Computational NeuroEthology

Statistical approaches: From P-Values to Pipelines

Albert Einstein College of Medicine

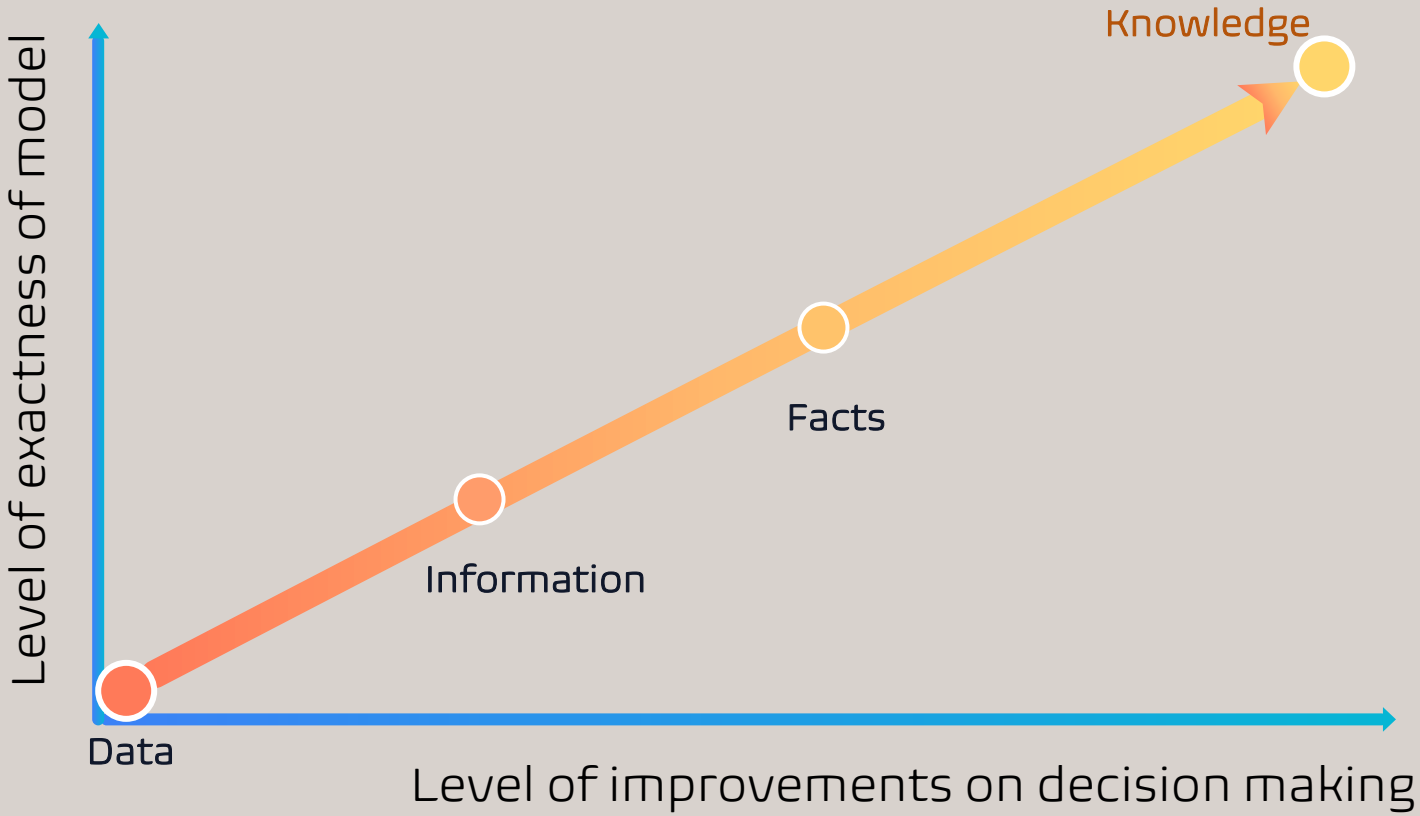
Class 2
October 6th, 2025

Mikhail Kislin

Overview

- Statistical Modeling Frameworks
- Modern Statistical Workflows
- Best Practices and Advanced Topics

