Guide on Statistical Methods

Author: Kamil Rafikov
Email: mailbox@kamil-rafik.com

Updated at September 18th, 2026.
Edition: 9.

Table of Contents

Introduction

Many years ago, when I studied in university, and later, when I entered independent research field, I looked for a simple short guide on statistics that provides full review of all types of data and methods that can be applied for analysis of these data. I did not find any and decided to write it by myself.

In the process of writing the guide few AI systems were used as search engines and validators. All information given further must be rechecked in official mathematical and research tutorials written by professional mathematicians and researchers. I reviewed in details only minor part of knowledge included into the guide. (Full final auto-validation finds about 10 issues in the text that I cannot check because of lack of qualification. Feel free to do it by yourself. However, I guess it is acceptable amount for a guide with such density of information.)

All used abbreviations of data types except MCAR, MAR, MNAR are self-invented; they were necessary to make the text maximally compact.

The guide should be used for introduction purposes only to understand the full range of available statistical research methods.

It is the final edition. Enjoy!

Classification of data

Scale and nature

NAME	ABBR.
Qualitative nominal	QN
Qualitative nominal binary	QNB
Qualitative ordinal	QO
Quantitative discrete	QD
Quantitative discrete compositional	QDC
Quantitative continuous	QC
Quantitative continuous compositional	QCC
Quantitative interval without true zero point, discrete	QID
Quantitative interval without true zero point, continuous	QIC
Quantitative ratio with true zero point, discrete	QRD
Quantitative ratio with true zero point, continuous	QRC

Mathematical structure

NAME	ABBR.
Directional	DIR
Functional	FUN

Origin and generation mode

NAME	ABBR.
Empirical	EMP
Experimental	EXP
Synthetic	SYN
Streaming	STR
Administrative/register data	ADM

Observational and temporal strucure

NAME	ABBR.
Cross-sectional	CS
Time-series	TS
Longitudinal/Panel	LP
Pooled cross-sectional	PCS
Time-to-event	TTE

Structural format

NAME	ABBR.
Structured	ST
Semi-structured	SST
Unstructured	UST

Dimensionality and variables count

NAME	ABBR.
Univariate	UV
Bivariate	BV
Multivariate	MV
High-dimensional with low sample size	HDLSS
Ultra high-dimensional	UHD

Relational and geometric topology

NAME	ABBR.
Graphs and networks	GN
Topological	TOP
Non-Euclidian/Manifold	NEUC
Tensor	TEN
Point pattern	PP
Sequence	SEQ

Completeness state

NAME	ABBR.
Fully observed	FO
Censored	CEN
Truncated	TRU
Missing randomly	MCAR
Missing depending on observed variable	MAR
Missing depending on missing value itself	MNAR

Domain semantics

Only few examples are given to provide maximal coverage for methods.

NAME	ABBR.
Geospatial	GS
Biometric	BIO
Financial	FIN
Behavioral	BEH
Epidemiological	EPI
Industrial	IND
Psychometric	PSY
Political	POL
Sports	SPT

Uncertainty

NAME	ABBR.
Stochastic/probabilistic	STO
Rough-set-style imprecison	RST
Quantum	QUA
Fuzzy	FUZ
Interval-valued	IV
Distributional	DIS
Possibilistic/Dempster-Shafer	POSS

Unobservability

NAME	ABBR.
Latent	LAT
Counterfactual	CFAC

Methods

Methods to apply for analysis depending on data type

DATA TYPE	METHODS	DIAGNOSTIC APPROACH TO CHECK QUALITY OF RESULTS

QN	Chi-Square Test, Multinomial Logistic Regression, Frequency Mode, Naive Bayes/kNN, Apriori/association rules	Chi-Square: expected cell count check, standardized residuals; Multinomial Logistic Regression: pseudo-R^2 (McFadden), classification accuracy; Naive Bayes/kNN: cross-validated accuracy, confusion matrix; Apriori: support/confidence/lift thresholds, rule stability across resamples
QNB	Binary Logistic Regression, Fisher's Exact Test, Odds Ratio	Binary Logistic Regression: ROC-AUC, Hosmer-Lemeshow test, confusion matrix; Fisher's Exact Test: exact p-value (no large-sample approximation to check); Odds Ratio: confidence interval width, consistency across strata
QO	Ordinal Logistic Regression, Spearman Rank Correlation, Median, Mann-Whitney Test, Wilcoxon Test, Kruskal-Wallis Test, Friedman's Test	Ordinal Logistic Regression: proportional odds test (Brant test), pseudo-R^2; Spearman: significance test, monotonicity check via scatter/rank plot; Mann-Whitney/Wilcoxon/Kruskal-Wallis/Friedman: rank-sum test statistic and p-value, effect size (rank-biserial correlation or epsilon-squared)
QD	Poisson Regression, Negative Binomial Regression, Descriptive Stats	Poisson Regression: overdispersion check (variance/mean ratio), deviance goodness-of-fit; Negative Binomial Regression: likelihood ratio test vs. Poisson, dispersion parameter significance; Descriptive Stats: no formal diagnostic (validity check is data completeness/outlier scan)
QDC	Dirichlet Multinomial Regression, Compositional Data Analysis (CoDa)	Dirichlet Multinomial Regression: goodness-of-fit via deviance, overdispersion parameter check; CoDa: log-ratio residual inspection, compositional balance checks
QC	Linear Regression, ANOVA, T-Test, Kernel Density Estimation, GAM, Spline/LOESS, Bayesian regression/MCMC, Hierarchical Bayesian models	Linear Regression: residual plots, Q-Q plot, R^2/adjusted R^2, Durbin-Watson; ANOVA: residual normality, Levene's test for variance homogeneity, F-statistic; T-Test: normality check, variance equality (Levene's); Kernel Density Estimation: bandwidth sensitivity check, cross-validated bandwidth selection; GAM: effective degrees of freedom, deviance explained, residual plots; Spline/LOESS: cross-validated smoothing-parameter selection, residual plots; Bayesian regression/MCMC: trace plots, R-hat convergence statistic, effective sample size; Hierarchical Bayesian models: same as MCMC plus posterior predictive checks

QCC	Aitchison Geometry, Log-ratio Transformations, CoDa Methods	Residual inspection in log-ratio space, compositional closure/balance verification
QID	Linear Regression, Factor Analysis, Mean/Standard Deviation	Linear Regression: residual plots, Q-Q plot, R^2; Factor Analysis: KMO test, Bartlett's test of sphericity, eigenvalue scree plot (approximate fit for discrete integer input); Mean/SD: no formal diagnostic (check distributional shape/outliers)
QIC	Linear Regression, Factor Analysis, Mean/Standard Deviation	Same as QID, with cleaner fit − normality/linearity diagnostics apply more directly to continuous data
QRD	Geometric Mean (for positive value only), Coefficient of Variation, Linear Regression	Geometric Mean/CoV: positivity check (verify no zero/negative values) before computing; Linear Regression: residual plots, Q-Q plot, R^2
QRC	Geometric Mean, Coefficient of Variation, Linear Regression	Geometric Mean/CoV: positivity check (verify no zero/negative values) before computing; Linear Regression: residual plots, Q-Q plot, R^2
CS	OLS Regression, K-Means Clustering, Principal Component Analysis, DBSCAN/hierarchical clustering	OLS Regression: residual plots, R^2, VIF (multicollinearity); K-Means: silhouette score, elbow method, Davies-Bouldin index; PCA: scree plot, cumulative explained variance; DBSCAN/hierarchical: silhouette score, dendrogram inspection, cophenetic correlation
TS	ARIMA, Exponential Smoothing, GARCH, LSTM Neural Networks, VECM/cointegration	ARIMA: ACF/PACF of residuals, Ljung-Box test, AIC/BIC; Exponential Smoothing: residual autocorrelation check, forecast error metrics (MAPE, RMSE); GARCH: standardized residual diagnostics, ARCH-LM test; LSTM: validation loss curves, out-of-sample forecast error; VECM/cointegration: Johansen trace/eigenvalue test, residual stationarity check
LP	Fixed Effects, Random Effects, Generalized Estimating Equations, Bayesian regression/MCMC, Hierarchical Bayesian models	Fixed vs. Random Effects: Hausman test; GEE: QIC (quasi-likelihood information criterion), working correlation structure check; Bayesian/Hierarchical: trace plots, R-hat, posterior predictive checks
PCS	Pooled OLS Regression, Difference-in-Differences (DiD)	Pooled OLS: residual diagnostics, poolability test (Chow test); DiD: parallel-trends pre-treatment check, placebo/falsification tests
TTE	Kaplan-Meier Survival Estimator, Cox Proportional Hazards Model	Kaplan-Meier: log-rank test for group comparison, confidence bands; Cox Model: Schoenfeld residuals (proportional-hazards assumption), concordance index (C-index)

