RV Institute of Management ®
Autonomous Institution
Affiliated to Bengaluru City University

# APPLICATION OF STATISTICS IN BUSINESS

**23MBA211**

NOTES


AUTHORS

Dr. Vinay K S
Dr. Santhosh M
Dr. Jahnavi M



DATA DRIVEN DECISIONS | ANALYZE EFFICIENTLY | PREDICT ACCURATELY | PERFORM EXCEPTIONALLY

**RV INSTITUTE OF MANAGEMENT**

Master of Business Administration | Semester I | 23MBA211 | 4 Credits

# APPLICATION OF STATISTICS IN BUSINESS

## 23MBA211

**Authors : Dr. Vinay K S | Dr. Santhosh M | Dr. Jahnavi M**

| MODULE 1 | MODULE 2 | MODULE 3 | MODULE 4 | MODULE 5 |
|---|---|---|---|---|
| Central Tendency<br>8 Hrs | Dispersion, Skewness & Kurtosis<br>8 Hrs | Correlation & Regression<br>10 Hrs | Probability & Distributions<br>8 Hrs | Testing of Hypotheses<br>14 Hrs |

Reference: SP Gupta — Statistical Methods (Sultan Chand) | N.D. Vohra — Business Statistics | Levin & Rubin — Statistics for Management

# MODULE 1: Measures of Central Tendency (8 Hours)

A measure of central tendency is a single value that represents the entire dataset by identifying the central position within that dataset. It summarises a large set of data into a single representative value around which data tends to cluster.

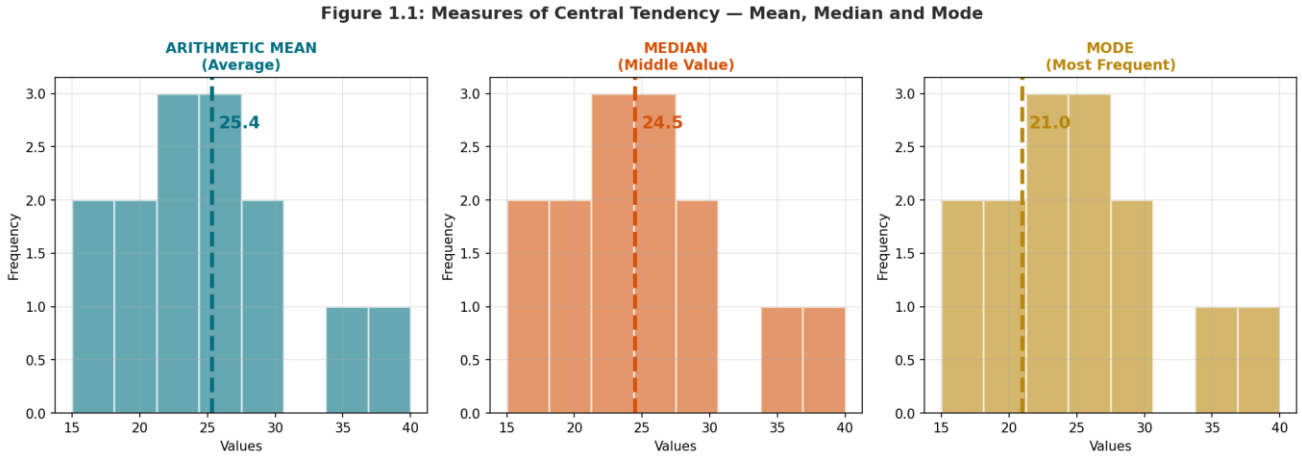


*Figure 1.1: Measures of Central Tendency — Mean, Median and Mode illustrated with frequency distributions*

## 1.1 Arithmetic Mean (AM)

The Arithmetic Mean is the most commonly used measure of central tendency. It is calculated by dividing the sum of all observations by the total number of observations.

### For Ungrouped (Individual) Data:

$$AM = \Sigma X / N \quad \text{(where N = total number of observations)}$$

### For Grouped (Discrete / Continuous) Data:

$$AM = \Sigma fX / \Sigma f \quad \text{(Direct Method)}$$

$$AM = A + (\Sigma fd / \Sigma f) \quad \text{(Short-Cut Method, d = X - A)}$$

$$AM = A + (\Sigma fd' / \Sigma f) \times h \quad \text{(Step-Deviation, d' = (X-A)/h)}$$

where A = Assumed Mean, d = Deviation from Assumed Mean, h = Class Width, f = Frequency

**Properties of Arithmetic Mean:**

- Sum of deviations from AM is always ZERO: $\Sigma(X - \bar{X}) = 0$
- AM is affected by every observation — a change in any value changes the mean
- AM can be calculated for all types of data (individual, discrete, continuous)
- Suitable for further algebraic and statistical calculations
- Most affected by extreme values (outliers) among all measures

### Weighted Arithmetic Mean

When observations have different importance (weights), the weighted mean is used:

```
Weighted AM = ΣwX / Σw   (where w = weight assigned to each observation)
```

Applications: Index numbers, GPA calculation, stock portfolio returns.

## 1.2 Geometric Mean (GM)

The Geometric Mean is the nth root of the product of n observations. It is used when data involves rates, ratios, percentages, or exponential growth.

$$GM = (X_1 \times X_2 \times X_3 \times \ldots \times Xn)^{(1/n)}$$

$$\log GM = (1/N) \times \Sigma \log X \quad \text{[Using logarithms — more practical]}$$

$$\text{For Grouped Data: } \log GM = \Sigma f \cdot \log X / \Sigma f$$

**Applications of Geometric Mean:**

- Computing average growth rates (population growth, compound interest)
- Averaging ratios, rates of change, and percentage increases
- Index number construction (Fisher's Ideal Price Index)
- Averaging logarithmically-distributed data (earthquake magnitudes, stock returns)

## 1.3 Harmonic Mean (HM)

The Harmonic Mean is the reciprocal of the arithmetic mean of the reciprocals of observations. It is particularly useful for averaging rates where the denominator varies.

$$HM = N / \Sigma(1/X) \quad \text{[For ungrouped data]}$$

$$HM = \Sigma f / \Sigma(f/X) \quad \text{[For grouped data]}$$

Best used for: Average speed when distances are equal, average price when quantities are equal.

**Relationship between AM, GM and HM:**

- AM ≥ GM ≥ HM (for all positive values)
- AM × HM = GM² (for two observations only)
- All three are equal only when all observations are identical
- GM is the geometric mean between AM and HM

## 1.4 Median

The Median is the middle value of an ordered dataset. It divides the data into two equal halves — 50% of values lie below and 50% above the median.

**For Ungrouped Data:**

```
If N is odd:  Median = [(N+1)/2]th observation
If N is even: Median = Average of (N/2)th and (N/2 + 1)th observations
```

**For Grouped (Continuous) Data:**

$$Median = L + [(N/2 - cf) / f] \times h$$

Where: L = Lower boundary of median class, N = Total frequency, cf = Cumulative frequency before median class, f = Frequency of median class, h = Class width

**Quartiles, Deciles and Percentiles:**

- Q1 (First Quartile): L + [(N/4 − cf)/f] × h — 25% below this value
- Q2 (Second Quartile): Same as Median — 50% below
- Q3 (Third Quartile): L + [(3N/4 − cf)/f] × h — 75% below
- Deciles (D1–D9): Divide data into 10 equal parts; D5 = Median
- Percentiles (P1–P99): Divide data into 100 equal parts; P50 = Median

## 1.5 Mode

The Mode is the value that occurs most frequently in a dataset. A dataset can have one mode (unimodal), two modes (bimodal), or more (multimodal).

### For Grouped (Continuous) Data — Czuprow's Formula:

$$\text{Mode} = L + [f_1 - f_0] \,/\, [2f_1 - f_0 - f_2] \times h$$

Where: L = Lower boundary of modal class, $f_1$ = Frequency of modal class, $f_0$ = Frequency of pre-modal class, $f_2$ = Frequency of post-modal class, h = Class width

## 1.6 Missing Value Cases

When one observation or one frequency is missing, we use the known total and the AM/Median/Mode formula to find the unknown.

### Missing Observation (when AM is given):

$$\text{Missing value} = (N \times AM) - \text{Sum of known values}$$

### Missing Frequency (when AM or Median is given):

Assign the missing frequency as 'f' and set up the equation using the given AM formula. Solve for f. This is one of the most commonly tested problems.

