}$, then

$$\bar{Y} = a\bar{X} + b$$

.

2. If we use a transformation $Y = X - \bar{X}$ then

$$\bar{Y} = 0$$

.

3. If we have two statistical series with means $\bar{X}_1$ and $\bar{X}_2$, and sizes n_1 and n_2 respectively, then the mean $\bar{X}$ of the combined series is given by:

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

Quantiles

- For $p \in [0, 1]$, the p-th quantile x_p is defined as:

$$x_p = \inf\{x \in \mathbb{R} \mid F(x) \geq p\}$$

where F is the empirical cumulative distribution function (ECDF) of X.

Particular Cases

- The Median : $\text{Med} = x_{0.5}$
- The First (lower) Quartile : $Q_1 = x_{0.25}$
- The Third (upper) Quartile : $Q_3 = x_{0.75}$

p-th quantile, descrete case

If the distribution function is discrete:

1. If np is an integer, then

$$x_p = \frac{x_{(np)} + x_{(np+1)}}{2}$$

2. If np is not an integer, then

$$x_p = x_{(\lceil np \rceil)}$$

where $\lceil np \rceil$ represents the smallest integer greater than or equal to np.

Median Calculation (continous case)

To compute the median using linear interpolation:

1. Identify i such that $F_{i-1} \leq 0.5 < F_i$.
2. Use the interpolation formula:

$$\frac{\text{Med} - l_{i-1}}{l_i - l_{i-1}} = \frac{0.5 - F(l_{i-1})}{F(l_i) - F(l_{i-1})}$$

3. Rearranging gives:

$$\text{Med} = l_{i-1} + (l_i - l_{i-1}) \times \frac{0.5 - F(l_{i-1})}{F(l_i) - F(l_{i-1})}$$

