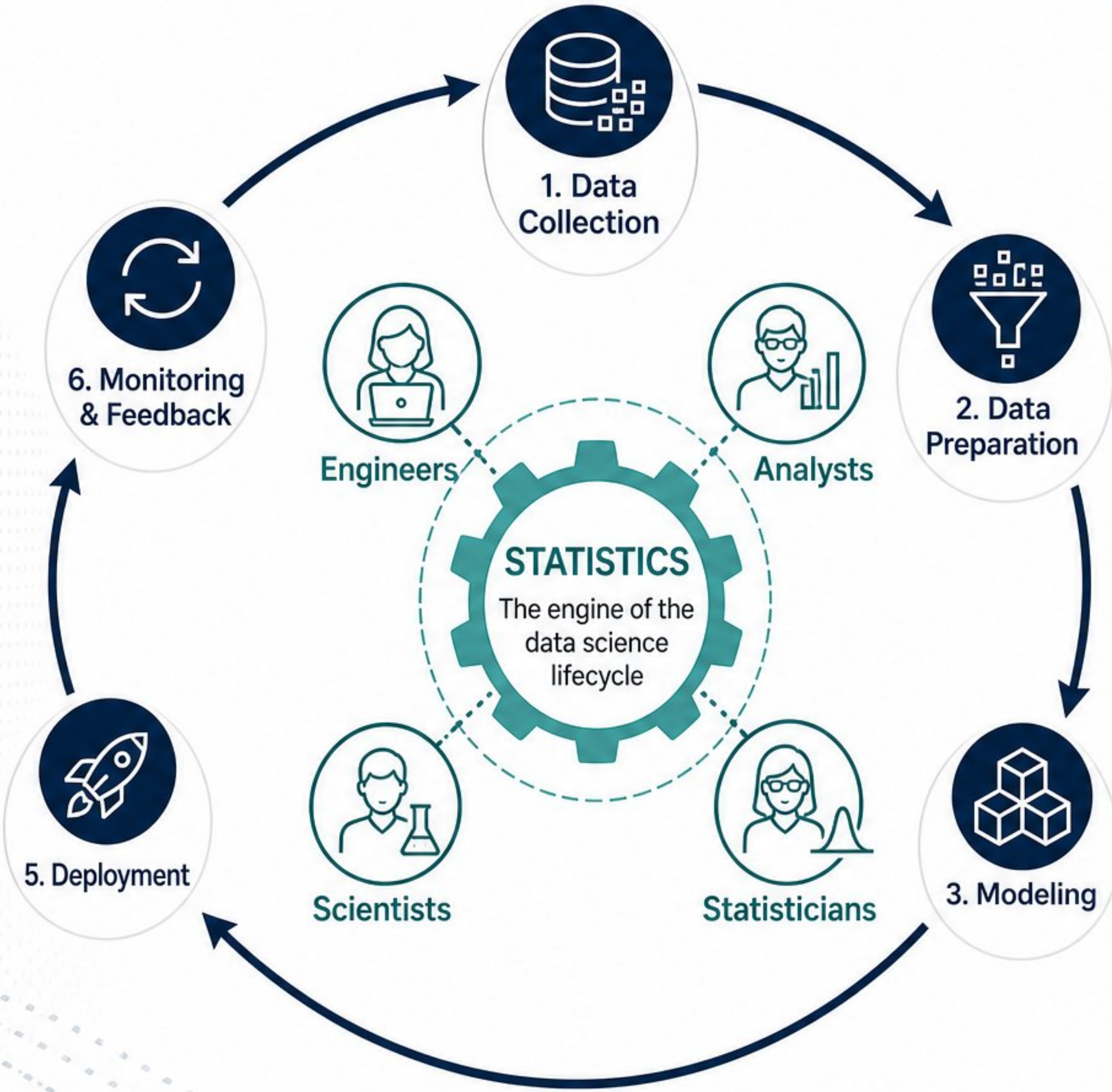
Fundamental Statistical Methods in Data Science

- Statistics powers the full data workflow: explore, infer, predict, validate, and communicate
- Descriptive stats summarize data; inferential stats draw conclusions from samples
- Predictive modeling learns patterns for forecasting future outcomes
- Statistical thinking guards against spurious correlations, bias, p-hacking, and overfitting