**Solved Example — Missing Frequency:**

- Given: AM = 28, N = 100, one frequency is unknown
- Step 1: Express ΣfX in terms of unknown f
- Step 2: Use AM = ΣfX / Σf → 28 = ΣfX / 100
- Step 3: Also Σf = 100, so second equation from total frequency
- Step 4: Solve the two equations simultaneously for f

## 1.7 Empirical Relationships

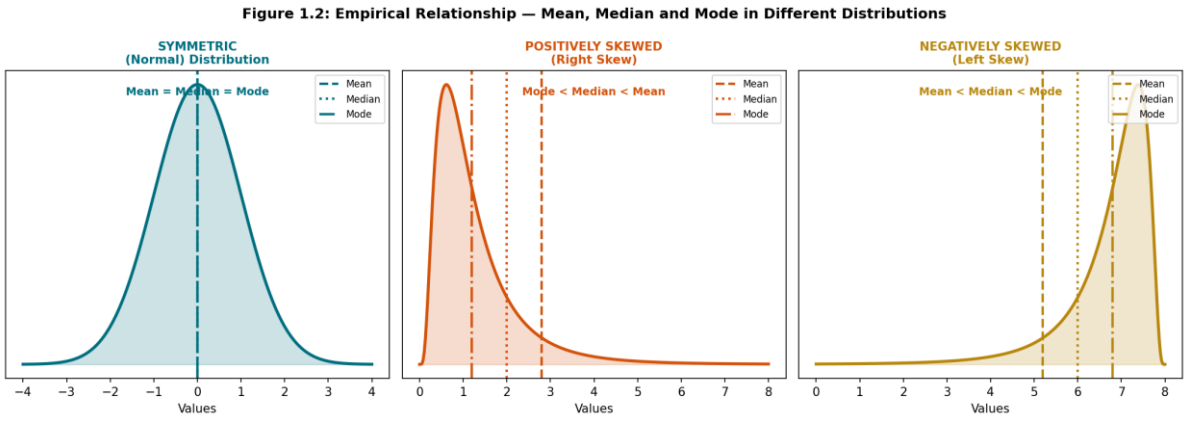


*Figure 1.2: Empirical Relationship — Mean, Median and Mode in symmetric and skewed distributions*

The empirical relationship between Mean, Median and Mode was established by Karl Pearson:

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

$$\text{Median} = \text{Mode} + (2/3)(\text{Mean} - \text{Mode})$$

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

| Symmetric Distribution | Skewed Distribution |
|---|---|
| Mean = Median = Mode | Mean ≠ Median ≠ Mode |
| Bell-shaped, balanced on both sides | Tail pulls one measure away |
| Empirical formula holds perfectly | Empirical formula is approximate |
| Positive skew: Mean > Median > Mode | Negative skew: Mean < Median < Mode |

## 1.8 Applications in Functional Areas of Business

| Functional Area | Measure Used | Application |
|---|---|---|
| **Finance** | AM, Weighted Mean | Average return on portfolio; weighted cost of capital |
| **Marketing** | Mode | Most popular product; most preferred brand |
| **HR Management** | Median | Median salary (avoids distortion by CEOs) |
| **Operations** | GM | Average growth rate; average productivity |
| **Economics** | HM | Average speed; average price indices |
| **Quality Control** | Mean, Median | Process capability analysis; control charts |

# MODULE 2: Measures of Dispersion, Skewness and Kurtosis (8 Hours)

Measures of central tendency alone do not fully describe a dataset. Two datasets can have the same mean but very different spreads. Measures of dispersion quantify the spread or variability around the central value.

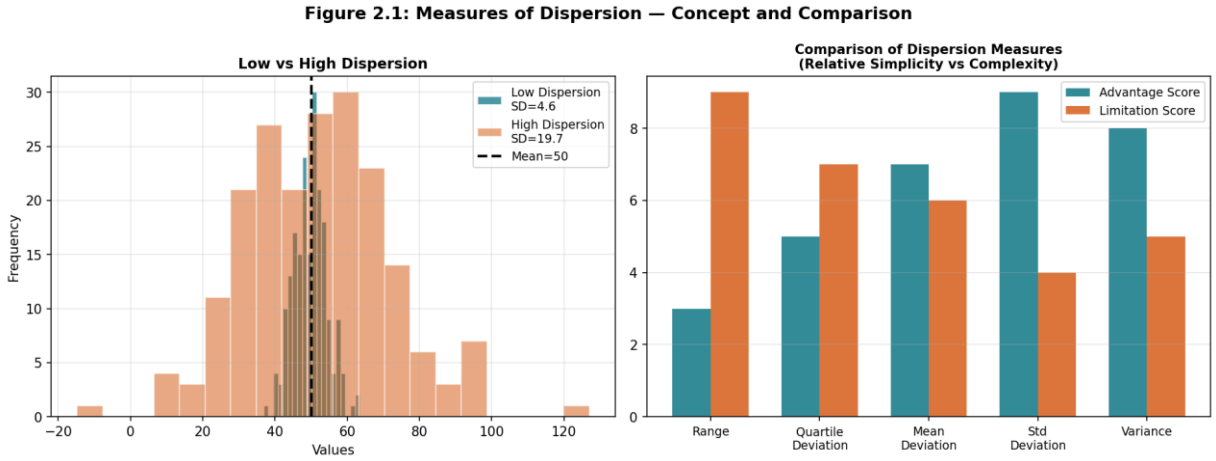


*Figure 2.1: Measures of Dispersion — Low vs High dispersion; Comparison of different measures*

## 2.1 Range

Range is the simplest measure of dispersion. It is the difference between the largest and smallest values in the dataset.

$$\text{Range} = \text{Maximum Value (L)} - \text{Minimum Value (S)}$$

$$\text{Coefficient of Range} = (L - S) / (L + S)$$

| Advantages of Range | Limitations of Range |
|---|---|
| Very simple to calculate and understand | Based on only two extreme values |
| Gives a quick idea of spread | Highly affected by outliers |
| Useful for quality control charts | Ignores all values between L and S |
| Works well for small datasets | Cannot be calculated for open-end classes |

## 2.2 Quartile Deviation (Semi-Interquartile Range)

Quartile Deviation (QD) measures the average spread of the middle 50% of data. It is based on Q1 and Q3 only, so it is not affected by extreme values.

$$\text{Quartile Deviation (QD)} = (Q3 - Q1) / 2$$

$$\text{Coefficient of QD} = (Q3 - Q1) / (Q3 + Q1)$$

Q1 = [(N+1)/4]th value for ungrouped data | Q3 = [3(N+1)/4]th value

## 2.3 Mean Deviation (Average Deviation)

Mean Deviation is the arithmetic mean of the absolute deviations of all observations from the mean (or median). It uses all values, unlike Range and QD.

$$\text{MD (about Mean)} = \Sigma|X - \bar{X}| / N \quad \text{[Ungrouped]}$$

$$\text{MD (about Mean)} = \Sigma f|X - \bar{X}| / \Sigma f \quad \text{[Grouped]}$$

$$\text{Coefficient of MD} = \text{MD} / \text{Mean (or Median)}$$

Mean deviation about the median is the minimum mean deviation — always less than MD about mean.

## 2.4 Standard Deviation (SD) and Variance

Standard Deviation is the most widely used measure of dispersion. It is the positive square root of the mean of squared deviations from the arithmetic mean. It considers all values and is suited for further mathematical operations.

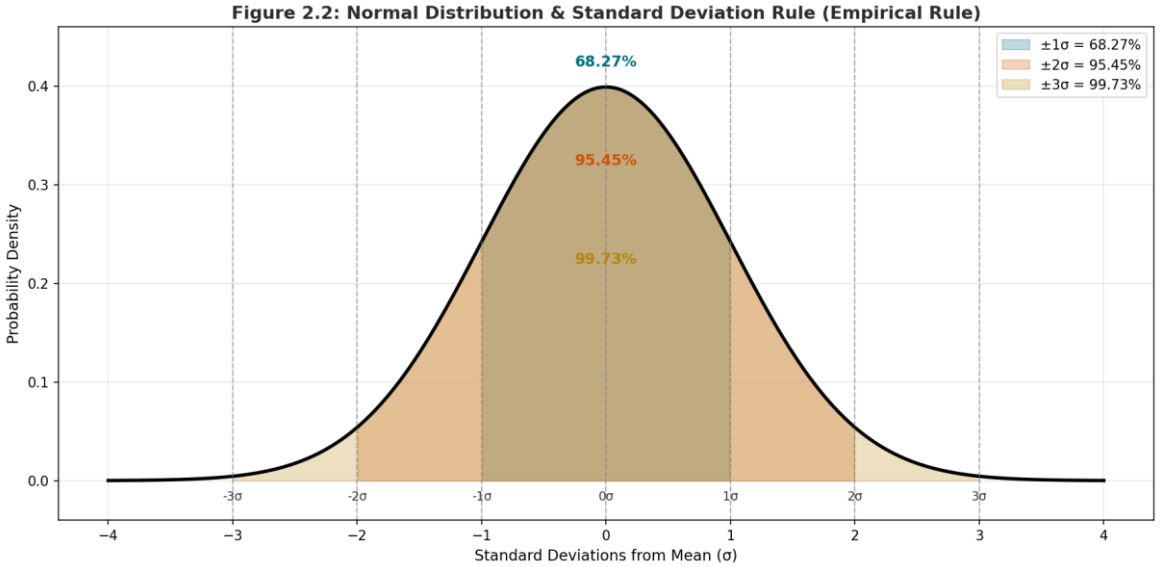


*Figure 2.2: The Empirical Rule (68-95-99.7) — Standard Deviation in Normal Distribution*

$$\sigma = \sqrt{[\Sigma(X - \bar{X})^2 / N] \quad \text{[Direct Method, Ungrouped]}}$$

$$\sigma = \sqrt{[(\Sigma X^2/N) - (\bar{X})^2] \quad \text{[Short-cut Method, Ungrouped]}}$$

$$\sigma = \sqrt{[\Sigma f(X - \bar{X})^2 / \Sigma f] \quad \text{[Grouped — Direct]}}$$

$$\sigma = (h/N) \times \sqrt{[N\Sigma fd'^2 - (\Sigma fd')^2] \quad \text{[Grouped — Step Deviation]}}$$

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

$$\text{Coefficient of Variation (CV)} = (\sigma / \bar{X}) \times 100\%$$

**Properties of Standard Deviation:**

- SD is always ≥ 0; SD = 0 only when all values are equal
- SD is independent of change of origin (adding/subtracting constant doesn't change SD)
- SD is affected by change of scale (multiplying by k multiplies SD by |k|)
- SD of a combined group: $\sigma_c = \sqrt{[(N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2)/(N_1+N_2)]}$
- CV is used to compare variability between datasets with different units or means