ST	SQL (it is a language but from a research perspective it may be considered as analysis method), Tabular Classifiers (XGBoost, Random Forests)	Tabular Classifiers: cross-validated accuracy/F1/AUC, feature importance stability, confusion matrix
SST	JSON/XML Flattening, NoSQL Aggregation, Regex Processing	Validation against schema/sample checks rather than statistical fit
UST	LLMs, Word Embeddings, CNNs, TF-IDF Text Mining	LLMs: held-out perplexity, task-specific accuracy/F1; Word Embeddings: intrinsic evaluation (analogy tasks, cosine similarity benchmarks); CNNs: validation accuracy/loss curves, confusion matrix; TF-IDF: downstream task performance (no intrinsic fit statistic)
UV	Descriptive Statistics, Histograms, Outlier Z-Score Detection	No formal model-fit diagnostic (these are exploratory tools); validity check is sample size adequacy and visual inspection
BV	Pearson Correlation, Simple Linear Regression, Scatter Plots	Pearson Correlation: significance test, linearity check via scatterplot; Simple Linear Regression: residual plots, R^2, Q-Q plot; Scatter Plots: visual inspection only (no formal diagnostic)
MV	Multiple Regression, Factor Analysis, MANOVA, Random Forests, SVM	Multiple Regression: residual plots, R^2/adjusted R^2, VIF; Factor Analysis: KMO, Bartlett's test, scree plot; MANOVA: Box's M test, Wilks' Lambda; Random Forests: out-of-bag error, feature importance, cross-validated accuracy; SVM: cross-validated accuracy/F1, margin/support-vector inspection
HDLSS	LASSO / Ridge Regression, Elastic Net, Regularized PCA, Bonferroni/FDR	LASSO/Ridge/Elastic Net: cross-validated prediction error, regularization-path stability, selected-variable consistency across resamples; Regularized PCA: cross-validated explained variance; Bonferroni/FDR: family-wise error rate check, adjusted p-value distribution
UHD	Sure Independence Screening (SIS), Penalized Regressions, Bonferroni/FDR	SIS: screening consistency across resamples; Penalized Regressions: same as LASSO/Ridge above; Bonferroni/FDR: same as HDLSS above

GS	Kriging, Spatial Autocorrelation (Moran's I), Geographically Weighted Regression, GAM, Spline/LOESS	Kriging: leave-one-out cross-validation, variogram fit quality; Moran's I: significance test (permutation-based), spatial residual mapping; GWR: local R^2, residual spatial autocorrelation check; GAM: effective degrees of freedom, deviance explained, residual plots; Spline/LOESS: cross-validated smoothing-parameter selection, residual plots
BIO	Survival Methods, Mixed Models for Repeated Biological Measures, Meta-analysis	Survival Methods: Schoenfeld residuals, C-index (as TTE); Mixed Models: residual plots by cluster, likelihood ratio test for random effects; Meta-analysis: heterogeneity statistics (I^2, Q-test), funnel plot for publication bias
FIN	Black-Scholes Model, Value at Risk (VaR), GARCH Forecasting, VECM/cointegration	Black-Scholes: backtesting against observed option prices, implied volatility consistency; VaR: backtesting (Kupiec test, exception counts); GARCH/VECM: standardized residual diagnostics, ARCH-LM test
BEH	Cohort Analysis, A/B Testing, Markov Chain Models, Apriori/association rules, Reinforcement learning	Cohort Analysis: retention-curve stability checks; A/B Testing: statistical power check, significance test (Z-test/Chi-Square); Markov Chain Models: transition-matrix stability, goodness-of-fit to observed transitions; Apriori: support/confidence/lift; Reinforcement Learning: cumulative reward curves, convergence of policy/value function
EPI	Compartmental Models (SIR/SEIR), Basic Reproduction Number (R_0) Estimation	Compartmental Models: fit residuals against observed case counts, parameter identifiability check; R_0 Estimation: sensitivity analysis to generation-interval assumptions, confidence interval width
IND	Statistical Process Control (Control Charts, Cpk/Ppk), Six Sigma Methodology	Control Charts: out-of-control point detection (Western Electric rules), false-alarm rate check; Cpk/Ppk: normality check of underlying process data (required for validity of the capability index)
PSY	Classical Test Theory Reliability (Cronbach's Alpha, Test-Retest/Inter-Rater Reliability)	Cronbach's Alpha: item-total correlation inspection, alpha-if-item-deleted check; Test-Retest/Inter-Rater: intraclass correlation coefficient (ICC) confidence interval
POL	Multilevel Regression with Poststratification (MRP)	Comparison of poststratified estimates against known population benchmarks (validation elections/censuses), cell-size adequacy check in poststratification cells

SPT	Bradley-Terry Model, Elo Rating System	Bradley-Terry: goodness-of-fit via deviance, out-of-sample prediction accuracy on held-out matchups; Elo: calibration check (predicted win probability vs. observed win rate), rating convergence stability
DIR	Directional Statistics, Von Mises Distribution Analysis	Goodness-of-fit test for circular distribution (e.g., Rayleigh test for uniformity), circular Q-Q plot
FUN	Functional Data Analysis (FDA), Functional PCA	FDA: cross-validated smoothing parameter, residual curve inspection; Functional PCA: proportion of variance explained by functional components
GN	Graph Neural Networks (GNNs), PageRank, Centrality Metrics	GNNs: held-out link-prediction accuracy, node-classification F1; PageRank: convergence check (rank stability across iterations); Centrality Metrics: sensitivity analysis to network perturbation
TOP	Topological Data Analysis (TDA)	Stability of persistence diagrams under noise/resampling (bottleneck/Wasserstein distance between diagrams)
NEUC	Manifold Learning, Isomap, t-SNE, UMAP Dimensionality Reduction	Trustworthiness/continuity metrics, reconstruction error, stability across random seeds/perplexity settings
TEN	Tensor Decomposition (CP/Tucker), Multi-way ANOVA	Tensor Decomposition: reconstruction error, core consistency diagnostic (CORCONDIA); Multi-way ANOVA: residual diagnostics, F-statistics per factor
PP	Ripley's K-function, Spatial Poisson Point Process Models	Ripley's K: simulation envelope (Monte Carlo confidence bands) comparison against complete spatial randomness; Point Process Models: residual point-pattern diagnostics
SEQ	Hidden Markov Models (HMM), Sequence Alignment (Needleman-Wunsch, BLAST), N-gram Models	HMM: log-likelihood convergence, held-out perplexity; Sequence Alignment: alignment score significance (E-value in BLAST); N-gram Models: held-out perplexity
EMP	Observational Studies, Descriptive Analysis	No formal statistical diagnostic − validity concerns are confounding and selection bias checks (design-level, not model-level)
EXP	Randomized Controlled Trials (RCT), Full/Fractional Factorial ANOVA	RCT: balance check across arms, intention-to-treat vs. per-protocol consistency; Factorial ANOVA: residual diagnostics, interaction-effect significance