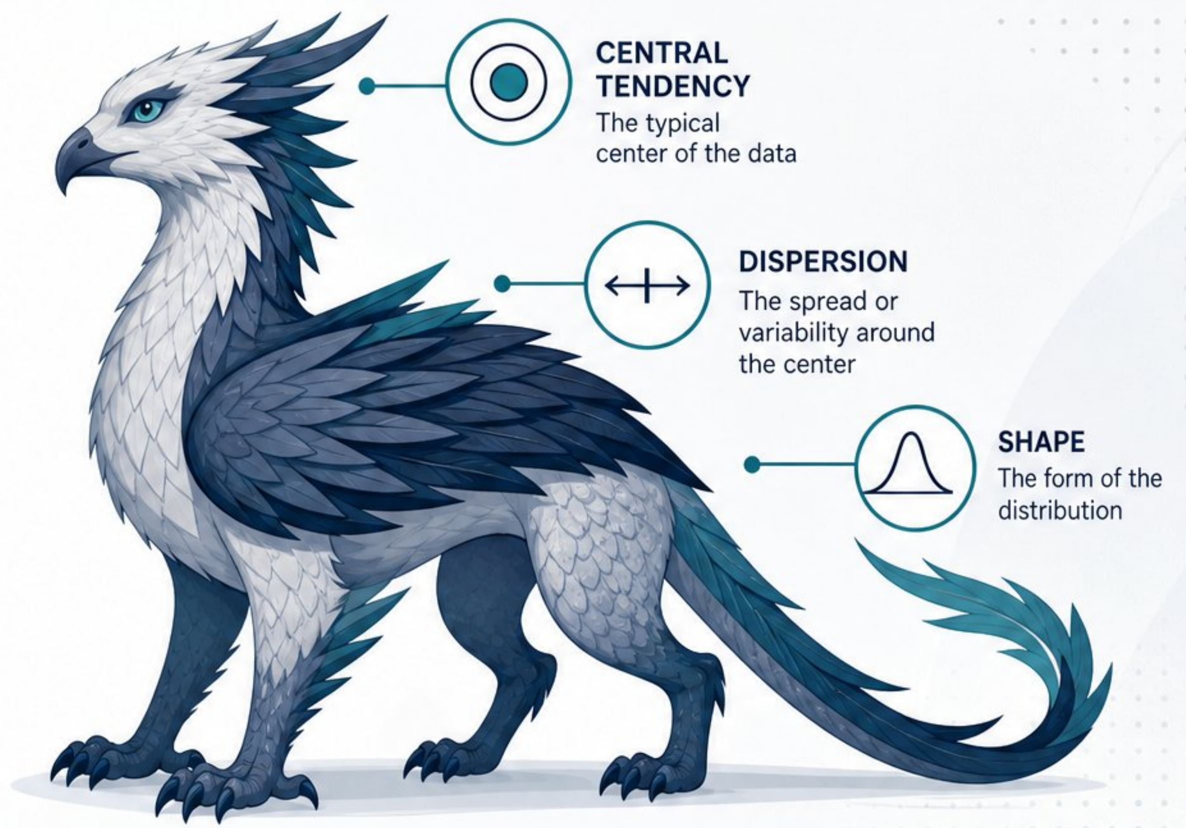
Statistics in the Data Science Workflow

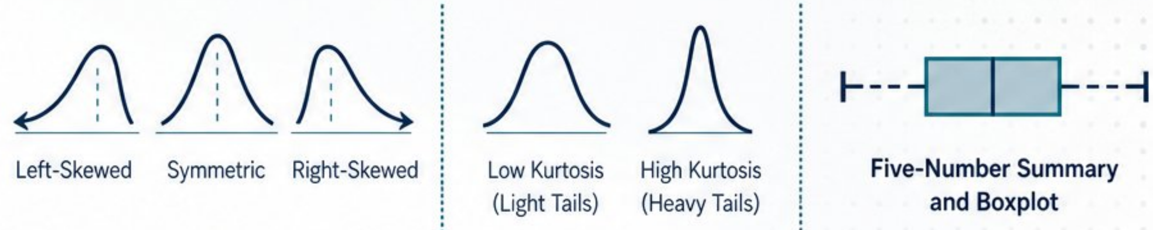


- Statistics: the engine of the full data science lifecycle
- Four complementary roles: engineers, analysts, scientists, and statisticians
- Shared foundations for classical, ML, and GenAI work in 2026
- Statistical literacy is the human safeguard against automated errors

Descriptive Statistics: Central Tendency, Dispersion, and Shape

- **Mean:** best for symmetric data without outliers; uses all values
- **Median:** robust choice for skewed data or outliers (income, prices)
- **Key dispersion:** range, IQR, variance, standard deviation
- **Skewness** shows tail direction; **kurtosis** indicates tail weight
- **Five-number summary** and **boxplots** give a robust distribution view