### Partition Values — Quartiles, Deciles, Percentiles (for grouped data)

$$Q_1 = L + [(N/4 - cf) / f] \times h$$

$$Q_3 = L + [(3N/4 - cf) / f] \times h$$

$$D_k = L + [(kN/10 - cf) / f] \times h \quad (k = 1 \text{ to } 9)$$

$$P_k = L + [(kN/100 - cf) / f] \times h \quad (k = 1 \text{ to } 99)$$

## 2.5 Skewness

**Figure 2.3: Types of Skewness — Symmetric, Positive and Negative**

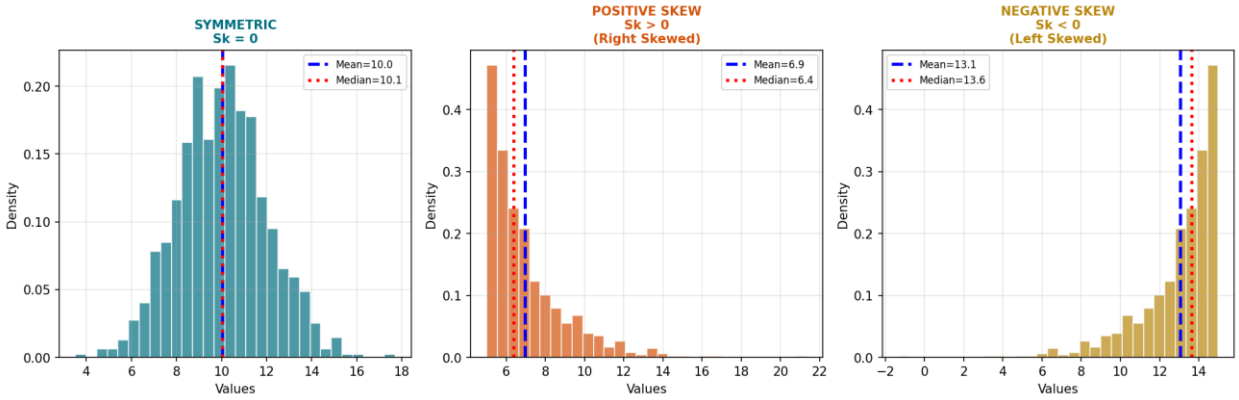


*Figure 2.3: Types of Skewness — Symmetric, Positively Skewed, and Negatively Skewed distributions*

Skewness measures the degree and direction of asymmetry of a frequency distribution. A distribution can be symmetric, positively skewed (right-skewed), or negatively skewed (left-skewed).

| Type | Description | Relationship |
|---|---|---|
| **Symmetric (Sk=0)** | Bell-shaped; equal tails on both sides | Mean = Median = Mode |
| **Positive Skew (Sk>0)** | Right tail longer; most data on left; pulled right | Mode < Median < Mean |
| **Negative Skew (Sk<0)** | Left tail longer; most data on right; pulled left | Mean < Median < Mode |

### Karl Pearson's Coefficient of Skewness

$$Skp = (Mean - Mode) / \sigma$$

$$Skp = 3(Mean - Median) / \sigma \quad [\text{When mode is ill-defined}]$$

Range of Karl Pearson's coefficient: $-3 \leq Skp \leq +3$ (approximately $\pm 1$ in practice)

### Bowley's Coefficient of Skewness (Quartile Skewness)

Bowley's method is based on quartiles, making it suitable when extreme values are present or for open-ended distributions.

$$Skb = (Q_3 + Q_1 - 2Q_2) / (Q_3 - Q_1)$$

Range of Bowley's coefficient: $-1 \leq Skb \leq +1$

**Comparison — Karl Pearson vs Bowley:**

- Karl Pearson: Uses AM and SD; affected by extreme values; range ≈ −3 to +3
- Bowley: Uses quartiles; not affected by extreme values; range −1 to +1
- Karl Pearson is preferred when data is not highly skewed
- Bowley is preferred for open-ended distributions or highly skewed data

## 2.6 Kurtosis

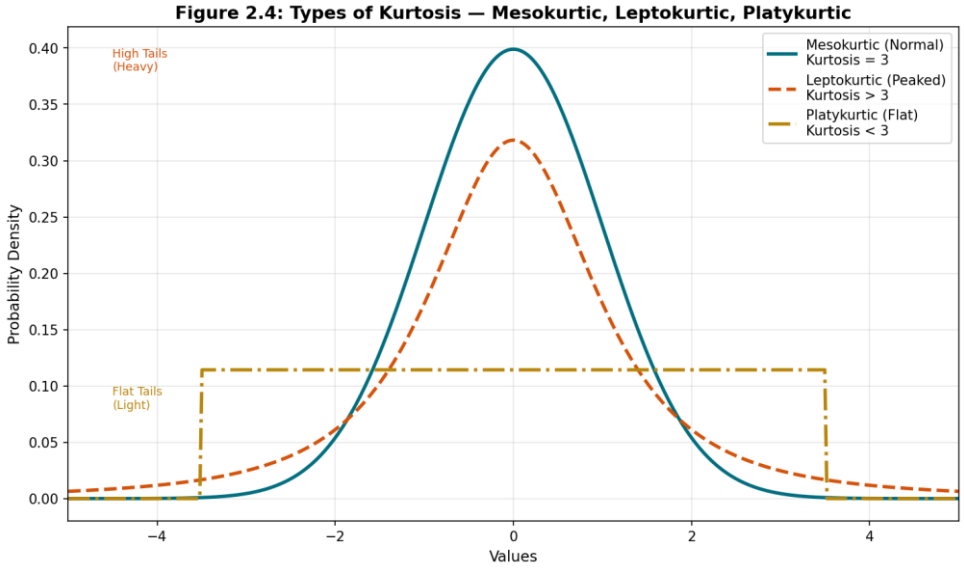


*Figure 2.4: Types of Kurtosis — Mesokurtic (Normal), Leptokurtic (Peaked), Platykurtic (Flat)*

Kurtosis measures the 'peakedness' or 'flatness' of a frequency distribution relative to the normal distribution. It describes the shape of the tails of the distribution.

$$\beta_2 \text{ (Beta-2)} = \mu_4 / \mu_2^2 \quad [\text{where } \mu_4 = \text{fourth central moment, } \mu_2 = \text{variance}]$$

$$\gamma_2 \text{ (Excess Kurtosis)} = \beta_2 - 3$$

| Type | $\beta_2$ Value | Description |
|---|---|---|
| **Mesokurtic (Normal)** | $\beta_2 = 3$ ($\gamma_2 = 0$) | Normal bell curve; moderate peak; moderate tails |
| **Leptokurtic (High Peak)** | $\beta_2 > 3$ ($\gamma_2 > 0$) | More peaked than normal; heavier/fatter tails; more outliers |
| **Platykurtic (Flat)** | $\beta_2 < 3$ ($\gamma_2 < 0$) | Flatter than normal; lighter tails; fewer extreme values |

# MODULE 3: Correlation & Regression Analysis (10 Hours)

Correlation measures the strength and direction of the linear relationship between two variables. Regression goes further by establishing the mathematical equation of that relationship for prediction purposes.

## 3.1 Introduction to Correlation

Correlation analysis studies the relationship between two quantitative variables X and Y. It tells us HOW STRONGLY and in WHICH DIRECTION they are related.

| TERM / FORMULA | MEANING & FORMULA |
|---|---|
| **Perfect Positive Correlation** | r = +1: As X increases, Y increases proportionally; all points on a straight line |
| **Strong Positive Correlation** | 0.7 < r < 1: Strong tendency for Y to increase with X |
| **Moderate Positive Correlation** | 0.3 < r < 0.7: Moderate tendency; some scatter around the trend line |
| **No Correlation** | r ≈ 0: No linear relationship between X and Y |
| **Moderate Negative Correlation** | −0.7 < r < −0.3: Moderate tendency for Y to decrease as X increases |
| **Perfect Negative Correlation** | r = −1: As X increases, Y decreases proportionally |

## 3.2 Scatter Diagram

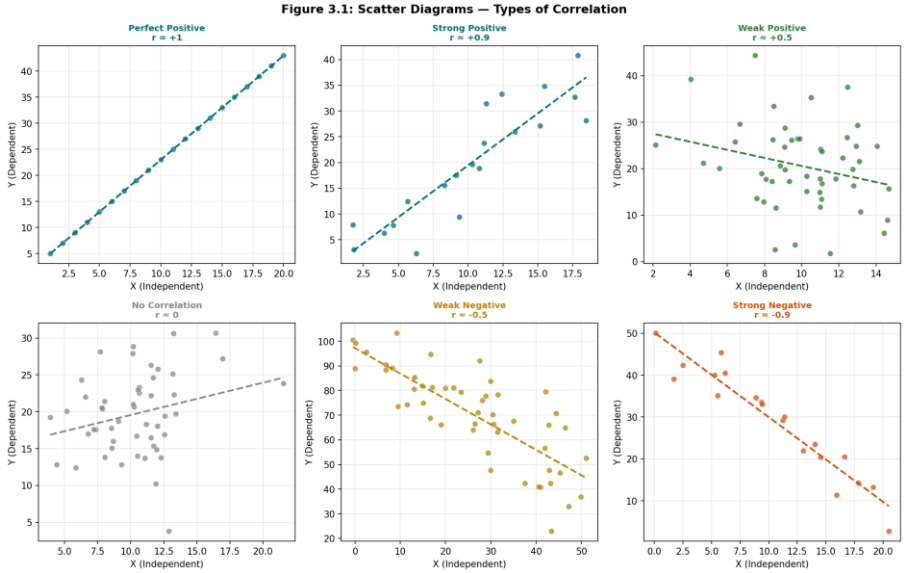


*Figure 3.1: Scatter Diagrams — Six types of correlation from perfect positive to strong negative*

A scatter diagram (scatter plot) is the simplest way to visually assess correlation. Each pair (X, Y) is plotted as a point. The pattern of points shows the type and strength of the relationship.

