SM

📊 Statistical Methods — Formula Sheet

Made By : Kashif Sayyad

🟡 Module I — Basic Concepts & Graphical Presentation

🔹 Class Interval & Class Width

$$\text{Class Width} = \frac{\text{Range}}{\text{Number of Classes}}$$

$$\text{Range} = X_{max} - X_{min}$$

$$\text{Number of Classes (Sturges' Rule)} = 1 + 3.322 \times \log_{10}(n)$$

> Where **n** = total number of observations. Class width should be a round number for ease of use. Equal class widths are preferred for histograms.

🔹 Class Midpoint (Mid-Value)

$$\text{Mid-value} = \frac{\text{Lower Limit} + \text{Upper Limit}}{2}$$

> Used as the representative value of a class in frequency polygon construction and further calculations like mean from grouped data.

🔹 Frequency, Relative Frequency & Percentage Frequency

$$\text{Relative Frequency} = \frac{f}{N}$$

$$\text{Percentage Frequency} = \frac{f}{N} \times 100$$

> Where **f** = frequency of a class, **N** = total number of observations. Relative frequency shows proportion, percentage frequency shows share out of 100.

Cumulative Frequency

$$\text{Cumulative Frequency} = \text{Sum of all frequencies up to and including that class}$$

$$CF_i = f_1 + f_2 + f_3 + \ldots + f_i$$

> Cumulative frequency is used to construct **ogive curves** (less than and more than ogives) and to find median, quartiles, and percentiles graphically.

Pie Diagram — Sector Angle

$$\text{Angle of Sector} = \frac{\text{Component Value}}{\text{Total Value}} \times 360^\circ$$

$$\text{Percentage of Component} = \frac{\text{Component Value}}{\text{Total Value}} \times 100$$

> Example: If salary expense = ₹40,000 and total expense = ₹1,00,000 → Angle = (40,000/1,00,000) × 360° = **144°**

Inclusive to Exclusive Class Conversion

$$\text{Correction Factor} = \frac{\text{Lower Limit of next class} - \text{Upper Limit of current class}}{2}$$

$$\text{New Lower Limit} = \text{Old Lower Limit} - \text{Correction Factor}$$

$$\text{New Upper Limit} = \text{Old Upper Limit} + \text{Correction Factor}$$

> Example: Class 10-19 and 20-29 → Correction = (20-19)/2 = 0.5 → New class becomes **9.5 - 19.5**

Less Than Cumulative Frequency Distribution

Class	Frequency (f)	Less Than CF
0 - 10	5	5
10 - 20	8	13
20 - 30	12	25
30 - 40	7	32

> **Less than CF** adds frequencies from top down. **More than CF** adds from bottom up. Both are used to draw ogive curves for graphical median and quartile finding.

🔹 Frequency Density (for unequal class widths)

$$\text{Frequency Density} = \frac{\text{Frequency}}{\text{Class Width}}$$

> When class widths are unequal, the y-axis of a histogram must show **frequency density** (not frequency) so that the **area** of each bar correctly represents frequency. Equal class widths allow using frequency directly on y-axis.

🔴 Module II — Measures of Central Tendency

🔹 Concept of Central Tendency

A **measure of central tendency** is a single value that represents the center or typical value of a dataset. The three main measures are **Mean**, **Median**, and **Mode** — each defines "center" differently and is suitable for different data types and distributions.

🔹 Arithmetic Mean (A.M.) — Ungrouped Data

$$\bar{X} = \frac{\sum X}{n}$$

> Where **ΣX** = sum of all values, **n** = number of observations. This is the simple mean — add all values and divide by count.

Weighted Mean:

$$\bar{X}_w = \frac{\sum wX}{\sum w}$$

> Where **w** = weight assigned to each value X. Used when different values have different levels of importance.

🔹 Arithmetic Mean — Grouped Data (Direct Method)

$$\bar{X} = \frac{\sum fm}{\sum f} = \frac{\sum fm}{N}$$

> Where **f** = frequency of each class, **m** = midpoint of each class, **N** = total frequency. Multiply each midpoint by its frequency, sum them all, divide by total frequency.

🔹 Arithmetic Mean — Grouped Data (Short Cut / Assumed Mean Method)

$$\bar{X} = A + \frac{\sum fd}{N}$$

> Where **A** = assumed mean (any convenient value, usually middle class midpoint), **d** = (m - A) = deviation of midpoint from assumed mean, **N** = total frequency.