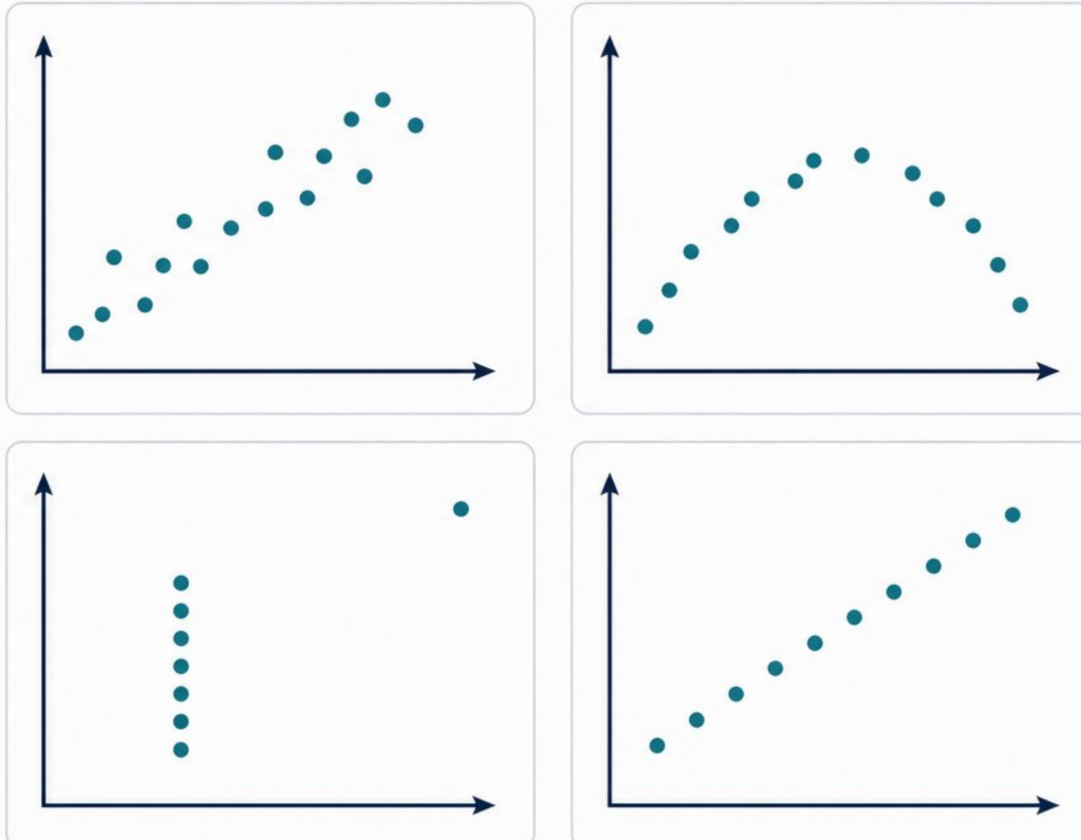




Visualizing Distributions and Relationships

- Choose chart types: histograms, density plots, scatter plots, and heatmaps for EDA
- Detect patterns, outliers, and gaps visually before running formal tests
- Honest visuals use proportional ink; avoid truncated axes and overplotting
- Pearson (r) measures linear strength; Spearman (r_s) captures monotonic rank-based trends
- Anscombe's quartet proves: always plot your data — summary statistics can mislead

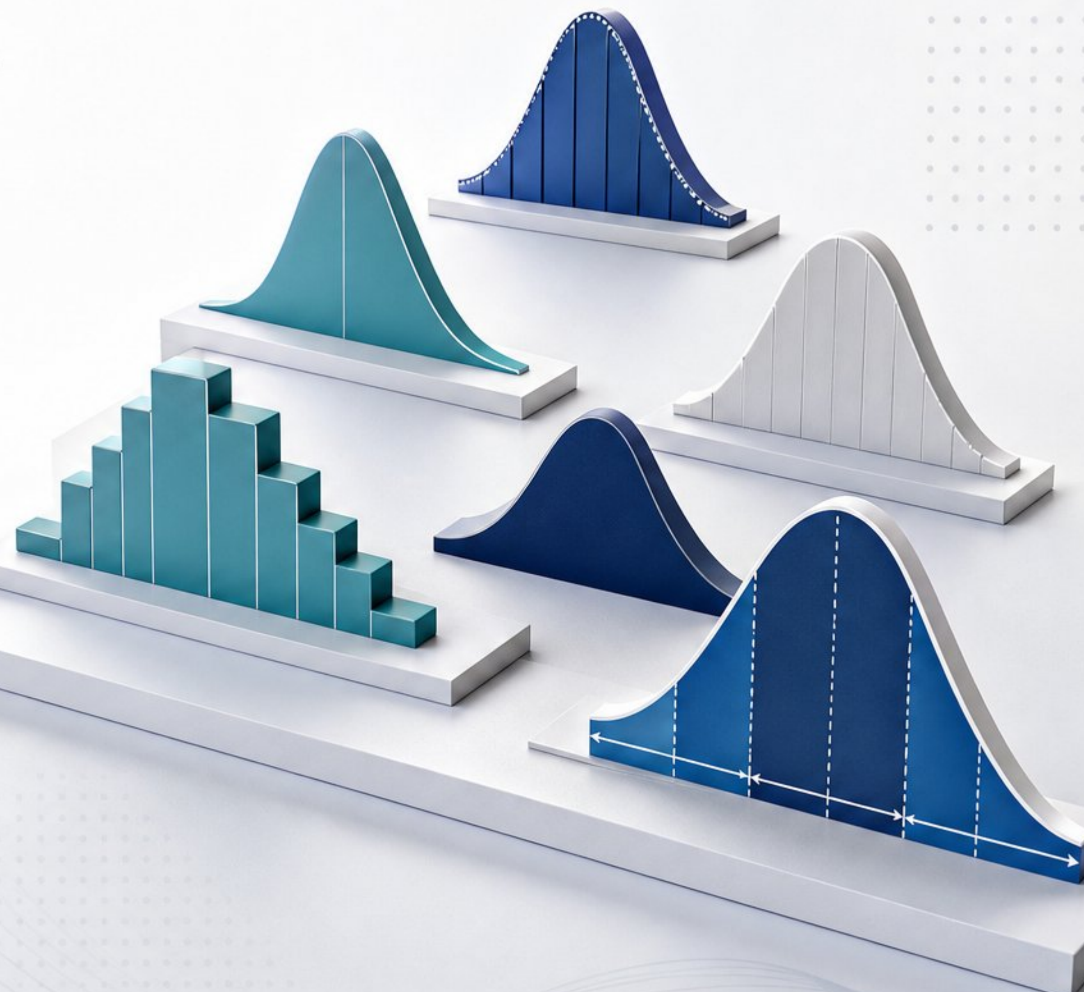
FOUR DATASETS. IDENTICAL SUMMARY.
DIFFERENT STORIES.