SYN	Generative Model Evaluation (FID Score), Augmentation Tests	FID Score: comparison against baseline/reference distributions; Augmentation Tests: downstream task performance with vs. without augmented data
STR	Online Machine Learning, Sliding Window Algorithms, Hoeffding Trees, Reinforcement learning	Online ML/Hoeffding Trees: prequential (online) accuracy tracking, concept-drift detection; Sliding Window: window-size sensitivity check; Reinforcement Learning: cumulative reward curves, convergence of policy/value function
ADM	Record Linkage, Small Area Estimation, Register-Based Cohort Analysis	Record Linkage: precision/recall against a gold-standard linked sample; Small Area Estimation: mean squared error of estimates, benchmarking against direct estimates
FO	Standard Parametric/Non-parametric Statistical Frameworks	Diagnostic depends entirely on which specific method is chosen (see corresponding row)
CEN	Survival Analysis, Tobit Regression, Accelerated Failure Time	Survival Analysis: as in TTE row; Tobit: likelihood ratio test, residual diagnostics adapted for censoring; AFT: residual diagnostics, distributional assumption check (e.g., Weibull fit)
TRU	Truncated Regression Models, Maximum Likelihood Estimation	Truncated Regression: likelihood ratio test, residual diagnostics adapted for truncation; MLE: convergence check, information matrix (standard error) validity
MCAR	Multiple Imputation by Chained Equations (MICE), EM Algorithm	MICE: convergence diagnostics (trace plots across imputations), comparison of imputed vs. observed distributions; EM Algorithm: log-likelihood convergence across iterations
MAR	Inverse Probability Weighting, Full Information ML (FIML)	IPW: covariate balance check after weighting, extreme-weight inspection; FIML: likelihood ratio test, convergence check
MNAR	Heckman Selection Models, Pattern-Mixture Models	Heckman: significance of inverse Mills ratio, exclusion-restriction validity; Pattern-Mixture: sensitivity analysis across assumed missingness patterns
STO	Monte Carlo Simulation, Stochastic Process Modeling, Markov Chain Analysis	Monte Carlo: convergence check (simulation count sensitivity), variance of estimate; Stochastic Process/Markov Chain: goodness-of-fit to observed transitions, stationarity check
RST	Rough Set Theory, Rule Induction, Lower/Upper Approximation Analysis	Rule accuracy/coverage on held-out data, consistency of approximation boundaries across resamples

QUA	Quantum Machine Learning, Quantum State Tomography	Quantum ML: held-out task accuracy; State Tomography: fidelity between reconstructed and true quantum state
FUZ	Fuzzy Logic Controllers, Fuzzy C-Means Clustering	Fuzzy Logic Controllers: performance validation against expected control outcomes; Fuzzy C-Means: partition coefficient, partition entropy, silhouette-style validity indices
IV	Interval Arithmetic, Robust Interval Regression	Interval coverage check (does predicted interval contain observed truth), interval width calibration
DIS	Wasserstein Distance Metrics, Information Geometry, Bayesian Inference	Wasserstein Distance: stability across resampling; Bayesian Inference: as in QC row (trace plots, R-hat, posterior predictive checks)
POSS	Dempster-Shafer Evidence Theory, Possibility Theory, Belief/Plausibility Function Combination	Consistency of combined belief/plausibility bounds across evidence sources, sensitivity to conflicting-evidence weighting
LAT	Structural Equation Modeling (SEM), Factor Analysis, HMMs	SEM: fit indices (CFI, RMSEA, SRMR), chi-square goodness-of-fit; Factor Analysis: KMO, Bartlett's test, scree plot; HMMs: log-likelihood convergence, held-out perplexity
CFAC	Propensity Score Matching, Causal Inference, Doubly Robust Estimation	Propensity Score Matching: covariate balance (standardized mean differences), common support overlap; Doubly Robust Estimation: sensitivity to model misspecification in either the outcome or propensity model

Methods to apply for analysis depending on data types combinations

As you can guess, number of combinations of data types may be overwhelming. That’s why only 100+ the most common combinations are provided further.

For analysis of some combinations of data different methods may be applied depending on the nature of data.

COMBINATION	EXAMPLES OF REAL WORLD DATA	METHODS
QC, CS, ST, MV, EMP, FO	Retail store daily sales figures, Real estate property valuation lists, Annual census demographic records	Multiple Linear Regression, Principal Component Analysis (PCA), K-Means Clustering
QNB, CS, ST, MV, EMP, FO	User newsletter subscription statuses (Yes/No), Customer churn records, Credit card default indicators (Pass/Fail)	Binary Logistic Regression, Random Forest Classifier, Chi-Square Test

QC, TS, ST, MV, FIN, EMP, FO	Daily closing stock prices, Historical currency exchange rates, Quarterly corporate revenue timelines	ARIMA/SARIMAX, GARCH Volatility Modeling, Vector Autoregression (VAR), LSTM
UST, CS, MV, BEH, EMP, FO	Product review text blocks, Customer service call transcriptions, Brand mention tweets	Sentiment Analysis, TF-IDF, Topic Modeling (LDA), LLM Embeddings
TTE, CS, ST, MV, BIO, EXP, CEN	Cancer clinical trial survival times, Medical hardware component failure rates, Clinical patient disease remission tracks	Cox Proportional Hazards Model, Kaplan-Meier Estimator, Log-Rank Test
QC, LP, ST, MV, FIN, EMP, MCAR	Multi-year corporate balance sheets with missing quarters, Multi-year household income survey panel data	Fixed Effects Models, Random Effects Models, MICE (Multiple Imputation)
QD, CS, ST, MV, BEH, EMP, FO	Weekly website page-view counts per user, Number of support tickets submitted per client, Count of monthly items purchased per user	Poisson Regression, Negative Binomial Regression, Descriptive Statistics
SST, CS, MV, BEH, EMP, FO	E-commerce product catalog metadata (in JSON), Web application server system logs (semi-structured), User profile configurations	JSON Flattening, NoSQL MapReduce Aggregations, XGBoost
QO, CS, ST, BV, BEH, EMP, FO	Five-star mobile app user satisfaction ratings, Customer feedback survey Likert scales	Ordinal Logistic Regression, Spearman Rank Correlation, Median Analysis
QC, CS, ST, MV, GS, EMP, FO	Global weather station temperature readings, Urban real estate coordinates and pricing, Regional air pollution density maps	Kriging Interpolation, Spatial Autocorrelation (Moran’s I), Spatial Regression
QC, STR, ST, MV, FIN, EMP, FO	Live high-frequency stock exchange ticker streaming, Real-time cryptocurrency order book changes	Online Stochastic Gradient Descent, Hoeffding Trees, Sliding Window Averaging
QNB, EXP, ST, BV, BEH, FO	Marketing email A/B test conversion outputs, Website user interface button-click clickthrough rates	Two-Sample Proportions Z-Test, Fisher's Exact Test, Chi-Square Test
UST, TS, MV, BEH, EMP, FO	Historical company email archival trail datasets, Weekly corporate support chat history logs	Time-Series Text Mining, Dynamic Topic Modeling, Recurrent Neural Networks
QC, TS, ST, UV, FIN, EMP, FO	Historical daily cryptocurrency exchange valuation track, Historical national monthly inflation indices	Exponential Smoothing, Automated ARIMA, Unit Root Testing (ADF)

QC, CS, ST, MV, BIO, EMP, FO	Standard electronic health laboratory lipid panels, Patient body mass index (BMI) baseline distributions	Factor Analysis, ANOVA, Multivariate Linear Regression
GN, CS, ST, MV, BEH, EMP, FO	Professional connection networks, User follower-following relational maps, Corporate supply chain dependency maps	Graph Neural Networks (GNN), PageRank, Centrality/Community Detection
QCC, CS, ST, MV, EMP, FO	Household monthly spending budget allocation percentages, Corporate asset portfolio distribution ratios	Log-ratio Transformations (clr/ilr), Dirichlet Regression
QC, PCS, ST, MV, EMP, FO	Multi-city annual employment survey combinations, Aggregated state-level education performance logs	Pooled OLS Regression, Difference-in-Differences (DiD) Estimation
QC, CS, ST, HDLSS, BIO, EXP, FO	Gene expression microarrays with low patient volume, Rare genetic mutation biomarker sequencing maps	LASSO Regression, Ridge Regression, Elastic Net, Regularized PCA
QNB, CS, ST, MV, FIN, EMP, MNAR	Loan application records missing income data selectively, High-net-worth individual banking survey portfolios	Heckman Selection Correction Model, Pattern-Mixture Modeling
PP, CS, ST, BV, GS, EMP, FO	Urban vehicle traffic accident crash location coordinates, Commercial retail chain store geographic footprints	Ripley's K-Function, Spatial Poisson Point Process Modeling
QC, TS, ST, MV, FIN, EMP, CEN	Stock trade execution delays bounding at microsecond minimums, Top-coded financial wealth index thresholds	Tobit Time-Series Regression, Censored Regression Analysis
QC, CS, ST, MV, CFAC, EMP, FO	Workforce job training program salary outcome evaluations, Policy adjustment economic performance metrics	Propensity Score Matching (PSM), Inverse Probability Weighting
QC, CS, ST, MV, LAT, EMP, FO	Customer brand loyalty survey index metrics, Employee workplace burnout diagnostic matrixes	Structural Equation Modeling (SEM), Exploratory Factor Analysis
BIO, CS, UST, MV, EXP, FO	MRI structural neuroimaging file sets, Clinical patient chest X-ray digital images	Convolutional Neural Networks (CNN), Vision Transformers (ViT)
QC, TS, ST, MV, BEH, SYN, FO	Generative AI synthesized web traffic load behaviors, Simulated consumer market transaction tracking rows	Generative Adversarial Network (GAN) Evaluation, FID Metric Testing
QC, CS, ST, MV, GS, EMP, MAR	Regional agricultural crop yield survey logs with regional gaps, Residential property pricing missing rural data points	Inverse Probability Weighting, Full Information ML (FIML)

