

UNIT 2

Descriptive Statistics

(6 Hours) — Probability and Statistics (CACS202)

1. Measures of Central Tendency

1.1 Introduction

A **measure of central tendency** (average) is a single value that represents the centre of a set of data. A good average is rigidly defined, easy to understand and calculate, based on all observations, not unduly affected by extreme values and suitable for further algebraic treatment. The main averages are the **arithmetic mean**, **median** and **mode**; **partition values** (quartiles, deciles, percentiles) divide data into equal parts.

1.2 Arithmetic Mean (AM)

Data type	Formula for $\bar{x}$
Individual observations	$\bar{x} = \Sigma x / n$
Discrete frequency distribution	$\bar{x} = \Sigma fx / N$, where $N = \Sigma f$
Continuous (grouped) — direct	$\bar{x} = \Sigma fm / N$, m = mid-value of class
Continuous — step deviation	$\bar{x} = A + (\Sigma fd / N) \times h$, $d = (m - A)/h$
Combined mean of two groups	$\bar{x}_c = (n_1\bar{x}_1 + n_2\bar{x}_2) / (n_1 + n_2)$
Weighted mean	$\bar{x}_w = \Sigma wx / \Sigma w$

- **Merits:** simple, uses every value, unique, suitable for algebraic treatment.
- **Demerits:** affected by extreme values, cannot be found for open-ended classes or qualitative data.
- **Properties:** $\Sigma(x - \bar{x}) = 0$; $\Sigma(x - \bar{x})^2$ is the minimum; if every value is changed as $y = a + bx$, then $\bar{y} = a + b\bar{x}$.

Example 2.1 — Mean of individual data

Find the mean of 12, 15, 18, 20, 25, 30, 35.

Solution

$$\Sigma x = 155, n = 7 \rightarrow \bar{x} = 155/7 = 22.14$$

1.3 Median

The **median** is the middle value when the data are arranged in order; half the observations lie above and half below it. For **individual data**: median = value of the $((n + 1)/2)$ th item. For a **continuous distribution**:

Median = $L + [(N/2 - cf) / f] \times h$, where L = lower limit of the median class (the class containing the $(N/2)$ th item), cf = cumulative frequency before it, f = its frequency, h = class width.

- **Merits**: not affected by extreme values, can be used with open-ended classes and found graphically (ogive).
- **Demerits**: ignores actual values, not suitable for algebraic treatment.

1.4 Partition Values (Quartiles, Deciles, Percentiles)

Partition values divide an ordered series into equal parts: **quartiles** Q_1, Q_2, Q_3 into 4 parts, **deciles** $D_1 \dots D_9$ into 10 parts and **percentiles** $P_1 \dots P_{99}$ into 100 parts. Note $Q_2 = D_5 = P_{50}$ = median. For grouped data:

$$Q_i = L + [(iN/4 - cf) / f] \times h, D_i = L + [(iN/10 - cf) / f] \times h, P_i = L + [(iN/100 - cf) / f] \times h$$

1.5 Mode

The **mode** is the value that occurs most frequently. A distribution may be unimodal, bimodal or multimodal. For grouped data, with modal class = class of highest frequency:

Mode = $L + [(f_1 - f_0) / (2f_1 - f_0 - f_2)] \times h$, where f_1 = modal-class frequency, f_0 = frequency before, f_2 = frequency after. Empirical relation for moderately skewed data: **Mode = 3 Median – 2 Mean**.