🔹 Arithmetic Mean — Step Deviation Method

$$\bar{X} = A + \frac{\sum fu}{N} \times h$$

$$u = \frac{m - A}{h}$$

> Where **h** = class width, **u** = step deviation. This simplifies calculation when class widths are equal — dividing deviations by h reduces numbers significantly.

Method	Formula	When to Use
Direct	ΣfM/N	Small values
Short Cut	A + Σfd/N	Large values
Step Deviation	A + (Σfu/N)×h	Equal class widths

🔹 Combined Mean

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

> Used to find the overall mean when two groups with known means and sizes are combined. Can be extended to more than two groups.

🔹 Merits and Demerits of A.M.

Merits — Based on all observations, easy to calculate, algebraically manageable, unique value, least affected by sampling fluctuations.

Demerits — Affected by extreme values (outliers), cannot be calculated for open-ended classes without assumption, may give non-integer result even for discrete data, may not

represent actual data point.

◆ Mode — Ungrouped Data

$$\text{Mode} = \text{Value that occurs most frequently}$$

> If two values occur equally often → **Bimodal**. If more than two → **Multimodal**. If all values occur once → **No mode**.

◆ Mode — Grouped Data (Czuprow's Formula)

$$Z = L + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times h$$

> Where:
>
> - **L** = lower boundary of modal class (class with highest frequency)
> - **f_1** = frequency of modal class
> - **f_0** = frequency of class before modal class
> - **f_2** = frequency of class after modal class
> - **h** = class width

Example: If modal class is 20-30, f_1=15, f_0=8, f_2=10, h=10:

$$Z = 20 + \frac{15 - 8}{(15 - 8) + (15 - 10)} \times 10 = 20 + \frac{7}{12} \times 10 = 25.83$$

◆ Merits and Demerits of Mode

Merits — Not affected by extreme values, can be found graphically, applicable to qualitative data, represents actual data value.

Demerits — May not be unique (bimodal/multimodal), not based on all observations, not algebraically manageable, may not exist.

◆ Empirical Relationship Between Mean, Median & Mode

$$\text{Mode} = 3 \times \text{Median} - 2 \times \text{Mean}$$

> This is Karl Pearson's empirical formula — very useful when one measure is unknown. Valid for moderately skewed distributions.

Median — Ungrouped Data

For odd n:

$$\text{Median} = \text{Value at position } \frac{n+1}{2}$$

For even n:

$$\text{Median} = \frac{\text{Value at } \frac{n}{2} + \text{Value at } \frac{n}{2}+1}{2}$$

> Data must be arranged in **ascending or descending order** before finding median. Median is the middle value that divides data into two equal halves.

Median — Grouped Data

$$M = L + \frac{\frac{N}{2} - CF}{f} \times h$$

> Where:
>
> - **L** = lower boundary of median class
> - **N** = total frequency
> - **CF** = cumulative frequency of class before median class
> - **f** = frequency of median class
> - **h** = class width

Median class = class where cumulative frequency first exceeds N/2

Example: N=50, so N/2=25. Find class where CF first crosses 25 — that is the median class.

Merits and Demerits of Median

Merits — Not affected by extreme values, can be found graphically (ogive), suitable for open-ended distributions, best for skewed data.

Demerits — Requires sorting data, not based on all values, not suitable for further algebraic treatment, affected by sampling fluctuations more than mean.

Geometric Mean (G.M.) — Ungrouped Data

$$G.M. = \sqrt[n]{X_1 \times X_2 \times X_3 \times \ldots \times X_n}$$

Using logarithms (easier for large n):

$$\log(G.M.) = \frac{\sum \log X}{n}$$

$$G.M. = \text{Antilog}\left(\frac{\sum \log X}{n}\right)$$

🔹 Geometric Mean — Grouped Data

$$G.M. = \text{Antilog}\left(\frac{\sum f \log m}{N}\right)$$

> Where **m** = midpoint of each class, **f** = frequency, **N** = total frequency. Take log of each midpoint, multiply by frequency, sum them, divide by N, take antilog.

🔹 Merits, Demerits and Applications of G.M.

Merits — Suitable for averaging ratios, rates, and percentages. Less affected by extreme large values than A.M. Gives equal weight to equal ratios.

Demerits — Cannot be calculated if any value is zero or negative. More complex to calculate. Less intuitive to interpret.

Applications — Average growth rates (population, GDP), average rates of return on investments, index numbers, biological growth studies.