QD, TS, ST, MV, BEH, EMP, FO	Daily website security login failure count metrics, Hourly warehouse product shipment tracking items	Integer-Valued Autoregressive (INAR) Models, Poisson VAR
TEN, TS, ST, MV, FIN, EMP, FO	Global macroeconomic country-by-industry trade flows, Multi-channel algorithmic financial market indices	Tensor Decomposition (CP / Tucker format), Multi-way ANOVA
NEUC, CS, ST, MV, BIO, EMP, FO	Structural 3D protein folding molecule coordinate matrices, Human single-cell RNA sequencing cellular maps	t-SNE, UMAP Dimensionality Reduction, Manifold Learning
DIR, CS, ST, UV, GS, EMP, FO	Ocean meteorology wind direction degree logs, Wildlife migration direction angle coordinates	Directional Statistics, Von Mises Distribution Curve Fitting
FUN, TS, ST, UV, FIN, EMP, FO	Intraday stock trading price continuity curves, Continuously monitored energy utility grid loads	Functional Data Analysis (FDA), Functional PCA
TOP, CS, ST, MV, BIO, EMP, FO	Medical brain arterial tree capillary network meshes, Complex biological genetic sequence configurations	Topological Data Analysis (TDA)
QD, LP, ST, MV, BEH, EMP, FO	Annual patient doctor-visit frequency logs over a decade, Yearly customer subscription renewal counts	Generalized Estimating Equations (GEE) with Poisson Log-Link
QO, LP, ST, MV, BEH, EMP, MCAR	Employee annual performance tier tracking metrics over time, Multi-year customer product satisfaction review logs	Longitudinal Ordinal Logistic Regression, Markov Transition Models
FUZ, CS, ST, MV, BEH, EMP, FO	Subjective corporate risk appetite scoring assessments, Ambiguous client sentiment categorization maps	Fuzzy C-Means Clustering, Fuzzy Logic System Rule Engines
IV, CS, ST, MV, FIN, EMP, FO	Daily corporate stock price spread windows (Min-Max), Daily local real estate transaction valuation brackets	Interval-Valued Regression, Interval Arithmetic Operations
DIS, CS, ST, MV, FIN, EMP, FO	Operational asset performance risk variance arrays, Consumer portfolio income distribution shape sets	Wasserstein Distance Profiling, Information Geometry Metrics
TTE, LP, ST, MV, BIO, EMP, CEN	Longitudinal patient health study tracking heart failure occurrences, Long-term machine tool degradation logs	Joint Modeling of Longitudinal & Survival Data, Extended Cox Models
QNB, CS, ST, MV, BEH, CFAC, FO	User checkout interaction tracking with premium feature options, Digital advertising campaign targeted interaction sets	Causal Forest Analysis, Doubly Robust Estimation Formulas

QC, CS, ST, UV, EMP, TRU	Insurance accident claims logged strictly above deductible floors, Employee recruitment screening test score distributions	Truncated Linear Regression, Maximum Likelihood Estimation
QD, CS, ST, MV, GS, PP, FO	Urban emergency service dispatch event GPS hotspots, Commercial forestry tree species inventory locations	Spatial Point Pattern Processes, Nearest Neighbor Distances
UST, STR, MV, BEH, EMP, FO	Live financial news wire text feeds streaming, Live online twitch video stream text chat lines	Online Word Embeddings, Continuous Sentiment Stream Transformers
QC, CS, ST, UHD, FIN, EMP, FO	Algorithmic trading market signal predictors (p >> n), Ultra-high dimensional financial risk factor matrices	Sure Independence Screening (SIS), Penalized Ridge Models
QCC, CS, ST, MV, EMP, FO	Household category purchase unit breakdowns (in percentages), Corporate operational budget line distributions	Compositional Data Analysis (CoDa), Log-Ratio Structural Models
QNB, TS, ST, MV, FIN, EMP, FO	Daily corporate credit rating change alerts (Downgrade/Do not downgrade), Central bank interest rate shift signals	Autoregressive Binary Probit Models, Markov Switching Regression
QC, TS, ST, MV, BIO, EMP, FO	Continuous ICU patient blood pressure monitor feeds, Laboratory wearable heart rate monitor recordings	Dynamic Linear Models, Fast Fourier Transform (FFT) Filtering
QNB, LP, ST, MV, BEH, EMP, FO	Annual user streaming subscription statuses over few years, Annual corporate credit default registry statuses	Panel Logistic Regression, Generalized Linear Mixed Models (GLMM)
UST, CS, MV, BIO, EMP, FO	Clinical pathology diagnostic text report archives, Hospital electronic health record physician text summaries	Clinical Named Entity Recognition (NER), BioBERT Embeddings
QC, EXP, ST, MV, BIO, FO	Laboratory drug dosage intervention response levels, Agricultural chemical crop yield trial metrics	Factorial ANOVA, Post-Hoc Tukey HSD Tests, Manova Models
SST, TS, MV, FIN, EMP, FO	Multi-currency crypto transaction ledger files (XBRL/JSON), Banking core application system activity logs	NoSQL Time-Series Transformations, Extreme Gradient Boosting
QC, CS, ST, MV, FIN, EMP, TRU	High-income tax bracket registry audits, Premium credit card user account spend behaviors	Truncated Normal Regression, Heckman Two-Step Adjustment
QD, CS, ST, BV, FIN, EMP, FO	Corporate board of directors seat sizes vs revenue, Corporate venture capital patent count distributions	Negative Binomial Correlation, Generalized Linear Models (GLM)

QN, CS, ST, MV, GS, EMP, FO	Municipal parcel land-use zoning category codes, National forest vegetation class classification maps	Multinomial Logistic Spatial Models, Spatial Random Forest
TTE, CS, ST, MV, FIN, EMP, CEN	Corporate venture capital portfolio startup bankruptcy horizons, Commercial housing mortgage prepayment loan cycles	Parametric Accelerated Failure Time (AFT) Models, Cox Models
QC, CS, ST, MV, LAT, CFAC, FO	Educational student test tier outcomes using latent abilities, Corporate workspace psychological training program outputs	Structural Equation Causal Models, Latent Class Matching
QNB, CS, ST, MV, BIO, LAT, FO	Clinical psychological patient depression indicators, Medical diagnostic syndrome checklist responses	Item Response Theory (IRT), Latent Class Analysis (LCA)
QC, TS, ST, MV, GS, EMP, FO	Regional agricultural weather station drought timelines, Ocean surface tracking buoy temperature sequences	Spatio-Temporal Kriging, Vector Autoregression with Spatial Weights
QD, STR, ST, MV, BEH, EMP, FO	Server connection error counts tracking hourly, Web application user check-out transaction items	Streaming Poisson Models, Online Page-Hinkley Change Isolation
UST, CS, MV, FIN, EMP, FO	Corporate annual financial earnings call transcripts, Investment bank text analyst summaries	Financial Sentiment Dictionary Analysis, FinBERT Transformers
QC, CS, ST, HDLSS, FIN, EXP, FO	Proprietary institutional trading signal simulations (high variables), Investment strategy factor optimization matrices	Elastic Net Regularization, Cross-Validated Ridge Models
QO, CS, ST, MV, BEH, EMP, MCAR	Consumer target market product preference rank choices, Employee engagement workspace health surveys with gaps	Ordinal Logistic Regression, Random Forest Imputation
QC, LP, ST, MV, BEH, EMP, FO	Annual customer credit card spend metrics tracked over few years, Annual corporate infrastructure emissions logs	Linear Mixed-Effects Models, Dynamic Panel Data GMM (Arellano-Bond)
QCC, TS, ST, MV, FIN, EMP, FO	National sovereign asset reserve allocation shifts over time, Institutional investment fund asset percentage shifts	Time-Series Compositional Data Analysis, Vector Log-Ratios
QC, CS, ST, MV, GS, PP, FO	Urban commercial property pricing paired with coordinates, Precision agriculture field soil nitrogen maps	Marked Point Pattern Analysis, Spatial Covariate Regression