00 What can statistics do for us?



1. Learning from data
stats by aggregation vs ML (all data)

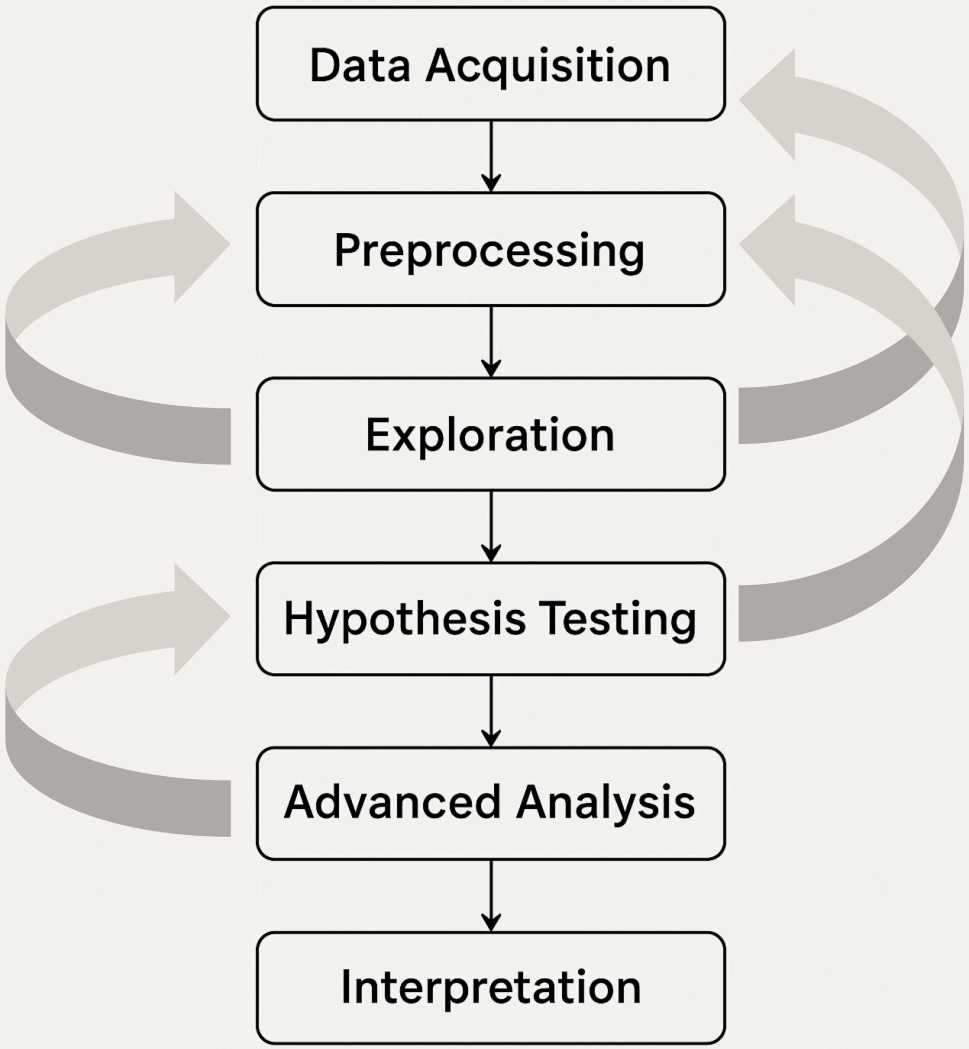
2. Deal with uncertainty and help to make decisions