Estimating the Median graphically : descrete case

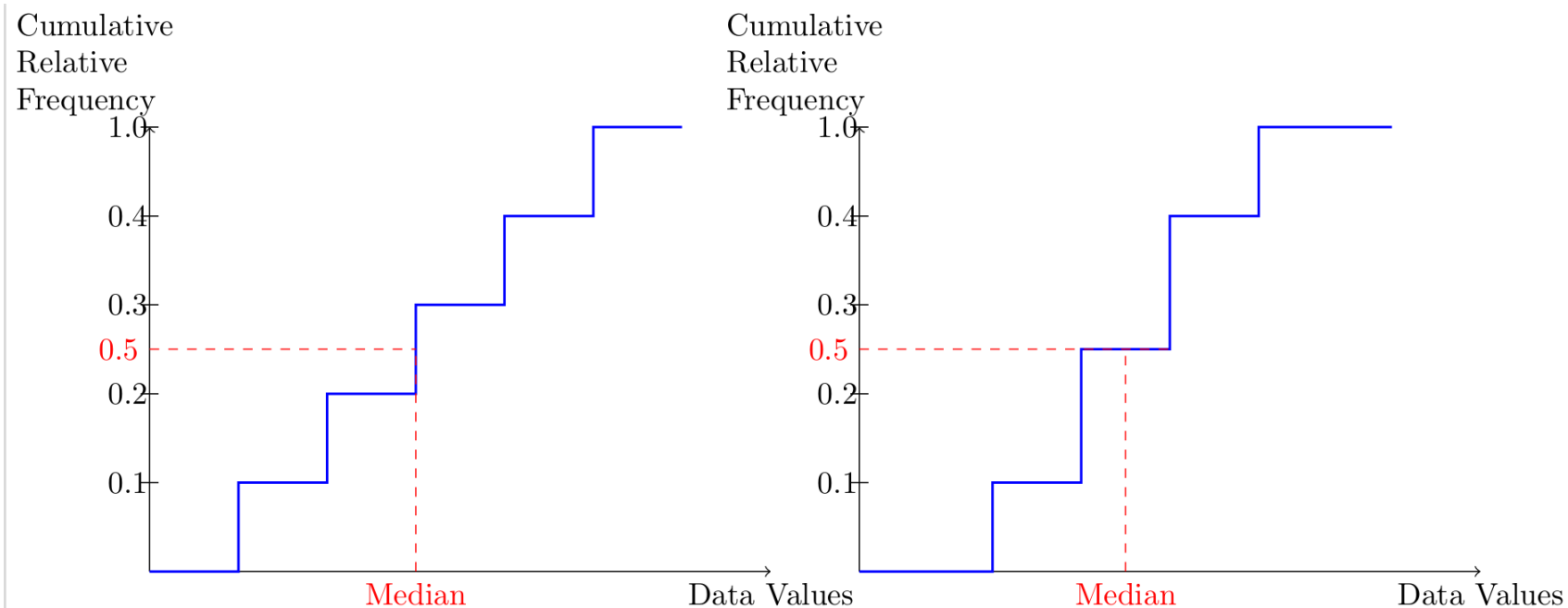


Figure: Estimating the Median graphically

Estimating the Median graphically : continous case

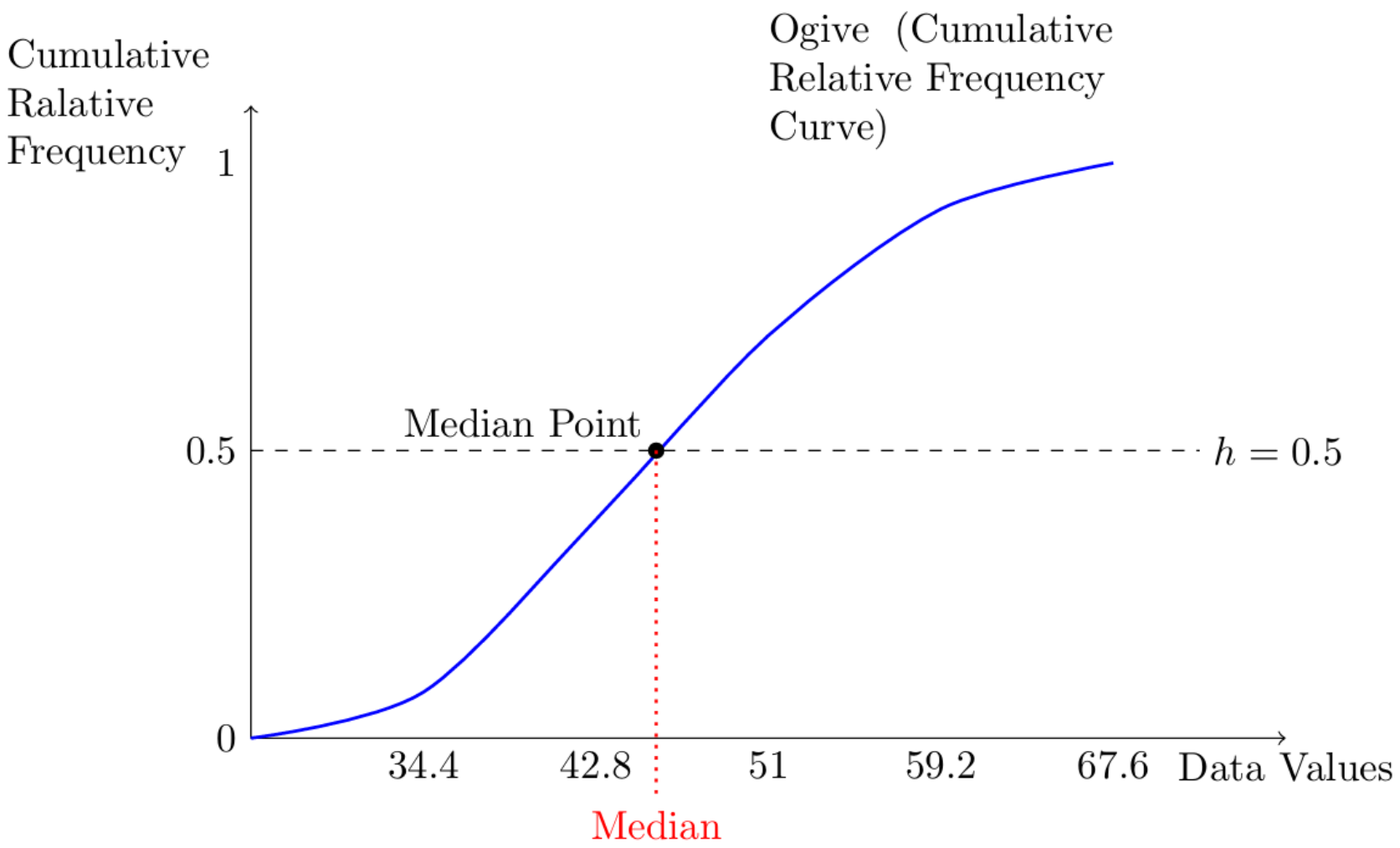


Figure: Estimating the Median graphically

Generalizing for Any p

For any quantile x_p:

1. Identify the appropriate interval for p.
2. Use the interpolation method:

$$x_p = l_{i-1} + (l_i - l_{i-1}) \times \frac{p - F(l_{i-1})}{F(l_i) - F(l_{i-1})}$$

Arithmetic Mean

The arithmetic mean $\bar{X}$ of a continuous random variable X is calculated as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$$

where:

- n_i: frequency of the i-th class
- f_i: relative frequency
- $c_i = \frac{l_{i-1}+l_i}{2}$: midpoint of the i-th class interval $[l_{i-1}, l_i[$

Modal Class

The modal class M is defined as:

- The class (or classes) with the highest density $\frac{n_i}{d_i}$ (or with relative frequency $\frac{f_i}{d_i}$).
- May not be unique (multimodal distributions).