GN, TS, ST, MV, BEH, EMP, FO	Daily transactional bank account wire transfer maps, E-commerce platform user-product interactive click graphs	Spatio-Temporal Graph Neural Networks, Dynamic Network Analysis
SST, STR, MV, BEH, EMP, FO	Website user clickstream event payloads (in JSON), Multi-user mobile gaming action telemetries	Online Random Forests
QC, CS, ST, MV, SYN, FO	Generative model synthesized consumer banking records, Simulated medical health outcome patient profiles	Wasserstein Distance, Empirical Distribution Kolmogorov-Smirnov Tests
TTE, CS, ST, MV, BEH, EMP, CEN	Digital streaming application customer subscription life lengths, Mobile gaming application user churn time metrics	Survival Analysis, Parametric Weibull Regression, Cox Proportional Models
QNB, CS, ST, MV, GS, EMP, FO	Municipal properties inside flooding hazard zones (Yes/No), Regional bank branch location profitability indicators	Spatial Logistic Regression, Autoregressive Spatial Probit
QC, CS, ST, BV, FIN, EMP, MCAR	Corporate capital expenditures vs net profits with missing items, Quarterly startup valuation logs with missing data lines	Simple Linear Regression, Expectation-Maximization (EM) Imputation
QD, CS, ST, MV, BIO, EMP, FO	Hospital patient historical pregnancy count metrics, Clinical patient bone fracture count records	Poisson Log-Linear Modeling, Standard Count Regression Classifiers
UST, LP, MV, BEH, EMP, FO	Multi-year corporate employee open-ended performance text logs, Patient multi-year psychiatric progress text files	Longitudinal Text Mining, Dynamic Word Embedding Evolution
QC, PCS, ST, MV, FIN, EMP, FO	Cross-regional industry labor financial data aggregates, Multi-state corporate tax structure compliance tables	Pooled OLS Regression, Fixed-Effect Difference-in-Differences
QNB, CS, ST, MV, BIO, EXP, FO	Clinical vaccine trial infection occurrences (Infected/Not infected), Medical treatment trial adverse event indicators	Logistic Regression, Odds Ratio Profiling, Wald Inference Tests
QC, TS, ST, MV, FIN, SYN, FO	Generative AI synthesized stock price asset paths, Monte Carlo financial scenario valuation iterations	Quantile-Quantile Distribution Matching, Volatility Backtesting
DIR, TS, ST, UV, GS, EMP, FO	Coastal weather buoy wind direction updates tracking hourly, Tectonic plate migration drift tracking lines	Circular Time-Series Analysis, Wrapped Autoregressive Models
FUN, CS, ST, MV, BIO, EMP, FO	Continuous clinical patient ECG heart wave curves, Human biometric speech articulation wavelength graphs	Functional Analysis of Variance (FANOVA), Functional PCA

TEN, CS, ST, MV, BIO, EMP, FO	Multi-channel EEG patient brain function grid arrays, Multi-spectral biological tissue biopsy scans	Tucker Decomposition, Higher-Order Principal Component Analysis
QO, CS, ST, MV, FIN, EMP, FO	Sovereign government credit rating classification bands, Corporate supply chain priority tier tags	Ordinal Probit Regression, Cumulative Link Models
QC, CS, ST, BV, BIO, EMP, TRU	Patient adult height registries setting floor limits, Adult health study metrics omitting juvenile entries	Truncated Maximum Likelihood, Corrected Linear Regression
PP, TS, ST, MV, GS, EMP, FO	Regional wildfire outbreak incident locations over a decade, Urban crime occurrence GPS logs over 12 months	Spatio-Temporal Point Process Models, Hawkes Processes
GN, LP, ST, MV, FIN, EMP, FO	Multi-year international banking transaction network maps, Corporate venture capitalist investment network records	Dynamic Graph Neural Networks, Link Prediction Algorithms
SST, LP, MV, BEH, EMP, FO	Multi-year client app configuration adjustment files (in NoSQL), Annual user profile historical logs	Document Flattening pipelines, Panel Tree-Based Classifiers
QC, TS, ST, MV, BEH, EMP, MAR	Smart home electricity grid consumption logs with systematic drops, Industrial factory machinery sensor feeds with routine down periods	Kalman Filter State-Space Imputation, Inverse Probability Model
QNB, CS, ST, MV, BEH, LAT, FO	Digital customer purchasing intention indicators, User web interface engagement state matrices	Latent Class Analysis (LCA), Choice Architecture Frameworks
QC, CS, ST, MV, FIN, CFAC, FO	Small business development grant economic payout reviews, Corporate carbon tax compliance financial effects	Regression Discontinuity Design (RDD), Synthetic Control Methods
UST, CS, MV, BEH, SYN, FO	Generative AI synthesized product review texts, LLM generated customer service dialogue samples	Perplexity Score Diagnostics, Turing Test Classification Models
TTE, TS, ST, MV, BIO, EMP, CEN	Clinical patient recurrent stroke event tracking timelines, Multi-stage industrial engine system component breakdowns	Recurrent Event Survival Analysis, Andersen-Gill Models
QC, CS, ST, MV, FIN, EMP, MNAR	High-earner compensation surveys where top tiers opt-out, Corporate debt collection performance audit logs	Heckman Selection Modeling, Non-Ignorable Missingness Imputation
QD, CS, ST, MV, FIN, EMP, FO	Corporate annual patent registration volume counts, Commercial banking branch location ATM installations	Poisson Regression, Hurdle Count Models, Zero-Inflated Poisson

QO, TS, ST, MV, FIN, EMP, FO	Weekly corporate credit ranking risk tiers over time, Monthly country macro financial stability levels	Ordered Autoregressive Probit, Markov Chain Transition Models
QC, LP, ST, MV, GS, EMP, FO	Multi-year regional climate station tracking logs, Annual satellite farm crop vegetation density scores	Linear Mixed Spatial Models, Spatio-Temporal Panel Regression
SST, CS, MV, FIN, EMP, FO	Institutional investment portfolio asset structures (in JSON/XML), Banking customer account profile metadata	Semi-Structured Parsing, Gradient Boosting Decision Trees (GBDT)
QC, CS, ST, BV, BEH, EMP, FO	Consumer income level distributions vs retail store purchases, Total web platform session durations vs order sizes	Pearson/Spearman Correlations, Ordinary Least Squares Regression
QNB, TS, ST, UV, FIN, EMP, FO	Daily market recession status markers (Recession/No), Stock index daily positive or negative direction closes	Autoregressive Logistic Forecast Models, Markov Switching Probit
QC, CS, ST, UV, FIN, EMP, FO	Individual corporate asset return baseline variances, Retail store individual checkout transaction totals	Basic Descriptive Analysis, Skewness/Kurtosis Tests, Z-Scores
UST, EXP, ST, MV, BEH, FO	Marketing headline design variation text performance outcomes, User interface generative layout feedback collections	Text-Based A/B Testing, Vector Embedding Distance Metrics
TTE, CS, ST, BV, FIN, EMP, CEN	Venture capital portfolio company duration tracking to IPO, Commercial building loan default timelines	Single-Variable Cox Regression, Life Table Survival Diagnostics
QC, LP, ST, MV, BIO, EXP, FO	Clinical drug trials tracking patient biological responses over weeks, Medical therapy recovery progress matrices over few months	Repeated Measures ANOVA, Linear Mixed-Effects Growth Models
QC, CS, ST, MV, BIO, EMP, FO	Tumor biopsy feature measurements for malignant/benign classification	Support Vector Machines (SVM), Linear Discriminant Analysis (LDA)
QN, CS, ST, MV, BEH, EMP, FO	Customer segment labels predicted from browsing/purchase features	k-Nearest Neighbors (kNN), Quadratic Discriminant Analysis (QDA)
UST, CS, MV, BEH, EMP, FO	Email spam/ham text classification, SMS phishing message flagging	Naive Bayes Classifier, TF-IDF Feature Vectors
QC, CS, ST, MV, FIN, EMP, FO	Customer segmentation by spend and frequency for a retail loyalty program	Hierarchical Agglomerative Clustering, Dendrogram Analysis
QC, CS, ST, MV, GS, EMP, FO	Geolocated crime incident clusters, irregularly shaped urban hotspots	DBSCAN, HDBSCAN Density-Based Clustering
QO, CS, ST, BV, BEH, EMP, FO	Two independent groups' Likert-scale satisfaction ratings (non-normal)	Mann-Whitney U Test