3. Make inferences about causation

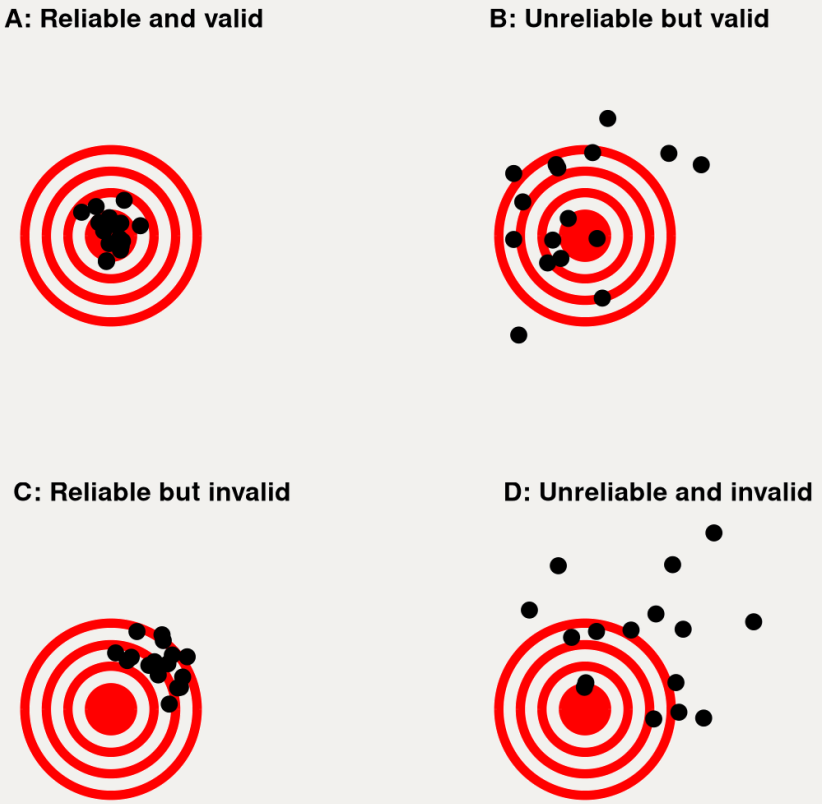
4. Make predictions about new situations based on priors

The Data Analysis Workflow

micro-meso-macro concept
of complex systems



What measurements produce a good data?



test-retest reliability

Face validity

Construct validity

Predictive validity

Computational validity: using computation to translate behaviors across species

Basic probability theory

Probability quantifies uncertainty:

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total outcomes}}$$

Complement Rule:

$$P(\text{not } A) = 1 - P(A)$$

Addition Rule (mutually exclusive):