These four datasets share the same summary statistics (mean, variance, correlation) yet reveal completely different patterns. **Visualization reveals the truth.**

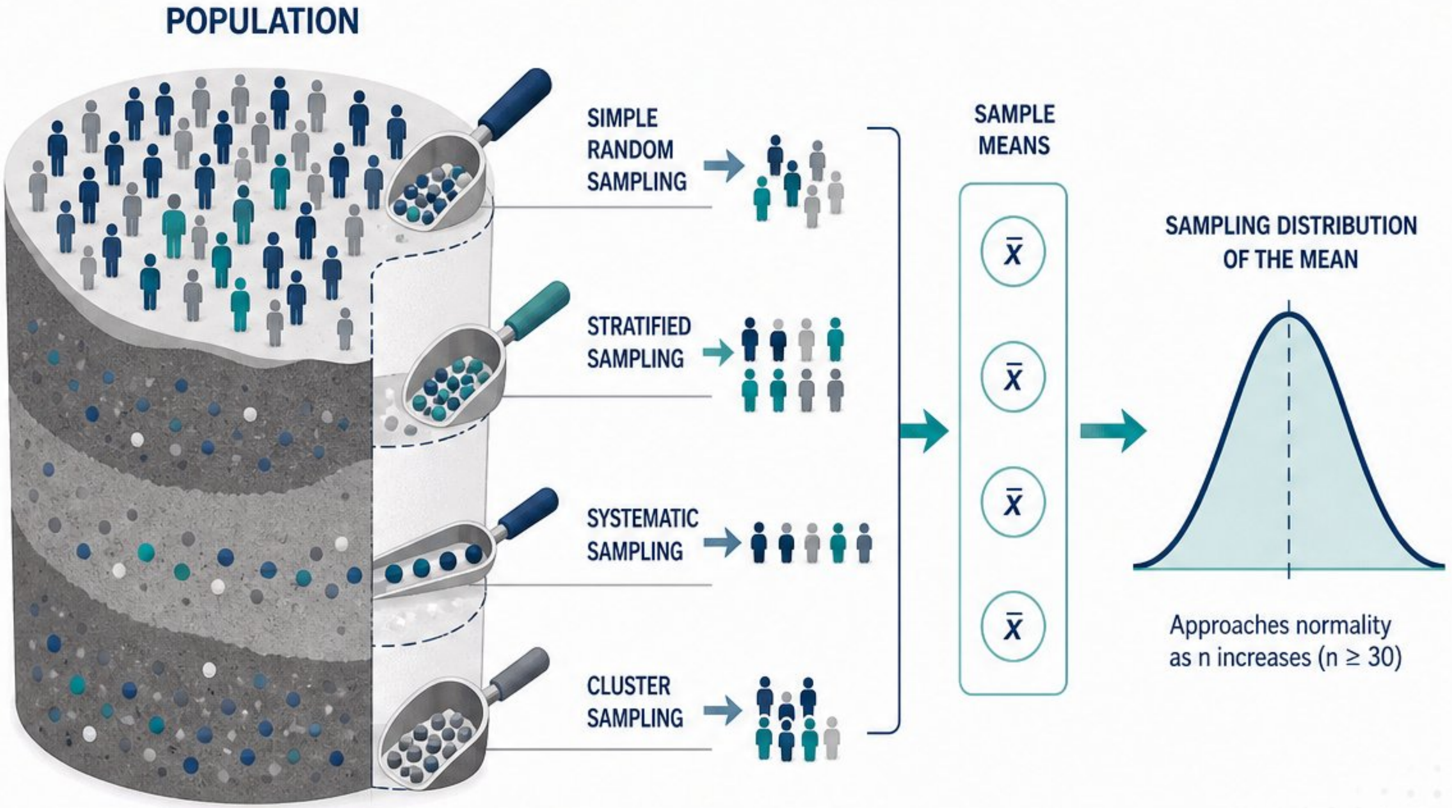
Probability Foundations for Data Analysis