QC, LP, ST, BV, BIO, EXP, FO	Paired pre/post treatment measurements on the same patients	Wilcoxon Signed-Rank Test
QO, CS, ST, MV, BEH, EMP, FO	Satisfaction rank comparisons across three or more store branches	Kruskal-Wallis H Test
QC, CS, ST, MV, GS, EMP, FO	Non-linear relationship between pollution exposure and health outcome across regions	Generalized Additive Models (GAM)
QC, TS, ST, UV, FIN, EMP, FO	Smoothing a noisy daily commodity price series for trend extraction	Spline Regression, Kernel-Weighted Local Regression (LOESS)
QC, CS, ST, MV, HDLSS/UHD, BIO, EXP, FO	Multiple gene-expression hypothesis tests run simultaneously across thousands of genes	Bonferroni Correction, Benjamini-Hochberg FDR Adjustment
QC, CS, ST, MV, FIN, EMP, FO	Estimating uncertain asset-return parameters using prior market beliefs	Bayesian Linear Regression, Markov Chain Monte Carlo (MCMC)
QC, LP, ST, MV, BEH, EMP, FO	Multi-level student test scores nested within schools and districts	Hierarchical (Multilevel) Bayesian Models
QC, TS, ST, MV, FIN, EMP, FO	Long-run co-moving relationship between interest rates and exchange rates	Vector Error Correction Model (VECM), Johansen Cointegration Test
QC, CS, ST, UV, BIO, EMP, FO	Estimating confidence intervals for a small clinical sample's mean effect	Bootstrap Resampling, Permutation Testing
QC, CS, ST, MV, FIN, EMP, FO	Comparing predictive accuracy of competing credit-risk models on holdout folds	K-Fold Cross-Validation, Leave-One-Out Cross-Validation
QN, CS, ST, MV, BEH, EMP, FO	Retail shopping-basket item co-purchase transaction logs	Apriori Algorithm, Market Basket / Association Rule Mining
QC, STR, ST, MV, BEH, EMP, FO	Sequential ad-recommendation decisions optimizing long-run click-through reward	Reinforcement Learning, Multi-Armed Bandit Algorithms
QC, CS, ST, MV, BIO, EMP, FO	Pooled effect sizes across dozens of published clinical drug-trial studies	Meta-Analysis, Forest Plot Effect-Size Pooling
QIC, CS, ST, UV, GS, EMP, FO	Daily ambient temperature readings in Celsius across weather stations	Linear Regression, Mean/Standard Deviation Analysis
QRC, CS, ST, UV, BIO, EMP, FO	Patient height and body weight clinical measurements	Geometric Mean, Coefficient of Variation
QUA, CS, ST, MV, EMP, FO	Qubit-state measurement outcomes from quantum computing experiments	Quantum State Tomography, Quantum Machine Learning

QDC, CS, ST, MV, BIO, EMP, FO	Genetic allele counts per locus summing to a fixed total per sample	Dirichlet-Multinomial Regression, Compositional Data Analysis
SEQ, CS, UST, MV, BIO, EMP, FO	DNA/protein sequence databases, Genome assembly read alignments	Sequence Alignment (BLAST), Hidden Markov Models
ADM, LP, ST, MV, BEH, EMP, FO	National tax registry records, Government benefit enrollment logs	Record Linkage, Small Area Estimation
STO, TS, ST, MV, FIN, EMP, FO	Stock price random-walk simulations, Queueing-system arrival processes	Monte Carlo Simulation, Stochastic Process Modeling
RST, CS, ST, MV, BEH, EMP, FO	Customer risk categorization with ambiguous, overlapping boundary cases	Rough Set Theory, Rule Induction
POSS, CS, ST, MV, BEH, EMP, FO	Multi-sensor fault diagnosis with conflicting sensor readings, Multi-expert risk assessment with disagreeing evidence sources	Dempster-Shafer Combination Rule, Possibilistic Reasoning
QID, CS, ST, UV, EMP, FO	Calendar year of construction across a building registry, Historical event years in a chronological archive	Linear Regression, Mean/Standard Deviation Analysis
QID, CS, ST, MV, BEH, EMP, FO	Standardized test scores (SAT/GRE, integer-scored) across a student cohort, Whole-number IQ test results across a clinical sample	Linear Regression, Factor Analysis
QRD, CS, ST, UV, GS, EMP, FO	Floor number in a multi-story building dataset (only positive numbers), Number of stories per building in an urban housing registry	Geometric Mean, Coefficient of Variation
QRD, CS, ST, BV, BEH, EMP, FO	Number of loyalty-program tier levels reached per customer, Employee job-grade level within a fixed hierarchy	Geometric Mean, Coefficient of Variation, Linear Regression
QD, TS, ST, MV, EPI, EMP, FO	Daily reported infection counts during a disease outbreak, Weekly hospitalization counts across regions	SIR/SEIR Compartmental Modeling, R_0 Estimation
QC, TS, ST, MV, IND, EMP, FO	Manufacturing line dimensional measurements sampled hourly, Assembly-line defect rate tracking	Statistical Process Control, Control Charts (X-bar/R), Cpk/Ppk Analysis
QO, CS, ST, MV, PSY, EMP, FO	Multi-item personality inventory responses, Standardized psychological assessment scale scores	Cronbach's Alpha, Confirmatory Factor Analysis
QNB, CS, ST,	National voter-intention survey with	Multilevel Regression with

MV, POL, EMP, FO	demographic breakdowns, State-level referendum polling data	Poststratification (MRP)
QNB, CS, ST, BV, SPT, EMP, FO	Head-to-head tennis match outcomes across a season, Chess tournament game results	Bradley-Terry Model, Elo Rating System

Cases when applied methods depend on amount of data

DATA TYPE	SMALL SAMPLE METHODS	LARGE SAMPLE METHODS
QNB	Fisher's Exact Test	Chi-Square Test, Cramér's V / Odds Ratio (effect size focus)
QN	Fisher's Exact Test (extensions), Exact multinomial tests	Chi-Square Test, Cramér's V
QC	Mann-Whitney/Wilcoxon (nonparametric), Exact permutation tests	T-Test/ANOVA (CLT-robust), Cohen's d / confidence intervals (effect size focus)
TS	Exponential Smoothing, naive/simple forecasts	ARIMA, GARCH, LSTM Neural Networks
Any, for estimates	Bayesian regression/MCMC	Frequentist methods converge with Bayesian as data grows; prior influence diminishes

General concepts underlying many listed methods

CONCEPT	AFFECTED DATA TYPES AND METHODS
n ≈ 30 rule of thumb (applies to mildly non-normal data only)	QC, QID, QIC, QRD, QRC; many CS/TS/LP combinations using T-Test, ANOVA, OLS
Normality assumption	QC, QID, QIC, QRD, QRC; T-Test, ANOVA, Linear Regression, Factor Analysis
Homogeneity of variance (homoscedasticity)	QC, QID, QIC, QRD, QRC; ANOVA, T-Test, Linear/Multiple Regression
Independence of observations	Almost every CS-based method
Linearity assumption	QC, QID, QIC, QRD, QRC, BV, MV rows using Linear/Multiple Regression, Pearson Correlation
No/low multicollinearity	MV, HDLSS rows
$p < 0.05$ significance threshold	Chi-Square Test, T-Test, ANOVA, Fisher's Exact Test, Mann-Whitney Test, Wilcoxon Test, Kruskal-Wallis Test, Friedman’s Test
Stationarity assumption	TS, FIN combinations
n > p requirement (samples > variables)	MV