- Upward slope from left to right → Positive correlation
- Downward slope from left to right → Negative correlation
- Points close to a line → Strong correlation

- Points widely scattered → Weak or no correlation

## 3.3 Karl Pearson's Coefficient of Correlation

Karl Pearson's coefficient (r) is the most widely used measure of correlation. It measures the linear correlation between two quantitative variables.

$$r = \Sigma(X - \bar{X})(Y - \bar{Y}) \;/\; \sqrt{[\Sigma(X - \bar{X})^2 \times \Sigma(Y - \bar{Y})^2]}$$

$$r = [N\Sigma dxdy - \Sigma dx\Sigma dy] \;/\; \sqrt{\{[N\Sigma dx^2 - (\Sigma dx)^2][N\Sigma dy^2 - (\Sigma dy)^2]\}} \quad \text{[Short-cut: } dx=X-A,\ dy=Y-B\text{]}$$

**Properties of Karl Pearson's r:**

- r lies between −1 and +1: $-1 \le r \le +1$
- r is dimensionless — it has no units
- r measures only LINEAR relationship; non-linear relationships may have $r \approx 0$
- r is symmetric: r(X,Y) = r(Y,X)
- r is not affected by change of origin or scale (adding/multiplying constants)
- $r^2$ = Coefficient of Determination = % of variation in Y explained by X

### For Bivariate (Grouped) Data

When data is available in a correlation table (bivariate frequency distribution), we use step-deviation method:

$$r = [N\Sigma fdxdy - \Sigma fdx\Sigma fdy] \;/\; \sqrt{\{[N\Sigma fdx^2 - (\Sigma fdx)^2][N\Sigma fdy^2 - (\Sigma fdy)^2]\}}$$

## 3.4 Spearman's Rank Correlation

Spearman's Rank Correlation (ρ) is used when data is in ordinal (ranked) form, or when Karl Pearson's assumptions are violated. It measures the strength of monotonic (not necessarily linear) relationships.

$$\rho = 1 - [6\Sigma D^2] \;/\; [N(N^2-1)] \quad \text{[Without ties]}$$

where D = Rank(X) − Rank(Y), N = number of pairs

### When There Are Ties (Equal Ranks)

When two or more observations have the same value, assign them the average of the ranks they would have received, and apply the correction factor:

$$\rho = 1 - 6[\Sigma D^2 + \Sigma m(m^2-1)/12] \;/\; [N(N^2-1)]$$

where m = number of items with the same rank in a group

| Karl Pearson's r | Spearman's ρ |
|---|---|
| Measures LINEAR correlation | Measures MONOTONIC correlation |
| Requires quantitative data (interval/ratio) | Suitable for ordinal (ranked) data |
| Uses actual values of X and Y | Uses ranks of X and Y |
| Affected by extreme values (outliers) | Not affected by extreme values |
| More informative for normal data | Preferred for skewed data or small samples |

## 3.5 Concurrent Deviation Method

This is the simplest method of finding correlation. We note whether each successive change in X and Y is an increase (+), decrease (−), or same (0), then compare directions.

$$r_c = \pm\sqrt{[\pm(2c - N) / N]}$$

where c = number of concurrent deviations (both X and Y change in same direction), N = number of pairs of deviations

The sign inside the square root must be ±. If (2c − N) is positive, take + outside; if negative, take −. This method gives only an approximate value of r.

## 3.6 Regression Analysis

**Figure 3.2: Regression Lines and Residuals**

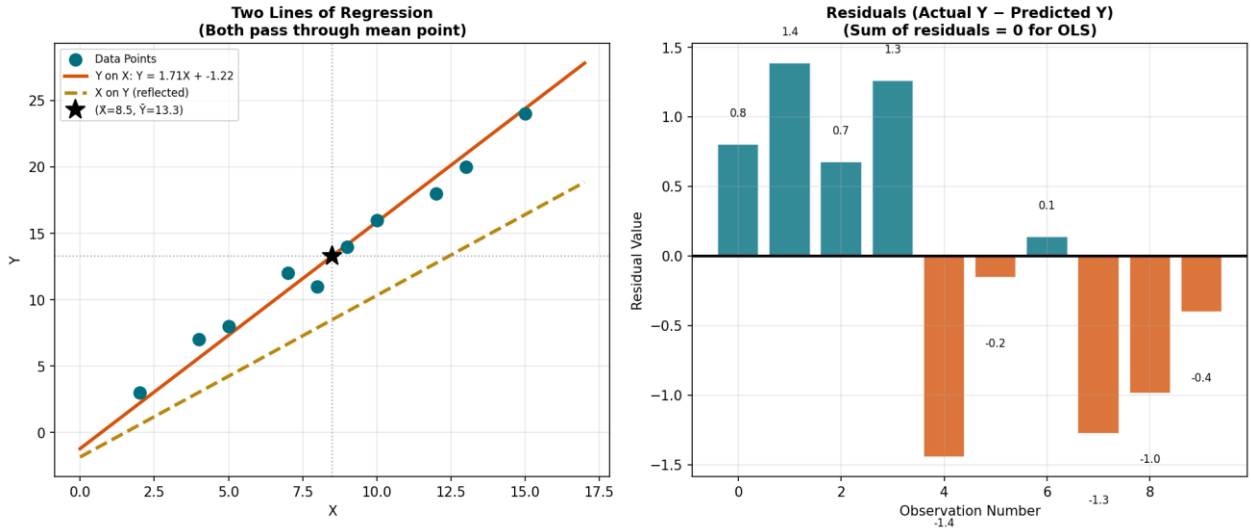


*Figure 3.2: Regression Lines — Line of Y on X (blue), and residuals showing actual vs predicted values*

Regression analysis establishes the mathematical equation of the relationship between variables for prediction. While correlation measures the strength, regression gives the equation to predict one variable from another.

### Two Lines of Regression

There are always two regression lines — one for predicting Y from X, and one for predicting X from Y. Both lines pass through the point ($\bar{X}$, $\bar{Y}$).

$$\text{Line of Y on X:} \quad Y - \bar{Y} = byx(X - \bar{X}) \quad \text{[Used to predict Y given X]}$$

$$\text{Line of X on Y:} \quad X - \bar{X} = bxy(Y - \bar{Y}) \quad \text{[Used to predict X given Y]}$$

### Regression Coefficients

$$byx = r \times (\sigma y / \sigma x) = \Sigma(X-\bar{X})(Y-\bar{Y}) / \Sigma(X-\bar{X})^2 \quad \text{[Slope of Y on X]}$$

$$bxy = r \times (\sigma x / \sigma y) = \Sigma(X-\bar{X})(Y-\bar{Y}) / \Sigma(Y-\bar{Y})^2 \quad \text{[Slope of X on Y]}$$

$$r = \sqrt{(byx \times bxy)} \quad \text{[Relation between r and regression coefficients]}$$

**Properties of Regression Coefficients:**

- byx and bxy always have the SAME SIGN (both + or both −)
- If byx > 1 then bxy < 1 (they cannot both be > 1 unless r = 1)
- $r^2 = byx \times bxy$ (coefficient of determination)
- Regression lines intersect at the point $(\bar{X}, \bar{Y})$ — the mean values
- Regression is not symmetric: byx ≠ bxy in general

## Coefficient of Determination ($r^2$)

$r^2$ measures the proportion of variation in Y that is explained by the variation in X. For example, $r^2$ = 0.81 means 81% of variation in Y is explained by X.

$$r^2 = \text{Explained Variation} / \text{Total Variation} = 1 - (SSE / SST)$$

# MODULE 4: Probability and Theoretical Distributions (8 Hours)

Probability is a numerical measure (between 0 and 1) of the likelihood that an event will occur. It forms the mathematical foundation for statistical inference and decision-making under uncertainty.