- Independent events, conditional probability, and the law of large numbers
- Discrete vs. continuous distributions and their real-world mappings
- Binomial (success/failure), Poisson (event rates), Normal, and Exponential
- Normal distribution: the 68-95-99.7 rule and z-scores as a universal yardstick
- Probability bridges descriptive statistics to inferential methods



Sampling Methods and the Central Limit Theorem

- Sample for cost, time, and feasibility; aim for an unbiased, representative subset of the population
- Common methods: simple random, stratified, systematic, and cluster sampling
- Central Limit Theorem: Sampling distribution of the mean approaches normality as n increases ($n \geq 30$)
- CLT justifies normality for inference, even when raw data are non-normal
- Standard error of the mean decreases as sample size grows, improving estimate precision

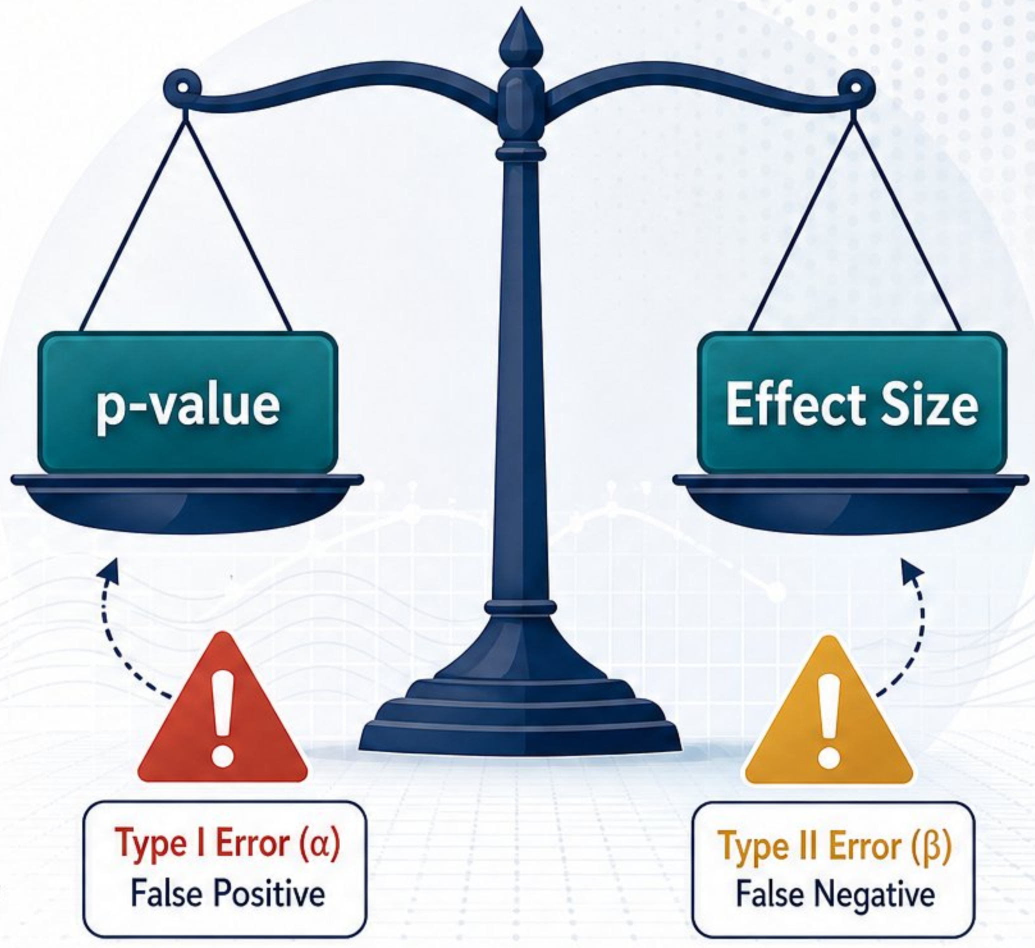


Confidence Intervals:
Estimation with Uncertainty