Interpolation Formula for the Mode

To calculate the mode Mo using linear interpolation:

$$Mo = L + \left(\frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} \right) \times h$$

where:

- L = lower boundary of the modal class.
- n_m = frequency of the modal class.
- n_{m-1} = frequency of the class before the modal class.
- n_{m+1} = frequency of the class after the modal class.
- h = width of the class intervals.

Interpolation Formula for the Mode

Remark : This last formula:

$$Mo = L + \left(\frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} \right) \times h$$

can be written as fellow :

$$Mo = L + \left(\frac{\Delta n_{m-1}}{\Delta n_{m-1} + \Delta n_{m+1}} \right) \times h$$

where:

- $\Delta n_{m-1} = n_m - n_{m-1}$.
- $\Delta n_{m+1} = n_m - n_{m+1}$.

Example of Finding the Mode

Consider the following statistical series:

Class Interval	**Frequency**
$[0, 10)$	5
$[10, 20)$	12
$[20, 30)$	20 **(Modal Class)**
$[30, 40)$	8
$[40, 50)$	3

- The modal class is $[20, 30)$ with a frequency of 20.
- We may need to apply linear interpolation to find the mode's exact value.

Example Calculation of Mode

Continuing with the earlier example:

- Modal class: $[20, 30)$
- $L = 20,\ n_m = 20,\ n_{m-1} = 12,\ n_{m+1} = 8,\ h = 10$

Plugging into the formula:

$$Mo = 20 + \left(\frac{20 - 12}{2 \times 20 - 12 - 8}\right) \times 10 = 20 + \left(\frac{(20 - 12)}{(20 - 12) + (20 - 8)}\right) \times 10$$

Simplifying:

$$Mo = 20 + \left(\frac{8}{20}\right) \times 10 = 20 + 4 = 24$$

The mode is approximately 24.

Introduction to Dispersion Parameters

- Dispersion parameters measure the spread or variability within a dataset.
- They help us understand how much data points differ from the central tendency.

Variance

- **Definition:** Average squared deviation from the mean.
- **Population Variance:** $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- **Sample Variance:** $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Interpretation:** Higher variance = more spread out data points.
- **Notation :** Variance of $X \equiv Var(X)$

Properties of Variance

- $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$
- If we use a transformation $Y = aX + b$ with $a, b \in \mathbb{R}$, then

$$Var(Y) = a^2 Var(X)$$

.

Standard Deviation

- **Definition:** Square root of the variance, in same units as data.
- **Population Standard Deviation:** $\sigma = \sqrt{\sigma^2}$
- **Sample Standard Deviation:** $s = \sqrt{s^2}$
- **Interpretation:** Larger standard deviation = greater variability.

Range

- **Definition:** Difference between maximum and minimum values.
- **Formula:** Range $= x_{\text{max}} - x_{\text{min}}$
- **Interpretation:** Quick sense of total spread, sensitive to outliers.

Interquartile Range (IQR)

- **Definition:** Difference between the Third quartile (Q_3) and the Fist quartile (Q_1).
- **Formula:** $\text{IQR} = Q_3 - Q_1$
- **Interpretation:** Represents spread of middle 50% of data, less sensitive to outliers.

Mean Absolute Deviation (MAD)

- **Definition:** Average of the absolute deviations from the mean.
- **Formula:** $\text{MAD} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu|$
- **Interpretation:** An alternative to variance that is less influenced by outliers.

Introduction to Form Parameters

- Form parameters describe the overall shape of a distribution, beyond its location and spread.
- They give crucial information about the symmetry and the "tailedness" of the data.