## 4.1 Basic Concepts of Probability

| TERM / FORMULA | MEANING & FORMULA |
|---|---|
| **Experiment** | Any action or process that generates outcomes with uncertainty (e.g., rolling a die) |
| **Sample Space (S)** | The set of ALL possible outcomes of an experiment (e.g., S={1,2,3,4,5,6} for a die) |
| **Event (A, B...)** | A subset of the sample space; one or more outcomes of interest |
| **Mutually Exclusive** | Events that cannot occur simultaneously: $P(A \cap B) = 0$ |
| **Exhaustive Events** | Events that cover all possible outcomes: $P(A_1 \cup A_2 \cup ... \cup A_n) = 1$ |
| **Independent Events** | Occurrence of one event does not affect the other: $P(A\|B) = P(A)$ |
| **Complementary Event** | $\bar{A}$ = everything except A; $P(A) + P(\bar{A}) = 1$ |

### Theories of Probability

| Theory | Definition | Formula |
|---|---|---|
| **Classical (A priori)** | Equal probability for each outcome; based on symmetry | P(A) = n(A)/n(S) = Favourable outcomes/Total outcomes |
| **Empirical (Statistical)** | Based on observed frequencies from experiments | P(A) = lim(f/n) as n→∞; relative frequency approach |
| **Subjective** | Based on personal judgment, experience, or belief | No fixed formula; used in business decisions under uncertainty |

## 4.2 Rules of Probability

**Figure 4.1: Probability Rules — Addition and Multiplication**

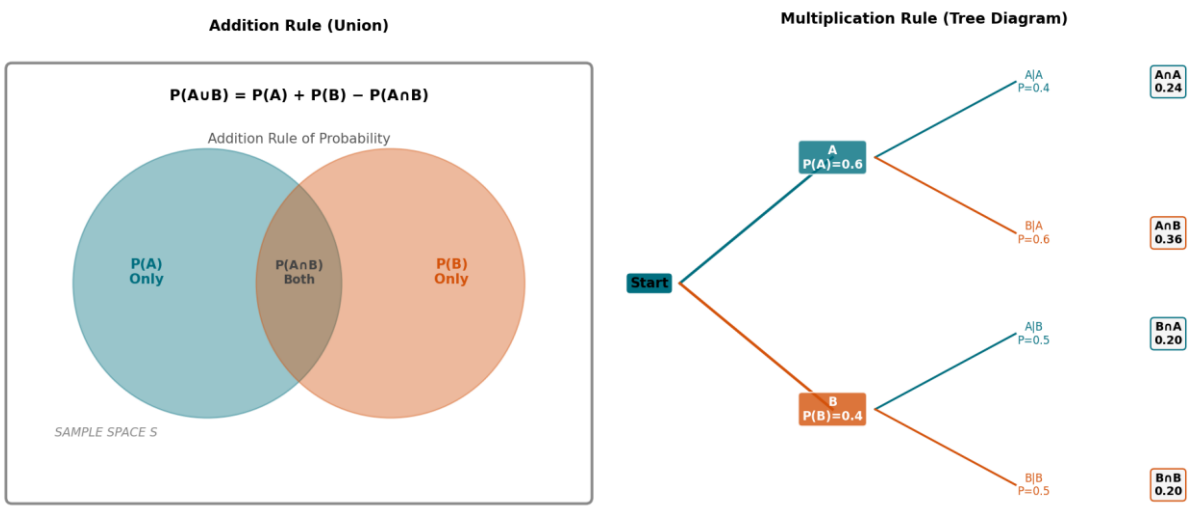


*Figure 4.1: Addition Rule (Venn diagram) and Multiplication Rule (Tree diagram)*

### Addition Rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{[General Addition Rule]}$$

$$P(A \cup B) = P(A) + P(B) \quad \text{[Mutually Exclusive events, since } P(A \cap B)=0\text{]}$$

### Multiplication Rule

$$P(A \cap B) = P(A) \times P(B|A) \quad \text{[General Multiplication Rule]}$$

$$P(A \cap B) = P(A) \times P(B) \quad \text{[Independent events]}$$

### Conditional Probability

$$P(A|B) = P(A \cap B) \,/\, P(B) \quad \text{[Probability of A given B has occurred]}$$

## 4.3 Permutations and Combinations

Permutations and Combinations are counting techniques essential for calculating probabilities in equally likely outcome experiments.

### Permutations (Order MATTERS)

$$^{n}P_{r} = n! \,/\, (n-r)! \quad \text{[Arranging r items from n distinct items]}$$

$$^{n}P_{n} = n! \quad \text{[All n items arranged]}$$

### Combinations (Order does NOT matter)

$$^{n}C_{r} = n! \,/\, [r! \times (n-r)!] \quad \text{[Choosing r items from n items]}$$

$$^{n}C_{0} = {}^{n}C_{n} = 1 \quad | \quad {}^{n}C_{1} = n \quad | \quad {}^{n}C_{r} = {}^{n}C(n-r)$$

**Permutation vs Combination — Key Difference:**

- Permutation: ABC and BAC are DIFFERENT (order matters) → Use for arrangements, rankings, sequences
- Combination: ABC and BAC are SAME (order doesn't matter) → Use for selections, committees, groups
- Example: Seating 3 people on 3 chairs = 3! = 6 permutations
- Example: Selecting 3 people from 5 for a committee = $^{5}C_{3}$ = 10 combinations

## 4.4 Bayes' Theorem

Bayes' Theorem allows us to update probabilities based on new evidence. It is extensively used in business for decision-making, medical diagnosis, and machine learning.

$$P(A_i|B) = P(A_i) \times P(B|A_i) \,/\, \Sigma_j [P(A_j) \times P(B|A_j)]$$

Where: $P(A_i)$ = Prior probability of $A_i$, $P(B|A_i)$ = Likelihood (conditional prob of B given $A_i$), $P(A_i|B)$ = Posterior probability (updated probability after observing B)

**Solved Example — Bayes' Theorem:**

- A factory has 3 machines: M1 (30% of output, 5% defective), M2 (45% of output, 3% defective), M3 (25% of output, 4% defective)
- A product is found defective. What is the probability it came from M2?
- P(Defective) = 0.30×0.05 + 0.45×0.03 + 0.25×0.04 = 0.015 + 0.0135 + 0.01 = 0.0385
- P(M2|Defective) = (0.45 × 0.03) / 0.0385 = 0.0135/0.0385 = 0.3506 = 35.06%

## 4.5 Binomial Distribution

**Figure 4.2: Binomial Distribution — PMF and Properties**

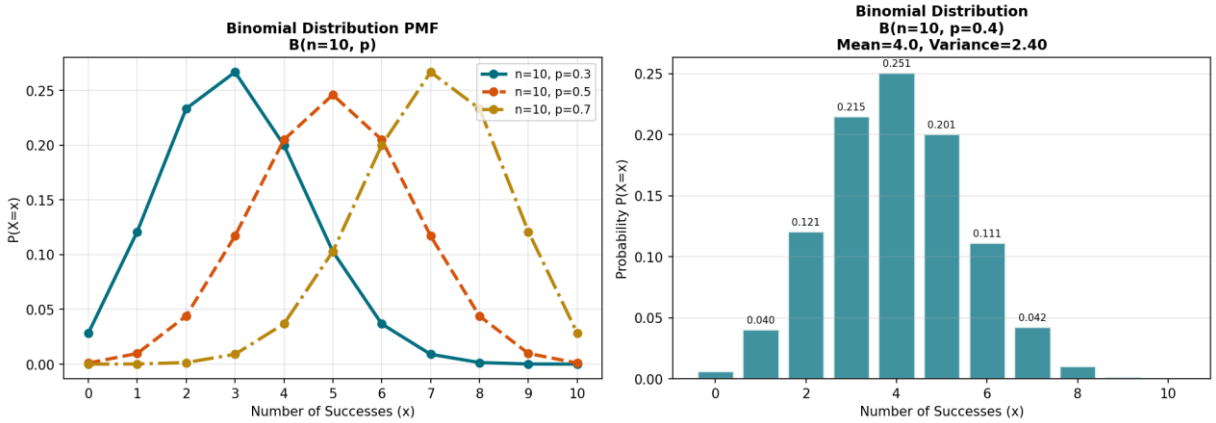


*Figure 4.2: Binomial Distribution — PMF for different p values (n=10) and B(10, 0.4)*

The Binomial Distribution is used when an experiment has exactly two outcomes (Success or Failure), is repeated n times under identical conditions, and each trial is independent with constant probability p of success.

$$P(X = x) = {}^{n}C_{x} \times p^{x} \times q^{n-x} \quad [\text{where } q = 1-p]$$

$$\text{Mean } (\mu) = np$$

$$\text{Variance } (\sigma^2) = npq$$

$$\text{Standard Deviation } (\sigma) = \sqrt{(npq)}$$

**Conditions for Binomial Distribution:**

- n = fixed number of trials (finite and definite)
- Each trial has exactly 2 outcomes: Success (p) or Failure (q=1−p)
- All trials are INDEPENDENT of each other
- p (probability of success) is CONSTANT across all trials
- Examples: Coin tosses, defective items in a batch, customers who buy vs don't buy

## 4.6 Poisson Distribution

**Figure 4.3: Poisson Distribution — PMF and Properties**

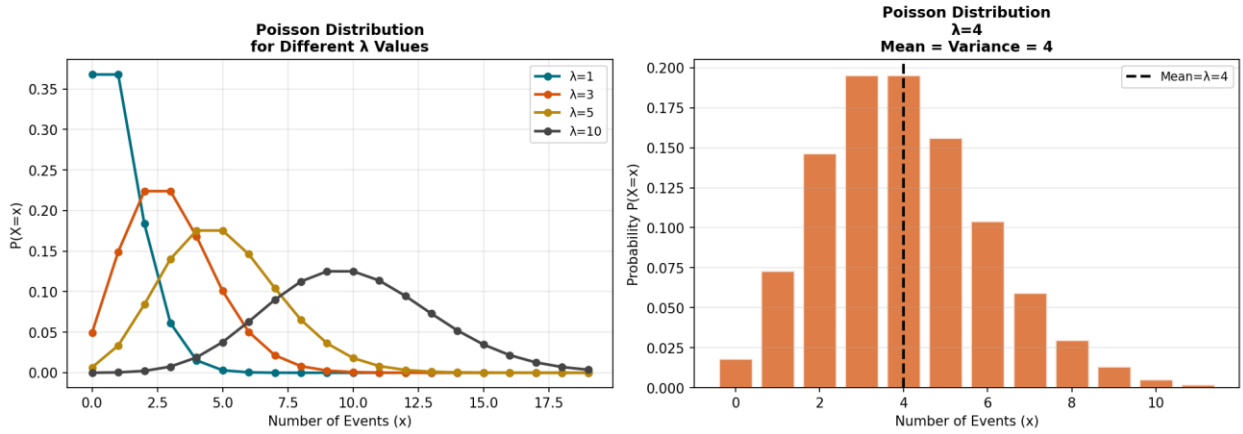


*Figure 4.3: Poisson Distribution — PMF for different λ values; Mean = Variance = λ*