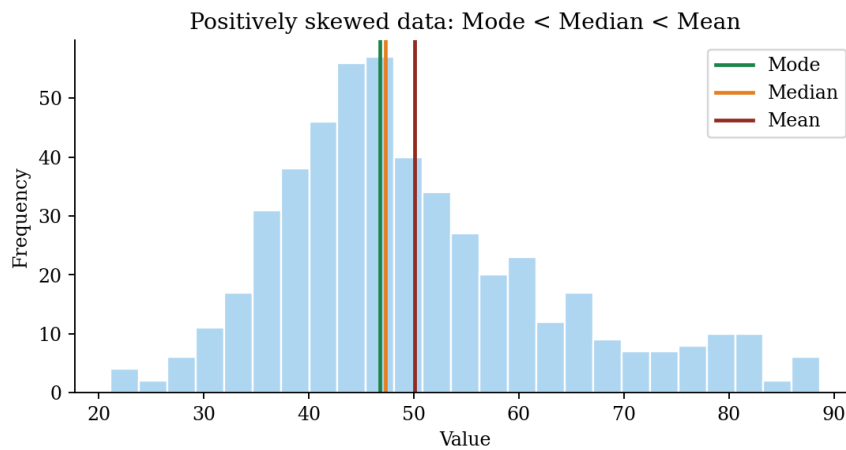


Figure 2.1 — Relative positions of mean, median and mode in a skewed distribution

Example 2.2 — Grouped data: mean, median, quartiles, mode

Marks of 50 students are given below. Calculate the mean, median, mode, Q_1 , Q_3 , D_7 and P_{90} .

Class	f	Mid-value m	fm	cf
0–10	5	5.00	25.00	5
10–20	8	15.00	120.00	13
20–30	15	25.00	375.00	28
30–40	16	35.00	560.00	44
40–50	6	45.00	270.00	50
Total	50		1350.00	

Solution

Mean = $\Sigma fm/N = 1350.00/50 = 27.00$

Median: $N/2 = 25 \rightarrow$ median class 20–30 (cf before = 13, $f = 15$). Median = $20 + (25 - 13)/15 \times 10 = 28.00$

Mode: modal class 30–40 ($f_1 = 16$, $f_0 = 15$, $f_2 = 6$). Mode = $30 + (16 - 15)/(32 - 15 - 6) \times 10 = 30.91$

Q_1 : $N/4 = 12.5 \rightarrow$ class 10–20 (cf = 5, $f = 8$): $10 + (12.5 - 5)/8 \times 10 = 19.38$

Q_3 : $3N/4 = 37.5 \rightarrow$ class 30–40 (cf = 28, $f = 16$): $30 + (37.5 - 28)/16 \times 10 = 35.94$

D_7 : $7N/10 = 35 \rightarrow$ class 30–40: $30 + (35 - 28)/16 \times 10 = 34.38$

P_{90} : $90N/100 = 45 \rightarrow$ class 40–50 (cf = 44, $f = 6$): $40 + (45 - 44)/6 \times 10 = 41.67$

2. Measures of Dispersion

2.1 Introduction

Averages alone do not describe a distribution. **Dispersion** (variation) measures how far the observations are scattered around the centre. Two series may have the same mean but very different spread. **Absolute measures** are expressed in the units of the data; **relative measures** (coefficients) are unit-free ratios used to compare different series.

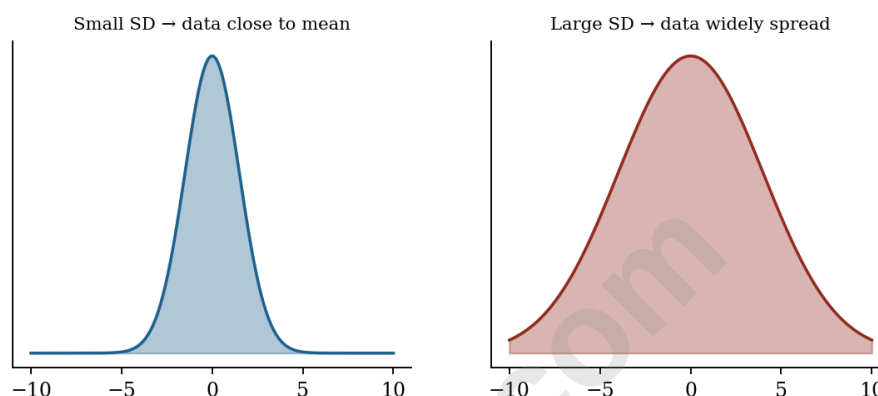


Figure 2.2 — Same mean, different dispersion

Measure	Absolute	Relative (coefficient)
Range	$R = L - S$ (largest – smallest)	$(L - S) / (L + S)$
Quartile deviation	$QD = (Q_3 - Q_1) / 2$	$(Q_3 - Q_1) / (Q_3 + Q_1)$
Mean deviation	$MD = \sum x - A / n$ (A = mean or median)	MD / A
Standard deviation	$\sigma = \sqrt{[\sum (x - \bar{x})^2 / n]} = \sqrt{[\sum x^2 / n - \bar{x}^2]}$	Coefficient of variation $CV = (\sigma / \bar{x}) \times 100 \%$
Variance	σ^2 (square of SD)	—

For a frequency distribution, replace Σ by Σf and n by N , with x replaced by the mid-value m . For a **sample**, the sample standard deviation uses $(n - 1)$ in the denominator: $s = \sqrt{[\sum (x - \bar{x})^2 / (n - 1)]}$.

