Main Topic	Sub-topic	Definition	Where to Use	Formulas	Use Case	Example	Main Goal
Descriptive Statistics	Central Tendency	Measures of average	EDA, reporting	Mean = Σx/n, Median, Mode	Summarize sales data	Mean revenue = ₹10K	Summarize data
	Dispersion	Spread of data	Quality control	Range, Variance, SD	Product dimension variability	SD of delivery time = 2 hrs	Measure consistency
	Skewness	Measures the **asymmetry** of a distribution	EDA, distribution check	Skewness = 3(Mean – Median)/SD or using 3rd moment	Income distribution skew	Positive skew (long right tail)	Identify direction of data imbalance
	Kurtosis	Measures the **tailedness** (peakedness) of a distribution	Risk analysis, QC	Excess Kurtosis = $(\mu_4/\sigma^4) - 3$	Financial returns sharp peaks	High kurtosis (leptokurtic)	Detect presence of outliers or heavy tails
Inferential Statistics	Estimation	Estimate population parameters	Survey sampling	CI = Mean ± Z*(SD/√n)	Predict avg. voter turnout	CI = 60% ± 3%	Generalize from sample
	Hypothesis Testing	Test population assumptions using sample data	A/B testing, drug trials, UX tests	p-value, t-test, z-test, chi-square	Compare control vs test group	$p < 0.05$ → reject null hypothesis	Make data-driven decisions
	Confidence Interval	Range of values likely to contain population parameter	Estimating unknown parameters	CI = mean ± Z * (σ / √n)	Estimating average income	CI = ₹52,000 ± ₹2,000	Express uncertainty in estimates
Hypothesis Testing	Null Hypothesis (H0)	Default assumption (no effect or no difference)	All tests begin with H0	–	Assume both drugs work equally	H0: μ1 = μ2	Benchmark to test against
	Alternative Hypothesis (H1)	Contradicts H0, represents a real effect or change	When evidence suggests a change	–	New drug better than old	H1: μ1 ≠ μ2	What we try to prove
	P-value	Probability of observing a test statistic as extreme as sample assuming H0	All statistical tests	P = P(T > t_obs_H0)		Compare with α to decide H0	p = 0.03 < 0.05 → Reject H0

Main Topic	Sub-topic	Definition	Where to Use	Formulas	Use Case	Example	Main Goal
Hypothesis Testing	Significance Level (α)	Threshold below which H0 is rejected	All statistical tests	Common $\alpha = 0.05$, 0.01	Control Type I error rate	$\alpha = 0.05$ means 5% false positive rate	Decide cutoff for significance
	One-Tailed Test	Tests effect in one direction	Direction-specific testing	–	Is new design better than old?	H1: $\mu1 > \mu2$	Test for improvement only
	Two-Tailed Test	Tests effect in both directions	General comparisons	–	Is there any difference?	H1: $\mu1 \neq \mu2$	Check for any change
	Type I Error (α)	Rejecting H0 when it's actually true (false positive)	Significance level	–	False claim drug works	$\alpha = 0.05$	Control false positives
	Type II Error (β)	Not rejecting H0 when it's false (false negative)	Power analysis	–	Miss a real drug effect	$\beta = 0.2$	Control false negatives
	T-Test	Compares means between groups (when population SD is unknown, small sample)	Compare two means	$t = (\bar{X}1 - \bar{X}2) / \sqrt{[(s_1^2/n_1) + (s_2^2/n_2)]}$	Drug A vs Drug B effect	Test if new drug has better mean recovery time	Compare means of small samples
	F-Test	Compares **variances** of two populations	Variance testing before ANOVA	$F = s_1^2 / s_2^2$ (s = sample variance)	Test if two processes have equal variability	F = Var(Group A) / Var(Group B)	Assess equality of variances
	Chi-Square Test	Tests **association** between categorical variables or goodness of fit	Categorical data, independence, fit	$\chi^2 = \Sigma [(O - E)^2 / E]$	Gender vs Preference, Dice fairness	$\chi^2 = 10.2$, df = 4, $p < 0.05$	Test independence or distribution shape
	ANOVA (One-Way)	Compare means across 3 or more groups	Group mean comparison	F = MS_between / MS_within	Compare A/B/C variants	$F = 3.6$, $p < 0.05$ → Significant difference	Generalize t-test to >2 groups
	Z-Test	Compare sample mean to population mean (large n or known σ)	Mean testing with known SD	$z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$	Population testing with known SD	$z = 2.5$, $p < 0.05$ → Significant result	Test population mean with known variance
Sampling Techniques	Simple Random	Equal chance for every item	Surveys	-	General voter polling	Lottery method	Reduce bias