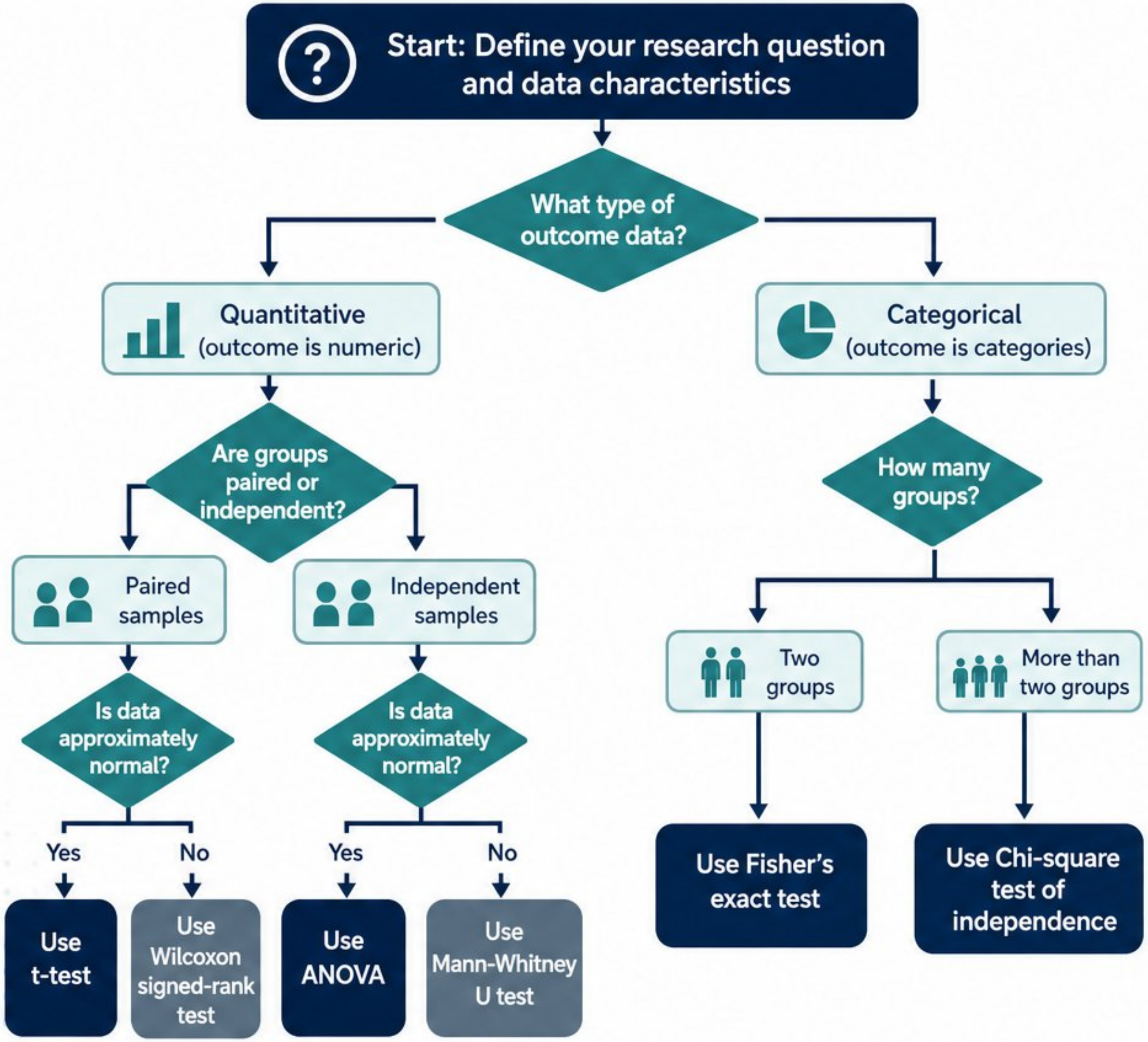
- Point estimates give a single value; interval estimates provide a plausible range
- Use z-distribution when σ is known, t-distribution when σ is unknown
- A 95% CI means 95% of repeated intervals would capture the true parameter
- Wider intervals result from smaller samples, higher variability, or higher confidence
- A specific computed interval does not have a 95% probability of containing the mean

Hypothesis Testing: Concepts, Errors, and p-Values

- Formulate null (H_0) and alternative (H_1) hypotheses in testable terms
- p-value: probability of data at least as extreme if H_0 is true
- Type I error (α): false positive; Type II (β): false negative
- Power ($1-\beta$) depends on sample size, effect size, and α
- Report effect sizes and CIs; avoid binary 'significant/not significant'



Choosing the Right Statistical Test



Always check assumptions and ensure the test aligns with your question and data.