- **Range** is the simplest but depends only on two extreme values.
- **Quartile deviation** uses the middle 50% of data; suitable for open-ended or skewed data.
- **Mean deviation** uses all values but ignores signs, which makes algebra difficult.
- **Standard deviation** is the most widely used measure; it uses all values and is suitable for further analysis. It is not affected by change of origin but is affected by change of scale: if $y = a + bx$ then $\sigma_y = |b| \sigma_x$.

- **Coefficient of variation** compares the consistency of series: **the smaller the CV, the more consistent (stable) the series.**

Example 2.3 — Dispersion of individual data

For the data 12, 15, 18, 20, 25, 30, 35 find the range, quartile deviation, mean deviation about the mean, standard deviation and coefficient of variation.

Solution

Mean = 22.14. **Range** = $35 - 12 = 23$; coefficient of range = $23/47 = 0.49$

Q_1 = value of $((n+1)/4)$ th = 2nd item = 15; Q_3 = $3(n+1)/4$ = 6th item = 30. **QD** = $(30 - 15)/2 = 7.5$; coeff. = 0.33

MD = $\sum|x - \bar{x}|/n = 47.14/7 = 6.73$; coeff. = 0.30

$\sum(x - \bar{x})^2 = 410.86$, so $\sigma = \sqrt{(58.69)} = 7.66$

CV = $(7.66/22.14) \times 100 = 34.60 \%$

Example 2.4 — Grouped data (from Example 2.2)

Using the same 50 students, find QD, MD, SD and CV.

Class	f	m	$m - \bar{x}$	$f m - \bar{x} $	$f(m - \bar{x})^2$
0–10	5	5.00	-22.00	110.00	2420.00
10–20	8	15.00	-12.00	96.00	1152.00
20–30	15	25.00	-2.00	30.00	60.00
30–40	16	35.00	8.00	128.00	1024.00
40–50	6	45.00	18.00	108.00	1944.00
Total	50			472.00	6600.00

Solution

QD = $(Q_3 - Q_1)/2 = (35.94 - 19.38)/2 = 8.28$; coefficient = 0.30

MD = $\sum f|m - \bar{x}|/N = 472.00/50 = 9.44$; coefficient = 0.35

SD $\sigma = \sqrt{[\sum f(m - \bar{x})^2/N]} = \sqrt{(132.00)} = 11.49$

CV = $(11.49/27.00) \times 100 = 42.55 \%$

Example 2.5 — Which series is more consistent?

Team A scores: mean 50, SD 5. Team B scores: mean 80, SD 6. Which is more consistent?

Solution

$$CV(A) = 5/50 \times 100 = 10\%; CV(B) = 6/80 \times 100 = 7.5\%$$

Team B has the smaller CV, so it is **more consistent**, even though its SD is larger.

Note: Use MD or SD about the mean for the exact formula; the SD shortcut $\sigma^2 = \Sigma x^2/n - \bar{x}^2$ gives the same result.

Key Takeaways

- ✓ Mean uses all data; median (middle value) is robust to extremes; mode is the most frequent value; Mode ≈ 3 Median $- 2$ Mean.
- ✓ Quartiles, deciles and percentiles divide data into 4, 10 and 100 equal parts; $Q_i = L + [(iN/4 - cf)/f] \times h$.
- ✓ Range, QD and MD are simple absolute measures; standard deviation is the most reliable and widely used.
- ✓ Relative measures (coefficients, $CV = \sigma/\bar{x} \times 100$) allow unit-free comparison; a smaller CV means greater consistency.
- ✓ For grouped data use class mid-values and the cumulative-frequency method.

— End of Unit 2 —