MCAR/MAR assumption for imputation	MCAR, MAR

Trees of relationships between methods

Regression family
- Continuous outcome
  - Linear Regression (fit via OLS)
    - Simple Linear Regression
    - Multiple Linear Regression
  - GAM
  - Spline/LOESS
    - Spline Regression
    - LOESS (Local Regression)
  - Regularized variants
    - LASSO
    - Ridge
    - Elastic Net
- Categorical outcome
  - Binary Logistic Regression
  - Multinomial Logistic Regression
  - Ordinal Logistic / Ordinal Probit
    - Ordinal Logistic Regression
    - Ordinal Probit Regression
  - Regularized variants
    - Regularized Logistic Regression
- Count outcome
  - Poisson Regression
    - Standard Poisson
    - Zero-Inflated Poisson
    - Hurdle Poisson
  - Negative Binomial Regression

Hypothesis testing (classical)
- Parametric
  - T-Test
  - ANOVA
  - MANOVA
  - Factorial / Multi-way ANOVA
  - Two-Sample Z-Test
- Categorical/count-based
  - Chi-Square Test
  - Fisher's Exact Test
  - Odds Ratio

Descriptive/EDA
- Central tendency & dispersion
  - Mean/Standard Deviation
  - Median
  - Mode
  - Descriptive Statistics
- Frequency distributions
  - Frequency Tables/Counts
- Shape diagnostics
  - Skewness/Kurtosis Tests
- Density estimation
  - Kernel Density Estimation
- Visualization
  - Histograms
- Outlier detection
  - Outlier Z-Score Detection

Classification & ML
- Classical classifiers
  - SVM
  - kNN
  - Naive Bayes
  - LDA/QDA
    - Linear Discriminant Analysis
    - Quadratic Discriminant Analysis
- Tree ensembles
  - Random Forest
  - XGBoost
- Neural
  - CNN
  - Vision Transformer (ViT)
  - LSTM
  - GNN
  - LLM
    - Encoder-only
    - Decoder-only
    - Encoder-decoder
- Text representation
  - Word Embeddings
  - TF-IDF

- Clustering & dimensionality reduction
  - Clustering
    - K-Means
    - DBSCAN
      - HDBSCAN
    - Hierarchical Clustering
  - Dimensionality reduction
    - PCA
    - Regularized PCA
    - Factor Analysis
    - Manifold-based
      - t-SNE
      - UMAP
      - Isomap
- Panel/longitudinal regression
  - Effects models
    - Fixed Effects
    - Random Effects
  - Generalized estimation
    - Generalized Estimating Equations (GEE)
  - Pooled cross-sectional
    - Pooled OLS Regression
- Time-series
  - Univariate forecasting
    - ARIMA
      - Standard ARIMA
      - SARIMAX
    - Exponential Smoothing
    - GARCH
  - Multivariate/structural
    - VECM/Cointegration
    - Kalman Filter State-Space
  - Dynamical systems modeling
    - Compartmental Models
      - SIR Model
      - SEIR Model
    - R_0 Estimation
- Distributional methods
  - Wasserstein Distance Metrics
  - Information Geometry
- High-dimensional screening
  - Sure Independence Screening (SIS)
- Survival / time-to-event
  - Nonparametric
    - Kaplan-Meier
  - Regression-based
    - Cox Proportional Hazards
      - Standard Cox Model
      - Extended Cox Model
    - Accelerated Failure Time
    - Tobit Regression
      - Standard Tobit
      - Tobit Time-Series Regression
    - Truncated Regression Models
- Compositional data
  - Transformation-based
    - Aitchison Geometry
    - Log-ratio Transformations
      - CLR Transformation
      - ILR Transformation
  - Regression-based
    - Dirichlet Regression
    - Dirichlet-Multinomial Regression
- Causal inference
  - Matching/weighting
    - Propensity Score Matching
    - Doubly Robust Estimation
  - Design-based
    - Regression Discontinuity Design
    - Synthetic Control Methods
    - Difference-in-Differences (DiD)
  - ML-based
    - Causal Forest Analysis
- Ranking & comparison models
  - Bradley-Terry Model
  - Elo Rating System
- Topological data analysis
  - Topological Data Analysis (TDA)

- Uncertainty / non-probabilistic
  - Set-based
    - Fuzzy Logic
      - Fuzzy Logic Controllers
      - Fuzzy C-Means Clustering
    - Rough Set Theory
      - Rule Induction
      - Lower/Upper Approximation Analysis
  - Evidence-based
    - Dempster-Shafer Theory
      - Belief/Plausibility Function Combination
      - Possibility Theory
    - Interval Arithmetic

- Process monitoring & quality control
  - Control Charts
    - X-bar Chart
    - R Chart
  - Capability Indices
    - Cpk
    - Ppk
  - Six Sigma Methodology

- Psychometric measurement
  - Reliability
    - Cronbach's Alpha
    - Test-Retest Reliability
    - Inter-Rater Reliability
  - Item Response Theory

- Nonparametric / rank-based
  - Two-group comparison
    - Mann-Whitney U
    - Wilcoxon Signed-Rank
  - Multi-group comparison
    - Kruskal-Wallis
    - Friedman's Test
  - Association
    - Spearman Rank Correlation

- Financial risk modeling
  - Derivative pricing
    - Black-Scholes Model
  - Risk measurement
    - Value at Risk (VaR)

- Bayesian
  - Estimation
    - Bayesian Regression/MCMC
  - Multilevel
    - Hierarchical Bayesian Models
    - Multilevel Regression with Poststratification

- Missing-data methods
  - MCAR/MAR-appropriate
    - MICE
    - EM Algorithm
    - Inverse Probability Weighting
    - FIML
  - MNAR-appropriate
    - Heckman Selection Model
    - Pattern-Mixture Models

- Spatial
  - Interpolation
    - Kriging
      - Ordinary Kriging
      - Spatio-Temporal Kriging
  - Autocorrelation
    - Moran's I
  - Regression-based
    - Geographically Weighted Regression
  - Point patterns
    - Ripley's K-function
    - Spatial Poisson Point Process

- Meta-analysis / evidence synthesis
  - Pooling
    - Fixed-Effect Pooling
    - Random-Effects Pooling
  - Bias/heterogeneity checks
    - Forest Plots
    - Heterogeneity Statistics
      - I^2 Statistic
      - Cochran's Q Test

- Multiple-comparison correction
  - Bonferroni Correction
  - Benjamini-Hochberg FDR Adjustment

- Association rule mining
  - Apriori Algorithm

Sequence models
- Probabilistic
  - Hidden Markov Models
- Alignment/frequency
  - Sequence Alignment
    - Needleman-Wunsch
    - BLAST
  - N-gram Models

Behavioral analytics
- Retention/lifecycle analysis
  - Cohort Analysis
- Experimental comparison
  - A/B Testing
- State-transition modeling
  - Markov Chain Models

Resampling & validation
- Interval estimation
  - Bootstrap Resampling
  - Permutation Testing
- Predictive validation
  - K-Fold Cross-Validation
  - Leave-One-Out Cross-Validation

Reinforcement learning
- Reinforcement Learning
- Multi-Armed Bandit Algorithms

Latent variable modeling
- Structural Equation Modeling (SEM)
- Latent Class Analysis (LCA)

Directional statistics
- Von Mises Distribution Analysis

Functional data analysis
- Functional Data Analysis (FDA)
- Functional PCA

Quantum methods
- Quantum Machine Learning
- Quantum State Tomography

Stochastic simulation & processes
- Monte Carlo Simulation
- Stochastic Process Modeling
- Markov Chain Analysis

Administrative data methods
- Record Linkage
- Small Area Estimation
- Register-Based Cohort Analysis

Streaming & online learning
- Online Machine Learning
- Sliding Window Algorithms
- Hoeffding Trees

Generative model evaluation
- FID Score
- Augmentation Tests

Network analysis
- Graph Neural Networks (GNN)
- PageRank
- Centrality Metrics

Study design
- Observational Studies
- Randomized Controlled Trials (RCT)

Estimation frameworks
- Maximum Likelihood Estimation (MLE)
- Generalized Method of Moments (GMM)

Tensor methods
- Tensor Decomposition

General approaches to development of new methods

- Relaxing an assumption of an existing method (it is the most common approach).
- Extending a method to a new data structure it wasn’t originally created for.
- Combining several existing methods to solve a problem neither solves alone.