Definition and Interpretation

Pearson's Skewness Coefficient

The **skewness coefficient** measures the asymmetry of a distribution around its mean:

$$g_1 = \frac{m_3}{s^3}$$

where:

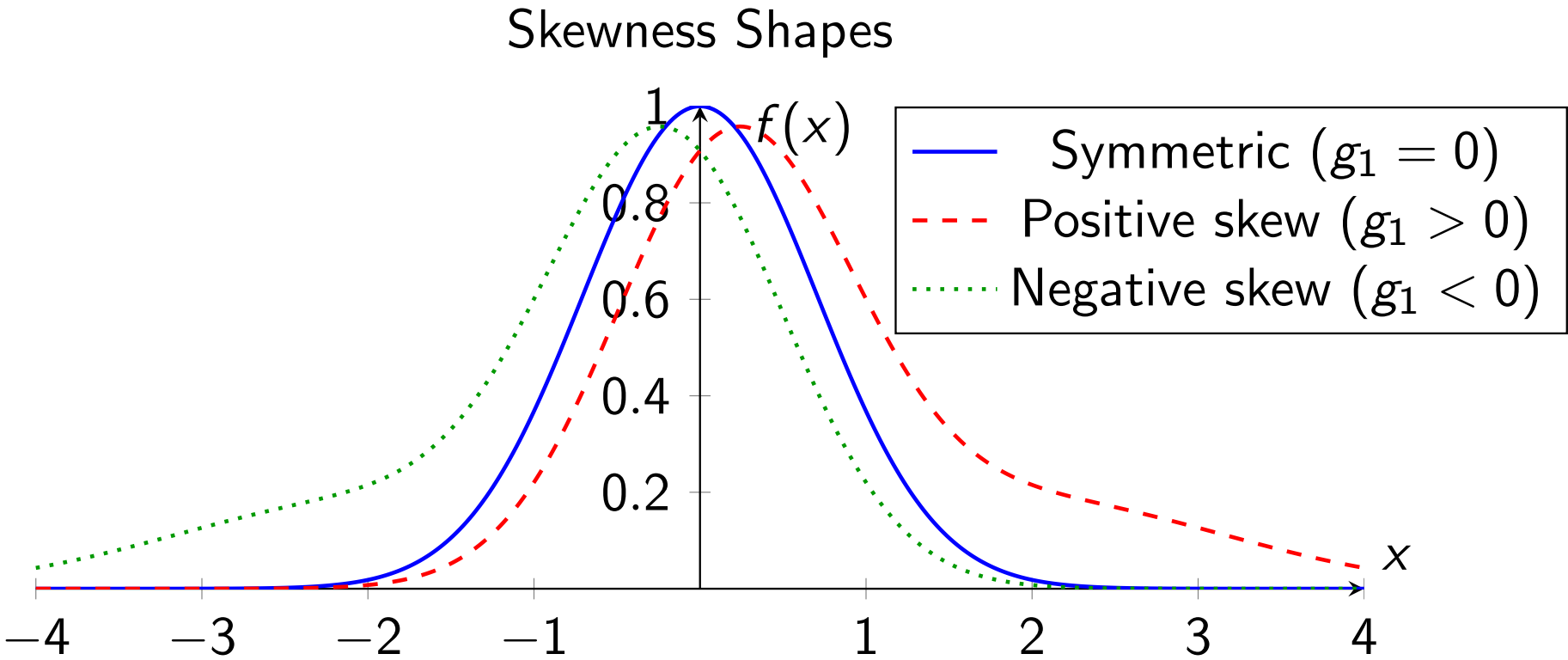
$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

and s is the sample standard deviation.

- $g_1 = 0$: symmetric distribution
- $g_1 > 0$: positive skew (right tail)
- $g_1 < 0$: negative skew (left tail)

Note: m_3 is the third central moment.

Graphical Interpretation



Interpretation:

- Symmetric distribution: mean = median= mode
- Positive skew: mean > median > mode
- Negative skew: mean < median < mode

Calculation Example

Detailed Example

Observed data: 2, 3, 4, 5, 12

$$\bar{x} = \frac{2+3+4+5+12}{5} = 5.2$$

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = 3.54$$

$$m_3 = \frac{1}{5}[(2-5.2)^3 + (3-5.2)^3 + (4-5.2)^3 + (5-5.2)^3 + (12-5.2)^3]$$

$$= \frac{1}{5}[-35.94 - 11.86 - 1.73 - 0.01 + 314.43] = 53.86$$

$$g_1 = \frac{45.78}{3.54^3} = \frac{45.78}{44.36} \approx 1.21$$

Interpretation: Positively skewed distribution (right tail).