🔹 Harmonic Mean (H.M.) — Ungrouped Data

$$H.M. = \frac{n}{\sum \frac{1}{X}} = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_n}}$$

🔹 Harmonic Mean — Grouped Data

$$H.M. = \frac{N}{\sum \frac{f}{m}}$$

> Where **f** = frequency, **m** = class midpoint, **N** = total frequency.

Merits and Demerits of H.M.

Merits — Suitable for averaging rates and speeds (when same distance covered at different speeds). Gives highest weight to smallest values.

Demerits — Cannot be used if any value is zero. Difficult to calculate. Rarely used in practice. Always ≤ G.M. ≤ A.M.

Relationship Between A.M., G.M. and H.M.

$$A.M. \geq G.M. \geq H.M.$$

$$G.M. = \sqrt{A.M. \times H.M.}$$

> The arithmetic mean is always the largest, harmonic mean always the smallest, and geometric mean lies between them. Equality holds only when all values are identical.

Quick Summary Table

Measure	Ungrouped Formula	Grouped Formula	Best Used For
A.M.	ΣX/n	ΣfM/N	Symmetric data
Mode	Most frequent	$L + [(f_1-f_0)/((f_1-f_0)+(f_1-f_2))]\times h$	Categorical, skewed
Median	Middle value	$L + [(N/2-CF)/f]\times h$	Skewed, open-ended
G.M.	Antilog(ΣlogX/n)	Antilog(ΣflogM/N)	Ratios, growth rates
H.M.	n/Σ(1/X)	N/Σ(f/m)	Rates, speeds

Module III — Measures of Dispersion

Concept of Dispersion

Dispersion measures how spread out or scattered the values in a dataset are around the central value. Two datasets can have the same mean but very different spreads — dispersion captures this difference. Higher dispersion = more variability = less consistency.

Absolute measures are in the same units as the data. **Relative measures** (coefficients) are unit-free — used for comparing two datasets.

🔹 Range

$$\text{Range} = X_{max} - X_{min}$$

$$\text{Coefficient of Range} = \frac{X_{max} - X_{min}}{X_{max} + X_{min}}$$

> Simplest measure of dispersion. Uses only two extreme values — highly affected by outliers. Quick to calculate but unreliable for detailed analysis.

Example: Data: 5, 8, 12, 20, 35 → Range = 35 - 5 = **30**, Coefficient = (35-5)/(35+5) = 30/40 = **0.75**

🔹 Quartile Deviation (Q.D.) — Ungrouped Data

First find **Q1** (lower quartile) and **Q3** (upper quartile):

$$Q_1 = \text{Value at position } \frac{n+1}{4}$$

$$Q_3 = \text{Value at position } \frac{3(n+1)}{4}$$

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

> Q.D. is also called **Semi-Interquartile Range**. It measures the spread of the middle 50% of data. Not affected by extreme values — good for skewed distributions and open-ended classes.

🔹 Quartile Deviation — Grouped Data

$$Q_1 = L + \frac{\frac{N}{4} - CF}{f} \times h$$

$$Q_3 = L + \frac{\frac{3N}{4} - CF}{f} \times h$$

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

> Where:
>
> - **L** = lower boundary of quartile class

- **N** = total frequency
- **CF** = cumulative frequency before the quartile class
- **f** = frequency of quartile class
- **h** = class width

Q1 class = class where CF first exceeds N/4 **Q3 class** = class where CF first exceeds 3N/4

◆ Deciles and Percentiles (Extended Quartile Logic)

$$D_k = L + \frac{\frac{kN}{10} - CF}{f} \times h \quad (k = 1, 2, \ldots, 9)$$

$$P_k = L + \frac{\frac{kN}{100} - CF}{f} \times h \quad (k = 1, 2, \ldots, 99)$$

Deciles divide data into 10 equal parts, **Percentiles** into 100 equal parts. D5 = P50 = Median. Same formula structure as median and quartiles — only the fraction changes.

◆ Mean Deviation (M.D.) — Ungrouped Data

$$M.D.(\bar{X}) = \frac{\sum |X - \bar{X}|}{n}$$

$$M.D.(M) = \frac{\sum |X - M|}{n}$$

$$\text{Coefficient of M.D.} = \frac{M.D.(\bar{X})}{\bar{X}} \quad \text{or} \quad \frac{M.D.(M)}{M}$$

Mean deviation is calculated from mean **or** median. Always use absolute values (ignore negative signs). M.D. from median is always minimum compared to M.D. from any other value.

◆ Mean Deviation — Grouped Data

$$M.D.(\bar{X}) = \frac{\sum f|m - \bar{X}|}{N}$$

$$M.D.(M) = \frac{\sum f|m - M|}{N}$$

Where **m** = class midpoint, **f** = frequency, **N** = total frequency. Calculate deviation of each midpoint from mean/median, take absolute value, multiply by frequency, sum, divide by N.