The Poisson Distribution models the number of occurrences of a rare event in a fixed interval of time or space. It is a limiting case of the Binomial distribution when n→∞ and p→0, such that np=λ.

$$P(X = x) = e^{-\lambda} \times \lambda^x / x! \quad [\text{where } \lambda = np = \text{mean number of occurrences}]$$

$$\text{Mean} = \lambda \quad | \quad \text{Variance} = \lambda \quad | \quad \text{SD} = \sqrt{\lambda}$$

**Conditions & Applications of Poisson Distribution:**

- Events occur independently and randomly in a fixed time/space interval
- Mean rate (λ) is constant; events are rare (small p, large n)
- Applications: Number of phone calls per hour, number of accidents per day
- Number of defects per unit of cloth, number of bacteria in a water sample
- Number of customers arriving at a bank in 5 minutes

Poisson as Approximation to Binomial: When n > 50 and p < 0.1 (or np < 5), use Poisson with λ = np.

## 4.7 Normal Distribution

**Figure 4.4: Standard Normal Distribution — Area Calculations**

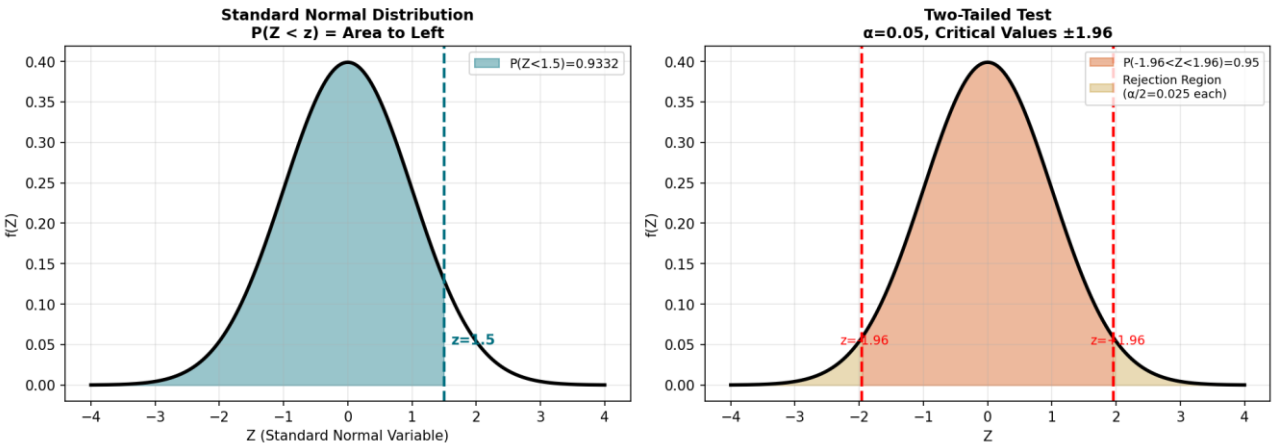


*Figure 4.4: Standard Normal Distribution — Area calculations; P(Z<1.5) and critical values for two-tailed test*

The Normal Distribution is the most important continuous probability distribution. It is symmetric, bell-shaped, and completely described by its mean (μ) and standard deviation (σ).

$$f(x) = [1/(\sigma\sqrt{2\pi})] \times e^{[-(x-\mu)^2/(2\sigma^2)]}$$

$$\text{Standardisation: } Z = (X - \mu) / \sigma \quad [Z \sim N(0,1) - \text{Standard Normal}]$$

**Properties of Normal Distribution:**

- Symmetric about the mean: Mean = Median = Mode = μ
- Bell-shaped curve; total area under curve = 1
- 68.27% of data lies within ±1σ; 95.45% within ±2σ; 99.73% within ±3σ (Empirical Rule)
- Completely defined by two parameters: μ (location) and σ² (spread)
- The standard normal (Z) table gives P(Z ≤ z) for any z value
- Used in: quality control, finance, height/weight distributions, test scores

## Using the Z-table (Standard Normal Table)

Steps: (1) Standardise: Z = (X−μ)/σ, (2) Look up Z value in table to get P(Z < z), (3) Use probability rules for required probability.

| Required Probability | How to Calculate | Formula |
|---|---|---|
| **P(X < a)** | Find Z = (a−μ)/σ; look up table | P(Z < z) directly from table |
| **P(X > a)** | Find Z; use complement | 1 − P(Z < z) |
| **P(a < X < b)** | Find $Z_1$ and $Z_2$ | $P(Z < z_2) - P(Z < z_1)$ |
| **P(X < −z)** | Symmetry of normal curve | P(Z > z) = 1 − P(Z < z) |

# MODULE 5: Testing of Hypotheses (14 Hours)

Hypothesis testing is a statistical procedure for making decisions about population parameters based on sample data. It provides a formal framework for determining whether sample evidence is sufficient to support or reject a claim about the population.

## 5.1 Hypothesis Testing Framework

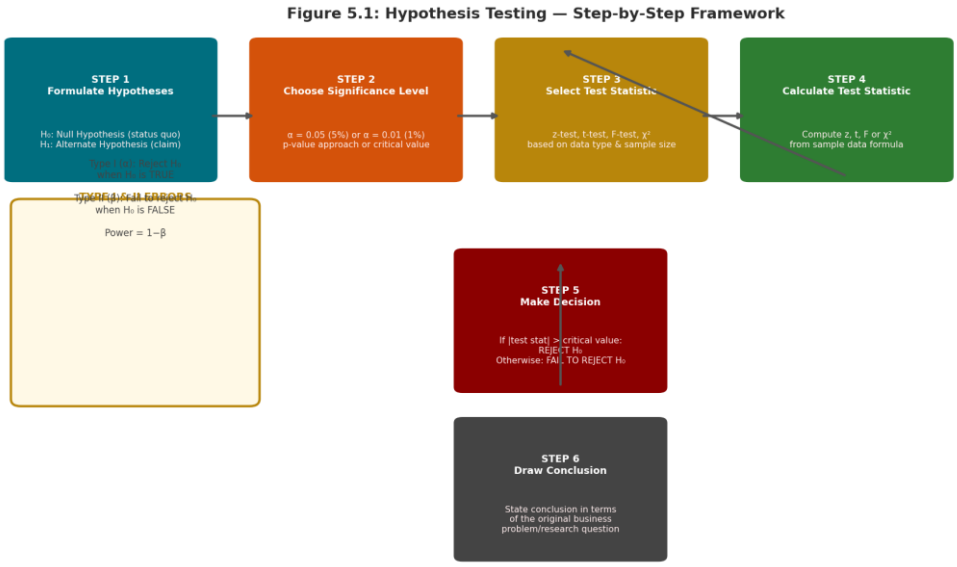


*Figure 5.1: Hypothesis Testing — Step-by-step framework from formulation to conclusion*

### Key Terms

| TERM / FORMULA | MEANING & FORMULA |
|---|---|
| **Null Hypothesis ($H_0$)** | The default assumption; claim of 'no effect' or 'no difference'; assumed TRUE until proven otherwise |
| **Alternate Hypothesis ($H_1$ or $H_a$)** | The research claim; what we want to prove; accepted only when $H_0$ is rejected |
| **Level of Significance (α)** | Probability of making Type I error; typically α = 0.05 (5%) or α = 0.01 (1%) |
| **Critical Value** | The boundary value that separates rejection region from acceptance region |
| **p-value** | Probability of getting test statistic as extreme or more extreme if $H_0$ is true; reject $H_0$ if $p < \alpha$ |
| **Test Statistic** | Calculated value (z, t, F, $\chi^2$) used to make the decision |
| **Degrees of Freedom (df)** | Number of independent values free to vary; affects critical values |

### Type I and Type II Errors

| Decision | $H_0$ is Actually TRUE | $H_0$ is Actually FALSE |
|---|---|---|
| **REJECT $H_0$** | **TYPE I ERROR (α) False Positive Probability = α** | **CORRECT DECISION Power = 1−β** |
| **FAIL TO REJECT $H_0$** | **CORRECT DECISION Confidence = 1−α** | **TYPE II ERROR (β) False Negative Probability = β** |

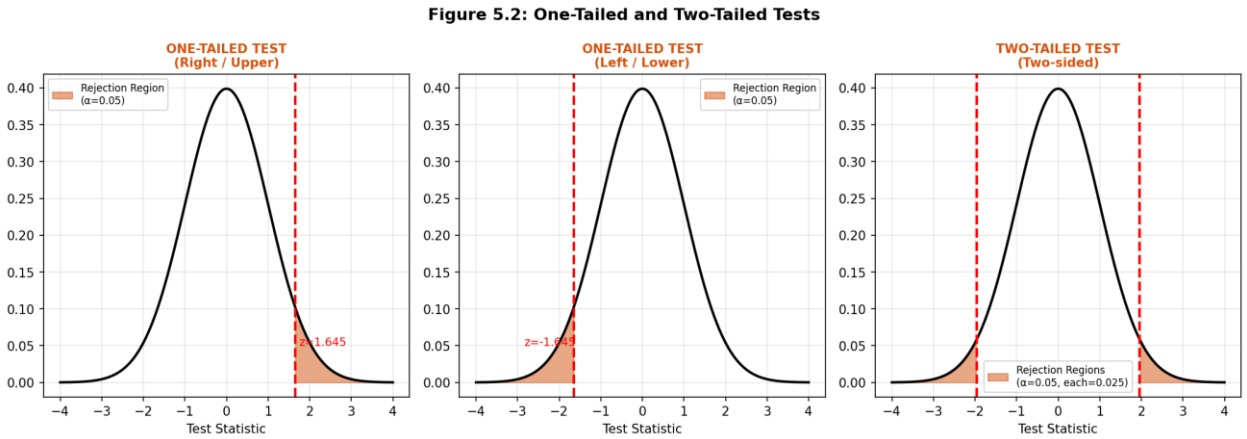


*Figure 5.2: One-Tailed and Two-Tailed Tests — Critical regions and rejection zones*

## 5.2 z-Test (Large Sample Test)

The z-test is used when the sample size is large (n ≥ 30) or when population standard deviation (σ) is known. The test statistic follows a Standard Normal distribution.