Main Topic	Sub-topic	Definition	Where to Use	Formulas	Use Case	Example	Main Goal
	Stratified	Divide by subgroups	Population analysis	-	Income-based study	Divide by salary ranges	Better representation
	Cluster	Random groups or clusters	Geographical sampling	-	City-wise polling	Random schools in districts	Cost-efficient sampling
	Systematic	Every kth item	Manufacturing, production	-	Quality control	Pick every 10th item	Simple implementation
Probability	Classical	Based on theory	Games, experiments	$P(A) = \text{Favorable/Total outcomes}$	Dice games	$P(6) = 1/6$	Quantify theoretical outcomes
	Empirical	Based on past data	Forecasting	$P = \text{freq}(A)/n$	Weather prediction	$P(\text{Rain}) = 0.6$	Use observed data
	Subjective	Based on belief	Expert systems	-	Stock forecast	$P(\text{Growth}) = 0.8$ (belief)	Model expert opinion
	Conditional	Probability of A given B	Risk analysis	$P(A/B) = P(A \cap B)/P(B)$		Defects by supplier	P(Defect
	Bayes' Theorem	Update beliefs with evidence	Medical diagnosis	$P(A/B) = [P(B/A) \cdot P(A)]/P(B)$			Test accuracy
Descriptive Summary	5-Number Summary	Statistical spread	Boxplots, visualization	-	Understand distribution	[Min, Q1, Median, Q3, Max]	Summarize distribution shape
Combinatorics	Permutations	Arrangements where order matters	Ranking, passwords	$nPr = n! / (n-r)!$	Arrange 3 of 5 books	$P(5,3) = 60$	Count ordered outcomes
	Combinations	Selections where order doesn't matter	Group selection, lottery	$nCr = n! / (r!(n-r)!)$	Choose 3 players from 5	$C(5,3) = 10$	Count unordered outcomes
Distributions	Normal Distribution	Bell-shaped, symmetric curve	Heights, exam scores	PDF formula	Model scores	Mean=Median=Mode	Model natural variation
	Binomial Distribution	Success/failure in fixed trials	Surveys, quality checks	$P(X=k) = C(n,k)p^k(1-p)^{(n-k)}$	Yes/No answers	P(3 heads in 5 flips)	Model binary outcomes
	Poisson Distribution	Count of rare events in interval	Traffic, calls	$P(X=k) = \lambda^k * e^{-\lambda} / k!$	Calls per min	$\lambda = 3 \rightarrow P(2 \text{ calls})$	Model rare events
	Standard Normal Distribution	Normal distribution with $\mu=0$, $\sigma=1$	Z-score analysis	$Z = (x-\mu)/\sigma$	Standardize marks	$Z = 1.5$	Compare across datasets

Main Topic	Sub-topic	Definition	Where to Use	Formulas	Use Case	Example	Main Goal
	Log-Normal Distribution	Distribution of log(x) is normal	Income, biology, stocks	$\log(X) \sim N(\mu,\sigma^2)$	Stock returns, income	log-normal distribution	Model skewed positive data
	Bernoulli Distribution	Only two outcomes: success/failure	Binary outcome modeling	$P(X=1)=p$, $P(X=0)=1-p$	Coin toss, trial success	P(success)=0.5	Model single binary event
Relationships	Correlation	Measures strength and direction of relationship	EDA, feature selection	$r = Cov(X,Y)/(\sigma X \cdot \sigma Y)$	Study time vs marks	r = 0.85 → strong +ve	Identify linear link
	Covariance	Measures direction of joint movement	Portfolio analysis	$Cov(X,Y) = \Sigma(x-\bar{X})(y-\bar{Y})/(n-1)$	Stocks A & B	Cov > 0 → both rise	Track co-movement
Data Scaling	Normalization	Scale data to 0–1 range	ML, image data	$(x-min)/(max-min)$	Pixel scaling	[0,1] range	Uniform scale
	Standardization	Scale data to mean=0, SD=1	ML models, z-scores	$(x-mean)/SD$	PCA, Z-score use	Z = 1.2	Handle different units
	Mean Normalization	Centers values around 0 using mean	ML preprocessing	$(x-mean)/(max-min)$	Salary data scaling	Normalized range: −1 to 1	Center around mean
	Robust Scaling	Uses median and IQR to scale	Skewed/outlier data	$(x-median)/IQR$	Income distribution	Less impacted by outliers	Handle outliers robustly
	Log Transformation	Reduces skew, compresses data range	Right-skewed data	log(x) or log(x+1)	Sales data	log(1000) = 3	Normalize positive skew
	Decimal Scaling	Scales down by powers of 10	Financial data	x / 10^j (j makes	x	< 1)	₹100000 becomes 1
Outlier Detection	Z-Score Method	Measures distance from mean in SDs	Cleaning data	$Z = (x-mean)/SD$	Remove extreme points	Z = 4 → outlier	Detect distant values
	IQR Method	Uses quartiles to detect extremes	Boxplots, visuals	Outlier: < Q1−1.5*IQR or > Q3+1.5*IQR	Identify anomalies	Value > Q3 + 1.5×IQR	Remove statistical extremes