- Use a decision framework: research question, outcome type, number of groups, paired vs. independent
- Parametric (t-test, ANOVA) vs. non-parametric (Mann-Whitney, Kruskal-Wallis, Friedman) alternatives
- Switch to non-parametric when data are ordinal, severely skewed, or samples are small and non-normal
- For categorical data, apply Chi-square test of independence or Fisher's exact test
- Define the test before seeing the data; post-hoc switching inflates false-positive rates

Correlation vs. Causation and Association Analysis

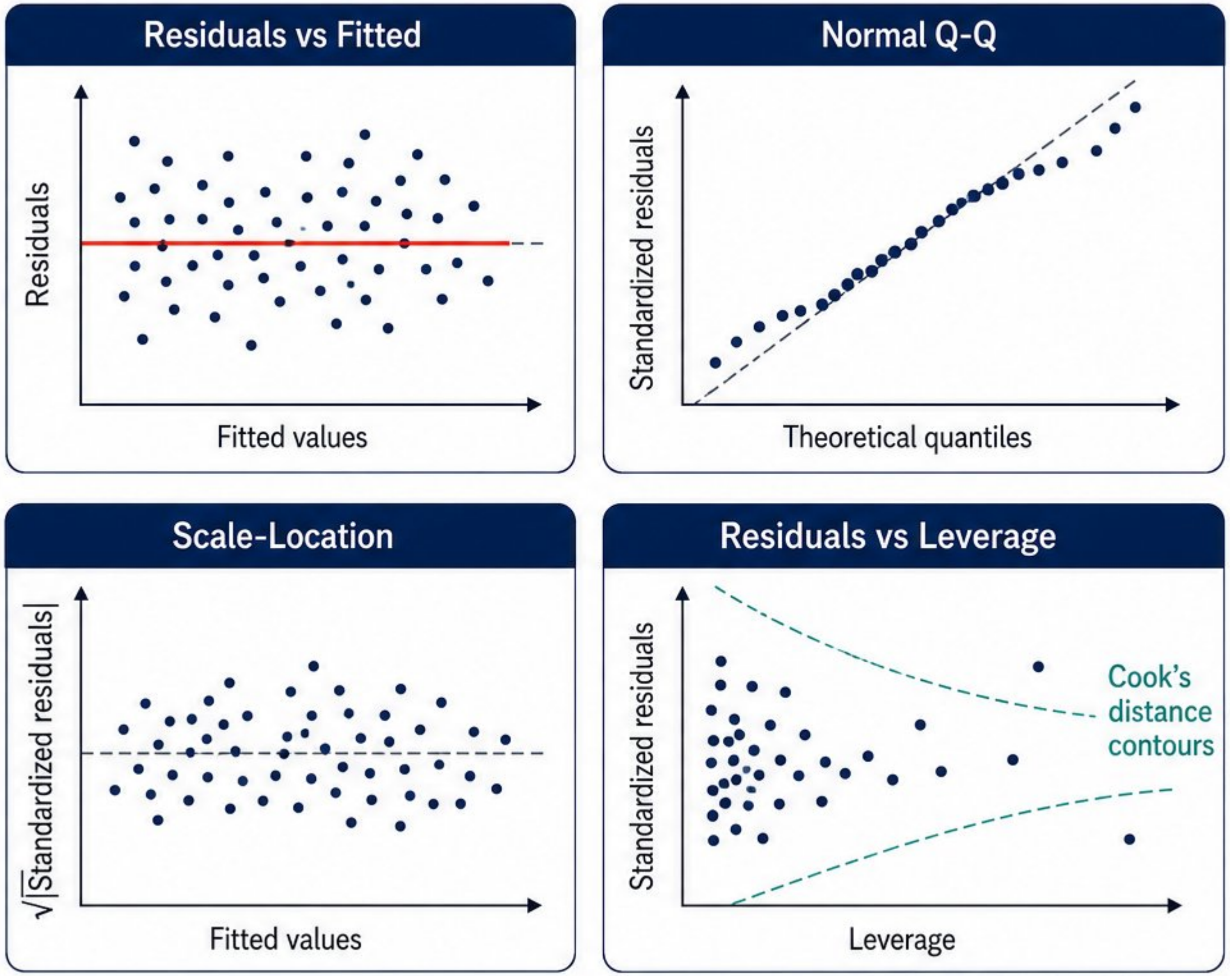
- Correlation measures strength/direction but never implies causation.
- Confounders, reverse causation, and spurious correlations are always possible.
- Pearson (r): linear relationships, sensitive to outliers, needs normality.
- Spearman (ρ): monotonic relationships, robust to outliers and skewed data.
- Correlation matrix is an exploration map, not a causal diagram.