### Test for Single Population Mean

$$z = (\bar{X} - \mu) / (\sigma/\sqrt{n}) \quad [\sigma \text{ known}]$$

$$z = (\bar{X} - \mu) / (s/\sqrt{n}) \quad [\sigma \text{ unknown, large } n\text{; use sample } s]$$

### Test for Difference Between Two Population Means

$$z = (\bar{X}_1 - \bar{X}_2) / \sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$$

### Test for Population Proportion

$$z = (\hat{p} - p) / \sqrt{(pq/n)} \quad [\text{where } \hat{p} = \text{sample proportion, } p = \text{hypothesised proportion}]$$

| Test Type | Critical Value (α=0.05) | Decision Rule |
|---|---|---|
| **Two-tailed** | ±1.96 | Reject $H_0$ if $\|z\| > 1.96$ |
| **One-tailed (Right)** | 1.645 | Reject $H_0$ if $z > 1.645$ |
| **One-tailed (Left)** | -1.645 | Reject $H_0$ if $z < -1.645$ |
| **Two-tailed (α=0.01)** | ±2.576 | Reject $H_0$ if $\|z\| > 2.576$ |

## 5.3 t-Test (Small Sample Test)

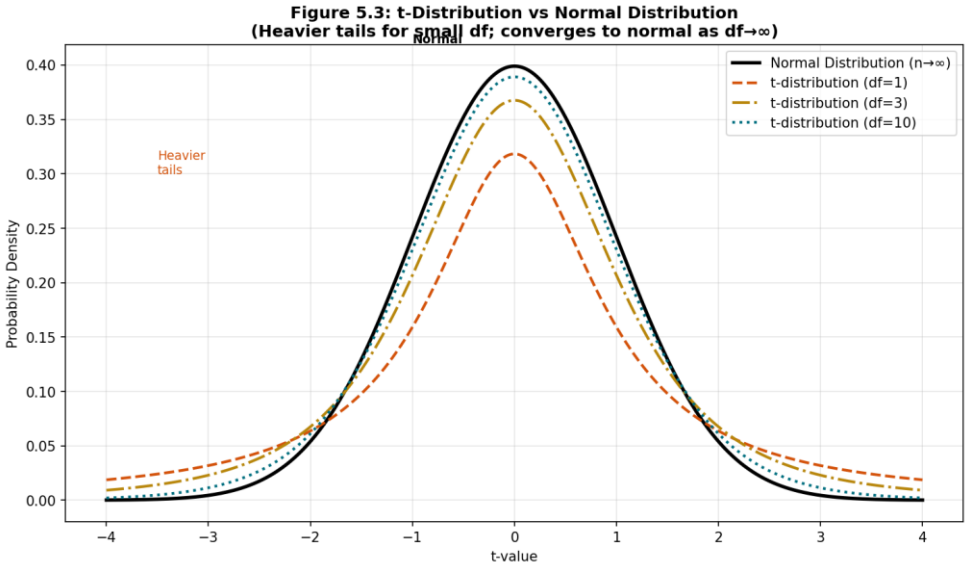


*Figure 5.3: t-Distribution — Heavier tails than normal; converges to normal as degrees of freedom increase*

The t-test is used when the sample size is small (n < 30) AND the population standard deviation is unknown. William Gosset ('Student') developed this distribution.

### One-Sample t-Test

$$t = (\bar{X} - \mu) / (s/\sqrt{n}) \quad \text{with } df = n-1$$

### Two-Sample Independent t-Test (Equal Variances)

$$t = (\bar{X}_1 - \bar{X}_2) / [Sp \times \sqrt{(1/n_1 + 1/n_2)}] \quad \text{with } df = n_1+n_2-2$$

$$\text{Pooled SD: } Sp = \sqrt{[(n_1-1)s_1^2 + (n_2-1)s_2^2] / (n_1+n_2-2)}$$

### Paired t-Test (Matched Samples)

$$t = \bar{D} / (sD/\sqrt{n}) \quad \text{with } df = n-1$$

$$\text{where } D = X_1 - X_2 \text{ (paired differences)}, \quad \bar{D} = \text{mean of differences}$$

**When to use which t-Test:**

- One-sample t: Compare sample mean to a known/hypothesised population mean
- Independent two-sample t: Compare means of two independent groups (e.g., two different cities)
- Paired t: Compare two related measurements on same subjects (before-after studies, matched pairs)
- Assumptions: Normally distributed population, independent observations, equal variances (for 2-sample)

## 5.4 F-Test (Variance Ratio Test)

The F-test compares the variances of two populations. It is also the foundation of ANOVA. The F-statistic is always the ratio of two variances.

$$F = s_1^2 / s_2^2 \quad [\text{Larger variance in numerator, so } F \geq 1]$$

$$df_1 = n_1 - 1 \text{ (numerator df)}, \quad df_2 = n_2 - 1 \text{ (denominator df)}$$

Reject $H_0$: $\sigma_1^2 = \sigma_2^2$ if $F > F\alpha(df_1, df_2)$ from F-table

## 5.5 Analysis of Variance (ANOVA)

**Figure 5.5: ANOVA — Analysis of Variance**

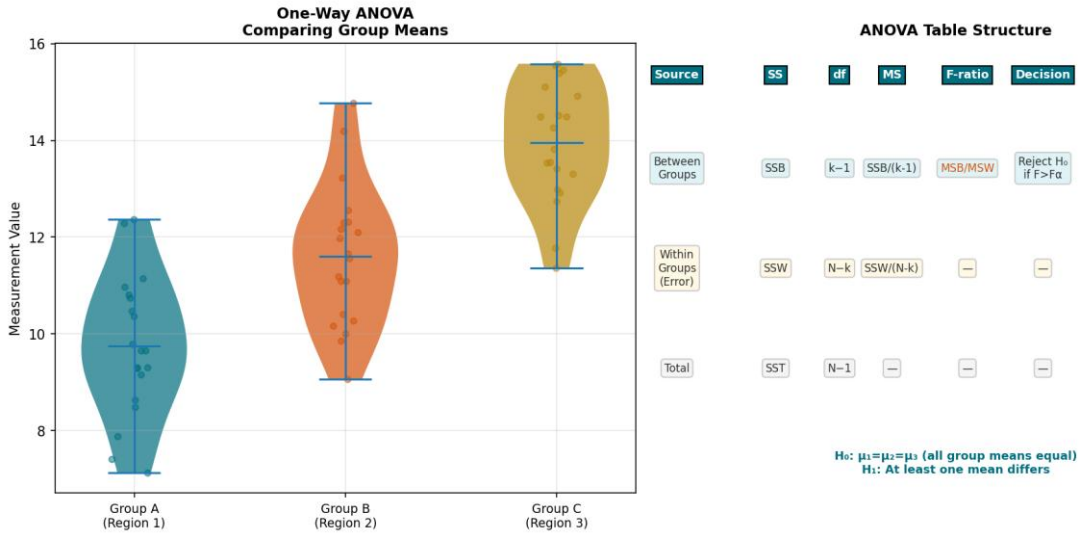


*Figure 5.5: ANOVA — Comparing three group means; ANOVA Table structure*

ANOVA tests whether the means of three or more groups are equal. Instead of doing multiple t-tests (which inflates Type I error), ANOVA tests all groups simultaneously using variance.