- Reframing a problem in a different mathematical language, then importing that language’s tools.
- Solving a computational bottleneck that was previously the limiting factor.
- Domain-driven necessity: a field hits a wall existing statistics can't cross, and builds a new tool.

Analysis

Standard sequence of data processing and methods application

1) Perform preliminary classification of data.
2) Check whether data are falsified or just have anomalies (optionally, see detection methods further in this guide).
3) Prepare data: a) resolve data structure issues and incompatibilities; b) resolve duplication issues and error values issues; c) check consistency of data types; d) perform domain-specific data preparation.
4) Perform working classification of data.
5) Perform exploratory data analysis by applying descriptive statistics to find central tendency, dispersion, frequency, distribution, and constructing histograms/plots based on extracted data.
6) Check structure of data to detect dependencies (temporal, spatial, categorical) and missingness (MCAR, MAR, MNAR). (Sometimes, it may be a part of the previous step, but I extracted it into separate step to highlight importance.)
7) Choose method for analysis.
8) Fit the model to the data.
9) Run diagnostics to check statistical assumptions on residuals to verify that residuals consist of “noise” only.
10) Interpret and report the results.

Solutions of diagnostic failures

DIAGNOSTIC FAILURE	SOLUTION
Non-normal residuals (any method producing residuals — Linear Regression, ANOVA, GAM, etc.)	Transform the variable (log, Box-Cox), or switch to the nonparametric alternative already listed under QO (Mann-Whitney/Wilcoxon/Kruskal-Wallis family)
Heteroscedasticity (Levene's test fails)	Weighted least squares, robust/heteroscedasticity-consistent standard errors, or variance-stabilizing transform

Autocorrelated residuals (Durbin-Watson, Ljung-Box)	Add lagged terms, switch to a model that accounts for the dependency structure (already-listed GEE, ARIMA, or Fixed/Random Effects depending on the data's temporal/panel structure)
High VIF (multicollinearity)	Drop or combine correlated predictors, or switch to the already-mentioned regularized methods (LASSO/Ridge/Elastic Net under HDLSS)
Overdispersion in Poisson (variance ⩾ mean)	Switch to Negative Binomial Regression − already mentioned as the QD alternative, just not linked to the diagnostic that triggers the switch
Proportional-hazards violation (Schoenfeld residuals)	Stratify the Cox model, or add a time-varying covariate/interaction
Non-stationary series (unit-root test fails) for any TS data	Differencing, or move to VECM/cointegration if multiple series are involved
Failed convergence (MCMC R-hat, EM algorithm)	Reparameterize, adjust priors, increase iterations/burn-in, or check for model non-identifiability

Detection of falsified data or data with anomalies

All listed detection methods just detect anomalies. They do not prove that conscious falsification really occurred.

FALSIFICATION PATTERN	DETECTION METHOD
Digit-frequency manipulation (fabricated numbers, altered financial/scientific figures)	Benford's Law (first-digit / first-two-digit distribution test); chi-square goodness-of-fit against expected digit frequencies
Implausibly uniform or "too clean" data (rounded values, suspiciously low variance)	Variance/dispersion check against domain-expected baseline; excess-uniformity test on the last digit or decimal distribution
Duplicated or copy-pasted records (fabricated by cloning real entries with minor edits)	Exact and near-duplicate detection (hashing, fuzzy matching / edit distance on records); duplicate-rate baseline comparison
Fabricated survey/subject responses (repeated identical response patterns, "straight-lining")	Response-pattern entropy analysis; identical-sequence detection across respondents; response-time-too-fast flags
Selectively reported results (cherry-picked favorable subsamples, p-hacked outcomes)	p-curve analysis; excess significance test (comparing observed proportion of significant results to the proportion expected given reported power)
Manipulated time-series (smoothed, backdated, or interpolated to hide gaps)	Autocorrelation-of-residuals check for artificial smoothness; changepoint detection at suspiciously convenient boundaries

Inconsistent internal totals (subtotals don't sum to reported totals, cross-table mismatches)	Cross-footing / reconciliation checks; internal consistency audits across related tables
Outlier suppression (implausible absence of natural outliers or tail events)	Comparison of observed tail behavior against the expected distribution (e.g., QQ-plot tail deviation, kurtosis check)
Coordinated/bot-generated data (unnatural regularity in timing or sequence)	Inter-arrival time distribution test (real human/process events are rarely perfectly regular); clustering of timestamp intervals
Image/data reuse across unrelated studies (duplicated figures, reused datasets misrepresented as new)	Perceptual hashing / image similarity matching; dataset fingerprinting and cross-study comparison

General approaches to perform meta-analysis

APPROACH	USE CASE
Fixed-effect meta-analysis	Assumes all studies estimate one true effect; pools via inverse-variance weighting
Random-effects meta-analysis	Assumes true effect varies across studies (heterogeneity); adds between-study variance to the pooling weights
Meta-regression	Explains why effect sizes vary across studies using study-level covariates
Forest plots	Standard visualization: each study's effect + CI, plus the pooled estimate
Heterogeneity statistics (I^2, Cochran's Q)	Quantify how much studies disagree beyond chance − decides fixed vs. random effect
Funnel plots / Egger's test	Detect publication bias (are small "null" studies missing?)
Network meta-analysis	Combines studies that compare different pairs of treatments, allowing indirect comparisons
Bayesian meta-analysis / hierarchical models	Treats each study as drawn from a shared prior

Automation of data analysis

The automation summary given further reflects the tooling landscape of the early 2020s. Since then, LLM-based agentic tools have been expanding aggressively into every stage of this pipeline. However, most available information on this topic online is vendor marketing rather than independent evaluation, and should be rechecked carefully. Treat the provided summary as a baseline, and verify current tooling through independent research rather than vendor claims alone.

Stages that can be automated only partially are highlighted with red dot symbol: ●.

STAGE	AUTOMATION APPROACH

● Classification of data (both preliminary and working)	Type inference on ingestion (pandas dtype inference, schema inference in Spark/Arrow); rule-based classifiers matching column properties (cardinality, value range, string patterns) to expected data types; however, new data types and combinations of data types may require human input to be classified.
Structure resolution and ingestion	Schema-on-read pipelines (Kafka with schema registry, Airflow/dbt transforms), automatic JSON/XML flattening.
● Falsification and anomaly screening	Benford's Law checks, duplicate-rate monitors, and response-pattern checks can all run as scheduled or streaming jobs; several are cheap enough to run on every new batch; however, automation detects anomalies only, it cannot decide whether anomaly is a result of conscious falsification.
Duplication and error-value detection	Deterministic and fuzzy dedup (Dedupe, Splink), constraint-based validation (Great Expectations, Deequ) running as a continuous data-quality gate.
● Missingness mechanism classification	MCAR-vs-MAR tests (Little's MCAR test) run automatically, but MAR-vs-MNAR is not statistically testable from the data alone, domain knowledge is necessary.
Dependency detection (temporal, spatial, categorical)	ACF/PACF computation, spatial autocorrelation (Moran's I), categorical association tests can all run unattended.
Exploratory data analysis	pandas-profiling/ydata-profiling, Sweetviz generate the full descriptive-stats/histogram suite with no human input.
● Method selection	Partially automatable via decision-rule engines or AutoML frameworks (auto-sklearn, H2O AutoML) that search a candidate set and pick by cross-validated performance.
Model fitting	Fully automated in any ML/stats pipeline.
● Diagnostics	Residual plots, VIF, R-hat, Ljung-Box can all be computed and thresholded automatically; MLOps monitoring platforms (Evidently, WhyLabs) run these continuously in production and alert on drift/violation. Interpretation of diagnostics results may require human input sometimes.
● Solution of issues detected at diagnostics	Can be automated as a rules engine ("if overdispersion detected then refit as Negative Binomial") but each such rule is itself a modeling decision that a human originally had to encode; full automation here is really "automating a human-authored decision tree," not the model discovering the right fix on its own.
● Interpretation and reporting	LLM can describe what is received in the result, but extracting real meaning may require human input and domain knowledge.