Simple Linear Regression: Modeling Relationships

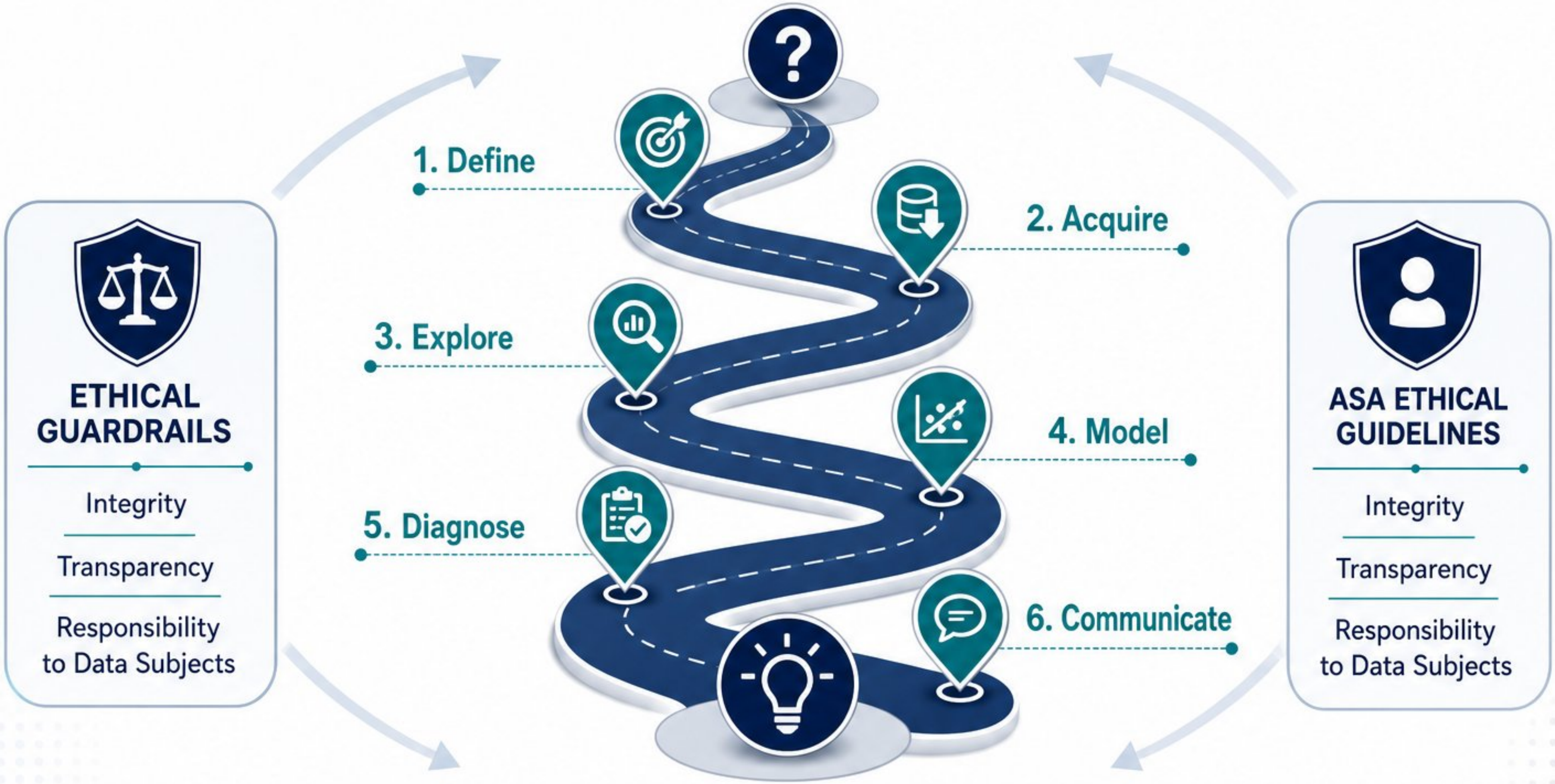
- Model form: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Slope β_1: mean change in Y per one-unit increase in X
- Intercept β_0: expected Y when X equals zero
- R-squared measures explained variance but does not confirm causation
- Four core LINE assumptions: Linearity, Independence, Normality, Equal variance

Regression Diagnostics: Checking Model Assumptions

- **Check four core assumptions:** linearity, independence, homoscedasticity, normality of residuals
- Residuals vs Fitted: flat red line = good; curves or funnels signal violations
- Normal Q-Q: points on diagonal line support normality; watch for heavy tails
- Scale-Location: horizontal band = constant variance; trend = heteroscedasticity
- Residuals vs Leverage: points beyond Cook's distance contours are influential



A Structured Statistical Analysis Framework



- Follow a 6-step reproducible workflow: define, acquire, explore, model, diagnose, communicate
- Uphold ASA ethical guidelines: integrity, transparency, and responsibility to data subjects
- Avoid data dredging and the 'significant/non-significant' binary
- Report effect sizes and confidence intervals alongside exact p-values
- Practice in R or Python with open datasets to prepare for causal inference

Key Takeaways and Continuing Your Learning Journey

- Question data sources, visualize first, and communicate uncertainty.
- Core foundations: descriptive stats, probability, CLT, intervals, tests, regression.
- Next steps: causal inference, predictive modeling, and Bayesian approaches.
- Stay current: engage with ASA/RSS, follow AI evaluation debates, practice on real data.