### One-Way ANOVA — Testing one factor (k groups)

Hypotheses: $H_0$: $\mu_1 = \mu_2 = ... = \mu_k$ vs $H_1$: At least one mean differs

$$SST = SSB + SSW \quad [\text{Total variation = Between-group + Within-group}]$$

$$SSB = \Sigma n_i (\bar{X}_i - \bar{X})^2 \quad [\text{Between-group Sum of Squares; df = k-1}]$$

$$SSW = \Sigma\Sigma (X_{ij} - \bar{X}_i)^2 \quad [\text{Within-group Sum of Squares; df = N-k}]$$

$$MSB = SSB/(k-1) \quad | \quad MSW = SSW/(N-k)$$

$$F = MSB / MSW \quad [\text{Test Statistic; compare to } F\alpha(k-1, N-k)]$$

**ANOVA Table Format:**

- Source | SS | df | MS | F
- Between Groups | SSB | k−1 | MSB=SSB/(k-1) | F=MSB/MSW
- Within Groups (Error) | SSW | N−k | MSW=SSW/(N-k) | —
- Total | SST | N−1 | — | —
- Decision: If F_calculated > F_table(α, k-1, N-k), REJECT $H_0$

### Two-Way ANOVA — Testing two factors

Two-way ANOVA tests the effect of two independent factors (rows and columns) and their interaction on the dependent variable.

$$SST = SSR + SSC + SSE \quad [\text{Total = Row effect + Column effect + Error}]$$

$$F_{rows} = MSR / MSE \quad | \quad F_{columns} = MSC / MSE$$

## 5.6 Chi-Square Test ($\chi^2$)

**Figure 5.4: Chi-Square Distribution and Test**

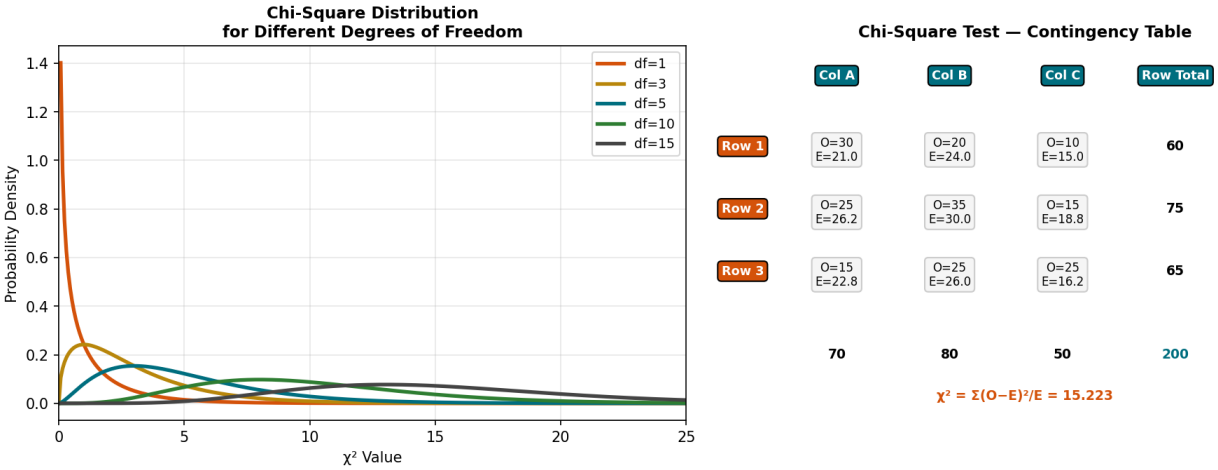


*Figure 5.4: Chi-Square Distribution and Contingency Table — observed vs expected frequencies*

The Chi-Square test is a non-parametric test used for categorical data. It tests whether observed frequencies differ significantly from expected frequencies.

### Chi-Square Test for Goodness of Fit

Tests whether sample data fits a specified distribution (normal, uniform, Poisson, etc.)

$$\chi^2 = \Sigma[(O - E)^2 / E] \quad \text{with } df = k-1-m$$

where O = Observed frequency, E = Expected frequency, k = number of categories, m = number of parameters estimated

### Chi-Square Test for Independence (Association)

Tests whether two categorical variables are independent of each other. Used with contingency tables.

$$\chi^2 = \Sigma[(O - E)^2 / E] \quad \text{with } df = (r-1)(c-1)$$

$$\text{Expected frequency: } E = (\text{Row Total} \times \text{Column Total}) / \text{Grand Total}$$

Yates' Correction (for 2×2 tables): $\chi^2 = \Sigma[(|O - E| - 0.5)^2 / E]$

**Assumptions for Chi-Square Test:**

- Expected frequency in each cell must be ≥ 5 (merge categories if necessary)
- Observations must be independent
- Data must be in frequency (count) form, not percentages
- Applies to categorical (nominal or ordinal) data only

## 5.7 Non-Parametric Tests

Non-parametric tests (distribution-free tests) do not assume any specific distribution of the population. They are used for ordinal or nominal data, small samples, or when parametric assumptions are violated.

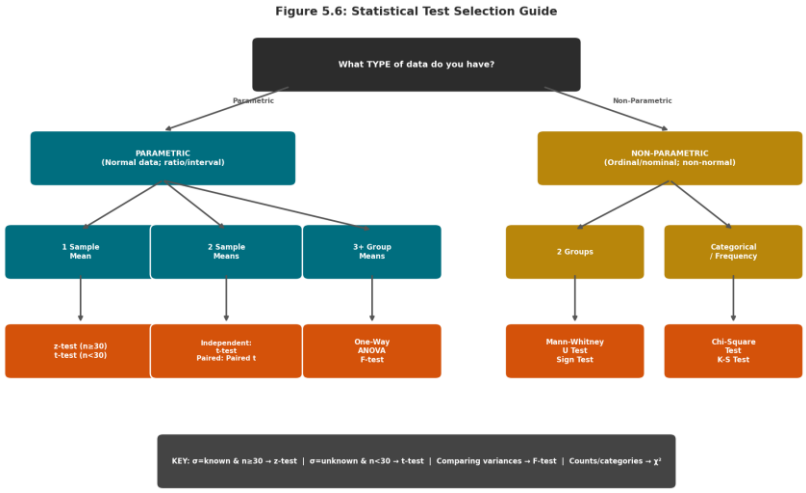


*Figure 5.6: Statistical Test Selection Guide — Parametric vs Non-parametric decision flowchart*

## Sign Test

The Sign Test is the simplest non-parametric test. It tests hypotheses about the population median using only the signs of differences (positive, negative, or zero).

- Replace each observation with + if above median, − if below, 0 if equal
- Count n+ (positive signs) and n− (negative signs); ignore zeros
- Test statistic: Use smaller of n+ or n− and compare with binomial/normal critical value
- Large sample: z = (min(n+, n−) − n/2) / √(n/4)

## Mann-Whitney U Test (Wilcoxon Rank-Sum Test)

The Mann-Whitney U Test is the non-parametric equivalent of the independent samples t-test. It compares two independent groups without assuming normality.

$$U_1 = n_1n_2 + n_1(n_1+1)/2 - R_1 \quad [R_1 = \text{sum of ranks for group 1}]$$

$$U_2 = n_1n_2 + n_2(n_2+1)/2 - R_2 \quad [\text{Check: } U_1 + U_2 = n_1 \times n_2]$$

$$z = (U - \mu_u) / \sigma_u \quad [\text{Large sample; } \mu_u = n_1n_2/2;\ \sigma_u = \sqrt{(n_1n_2(n_1+n_2+1)/12)}]$$

Steps: (1) Combine both samples and rank all values from smallest to largest, (2) Sum ranks for each group, (3) Calculate $U_1$ and $U_2$, (4) Use smaller U as test statistic, (5) Compare with critical value or use z-approximation for large n.

## Median Run Test (Wald-Wolfowitz Runs Test)

Tests whether a sequence of observations is random by counting the number of 'runs' (consecutive values above or below the median).

$$\text{Expected Runs: } E(r) = (2n_1n_2)/(n_1+n_2) + 1$$

$$\text{Variance: } \sigma^2(r) = 2n_1n_2(2n_1n_2 - n_1 - n_2) / [(n_1+n_2)^2(n_1+n_2-1)]$$

$$\text{Test Statistic: } z = [r - E(r)] / \sigma(r)$$

### Kolmogorov-Smirnov (K-S) One-Sample Test

The K-S test compares an observed (empirical) cumulative distribution with a specified theoretical distribution. It is used to test goodness of fit for continuous distributions.

$$D = \max|F_n(x) - F_0(x)|$$

where F_n(x) = empirical CDF from sample, F_0(x) = theoretical CDF. Reject $H_0$ if D > D_critical (from K-S table at significance level α).

| Parametric Tests | Non-Parametric Tests |
|---|---|
| Assume normal/specific distribution | No distributional assumption |
| Use actual measurements (interval/ratio) | Use ranks or signs (ordinal/nominal) |
| More powerful when assumptions met | Less powerful but more flexible |
| Large samples preferred | Suitable for small samples |
| z-test, t-test, F-test, ANOVA | Sign test, Mann-Whitney, KS test, $\chi^2$ |

## 5.8 Summary — Choosing the Right Test

| Situation | Test to Use | Key Formula / Statistic |
|---|---|---|
| **1 sample mean, σ known, large n** | z-test | $z = (\bar{X}-\mu)/(\sigma/\sqrt{n})$ |
| **1 sample mean, σ unknown, small n** | One-sample t-test | $t = (\bar{X}-\mu)/(s/\sqrt{n})$, df=n-1 |
| **2 independent means, small n** | Two-sample t-test | t uses pooled SD, $df=n_1+n_2-2$ |
| **2 matched/paired samples** | Paired t-test | $t = \bar{D}/(sD/\sqrt{n})$, df=n-1 |
| **3 or more group means** | One-Way ANOVA | F = MSB/MSW |
| **2 factors on one variable** | Two-Way ANOVA | Separate F for rows and columns |
| **2 population variances** | F-test | $F = s_1^2/s_2^2$ |
| **Goodness of fit / Independence** | Chi-Square test | $\chi^2 = \Sigma(O-E)^2/E$ |
| **2 independent groups (non-param)** | Mann-Whitney U | Use rank sums |
| **Test median / paired (non-param)** | Sign Test | Count + and − signs |
| **Randomness of sequence** | Median Run Test | Count runs; z-test |
| **Fit to distribution (non-param)** | K-S Test | D = max\|Fn(x)−F0(x)\| |