Steps:

1. Find mean or median
2. Find |m - mean| for each class
3. Multiply by frequency: f × |m - mean|
4. Sum all → Σf|m - mean|
5. Divide by N

🔹 Standard Deviation (S.D.) — Ungrouped Data

Direct Method:

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{n}}$$

Shortcut Method:

$$\sigma = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{\sum X^2}{n} - \bar{X}^2}$$

Assumed Mean Method:

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

> Where **d = X - A** (A = assumed mean). Standard deviation is the most important and widely used measure of dispersion — it uses all values and is algebraically manageable.

🔹 Standard Deviation — Grouped Data

Direct Method:

$$\sigma = \sqrt{\frac{\sum f(m - \bar{X})^2}{N}}$$

Shortcut Method:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Step Deviation Method:

$$\sigma = \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} \times h$$

> Where **u = (m - A)/h**. Step deviation method is easiest for grouped data with equal class widths — reduces large numbers to small integers before squaring.

🔹 Standard Deviation — Step by Step Table Method

Class	m	f	u=(m-A)/h	fu	fu²
10-20	15	5	-2	-10	20
20-30	25	8	-1	-8	8
30-40	35	12	0	0	0
40-50	45	7	1	7	7
Total		**32**		**-11**	**35**

$$\sigma = \sqrt{\frac{35}{32} - \left(\frac{-11}{32}\right)^2} \times 10$$

🔹 Coefficient of Variation (C.V.)

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

> C.V. is a **relative measure** of dispersion expressed as a percentage. Used to compare variability of two or more datasets with different units or different means. **Lower C.V. = More consistent/stable**. Higher C.V. = More variable.

Example: Dataset A: mean=50, σ=5 → C.V.=10%. Dataset B: mean=200, σ=15 → C.V.=7.5%. Dataset B is **more consistent** despite having larger σ.

🔹 Variance

$$\text{Variance} = \sigma^2$$

Different formulae for Variance:

$$\sigma^2 = \frac{\sum(X - \bar{X})^2}{n} \quad \text{(Ungrouped, Direct)}$$

$$\sigma^2 = \frac{\sum X^2}{n} - \bar{X}^2 \quad \text{(Ungrouped, Shortcut)}$$

$$\sigma^2 = \frac{\sum f(m - \bar{X})^2}{N} \quad \text{(Grouped, Direct)}$$

$$\sigma^2 = \frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2 \quad \text{(Grouped, Shortcut)}$$

$$\sigma^2 = \left[\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2\right] \times h^2 \quad \text{(Grouped, Step Deviation)}$$

> Variance is simply the **square of standard deviation**. It is always non-negative. Units of variance are the square of data units — which is why S.D. (square root of variance) is more interpretable.

🔹 Combined Standard Deviation

$$\sigma_{combined} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

> Where:
>
> - $\mathbf{d_1} = \bar{X}_1 - \bar{X}combined$
> - $\mathbf{d_2} = \bar{X}_2 - \bar{X}combined$
> - $\bar{X}combined = (n_1\bar{X}_1 + n_2\bar{X}_2)/(n_1+n_2)$

🔹 Complete Summary Table — All Dispersion Measures

Measure	Ungrouped Formula	Grouped Formula	Relative Measure
Range	Xmax - Xmin	Xmax - Xmin	(L-S)/(L+S)
Q.D.	(Q3-Q1)/2	Use quartile formula	(Q3-Q1)/(Q3+Q1)
M.D.	Σ|X-X̄|/n	Σf|m-X̄|/N	M.D./Mean
S.D.	√[ΣX²/n - X̄²]	√[Σfd²/N-(Σfd/N)²]	σ/X̄ × 100 (C.V.)
Variance	σ²	σ²	—

🔹 Properties and Comparison of Dispersion Measures

Property	Range	Q.D.	M.D.	S.D.
Uses all values	❌	❌	✅	✅

Property	Range	Q.D.	M.D.	S.D.
Affected by outliers	✅ Very	❌ No	Slightly	Moderately
Algebraically manageable	❌	❌	❌	✅
Suitable for open-ended	❌	✅	❌	❌
Most reliable	❌	❌	❌	✅

> **Standard Deviation** is the most reliable, most used, and mathematically most useful measure of dispersion. It is the foundation of many advanced statistical techniques including correlation, regression, and hypothesis testing.