Definition and Interpretation

Pearson's Kurtosis Coefficient

The **kurtosis coefficient** measures the peakedness and tail heaviness of a distribution:

$$g_2 = \frac{m_4}{s^4} - 3$$

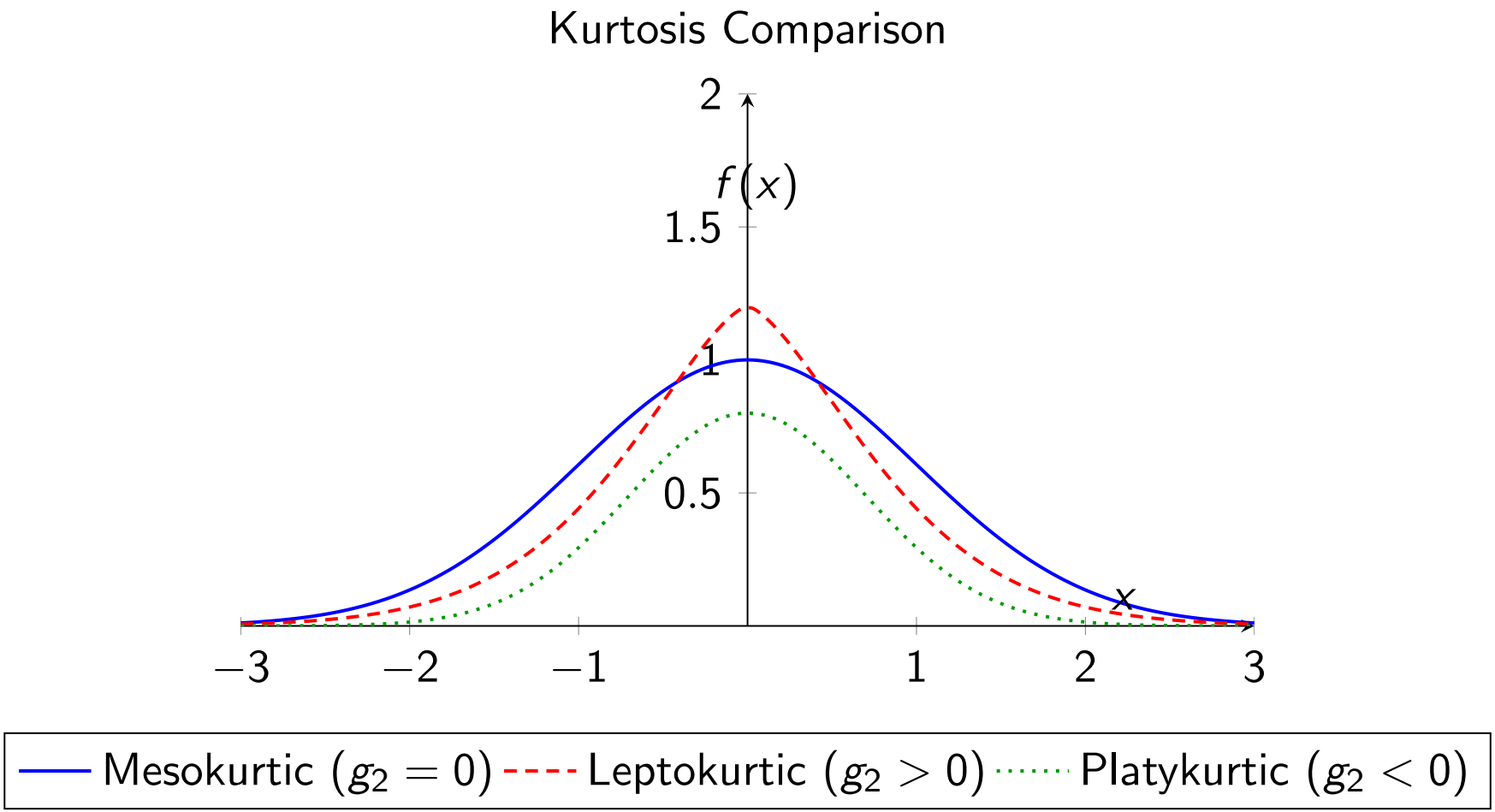
where:

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

- $g_2 = 0$: mesokurtic (like normal distribution)
- $g_2 > 0$: leptokurtic (more peaked, heavy tails)
- $g_2 < 0$: platykurtic (flatter, light tails)

Note: The "-3" centers the measure on the normal distribution.

Graphical Comparison



Practical importance:

- Leptokurtic distributions: more extreme values than normal
- Platykurtic distributions: fewer extreme values than normal

Definition and Applications

Coefficient of Variation

The **coefficient of variation (CV)** measures the relative dispersion of data:

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

Condition: $\bar{x} \neq 0$

- Allows comparison of variability between datasets with different scales
- Low CV indicates homogeneous data

Comparative Example

Group A: $\bar{x} = 100,\ s = 10 \Rightarrow CV = 10\%$
Group B: $\bar{x} = 50,\ s = 10 \Rightarrow CV = 20\%$
$\Rightarrow$ Group A is more homogeneous relative to its mean.