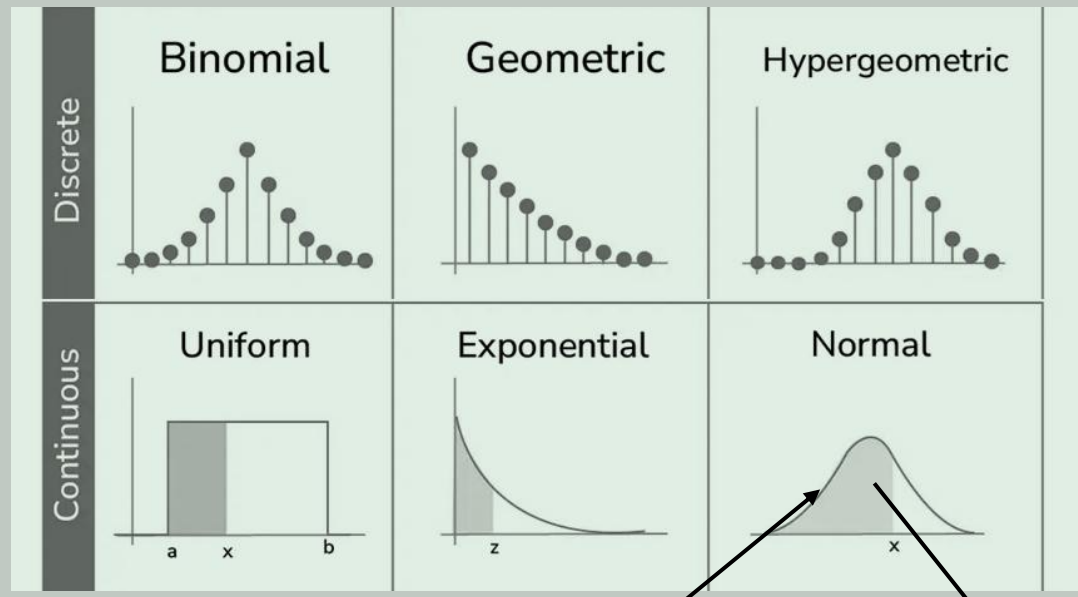
$$P(A \text{ or } B) = P(A) + P(B)$$

Multiplication Rule (independent events):

$$P(A \text{ and } B) = P(A) \times P(B)$$

Conditional Probability:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

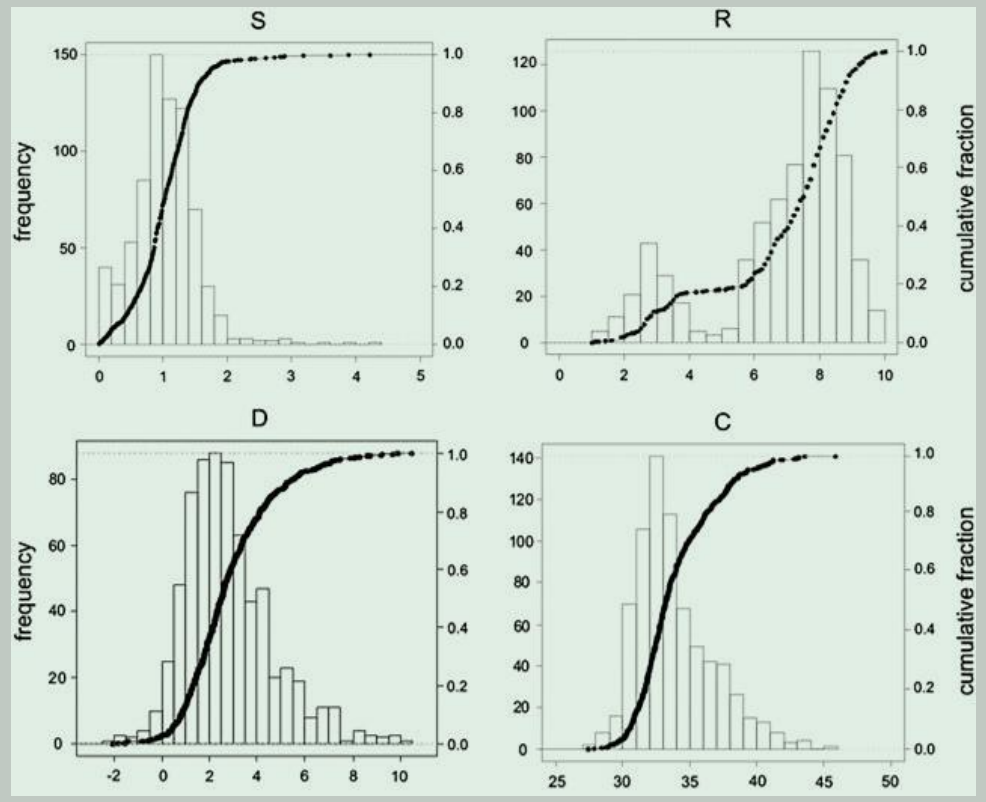


likelihood probability

Data sample

Frequentist Inference

Frequentist inference is the process of determining properties of an underlying distribution via the observation of data.



1. Summarize data

Expectation

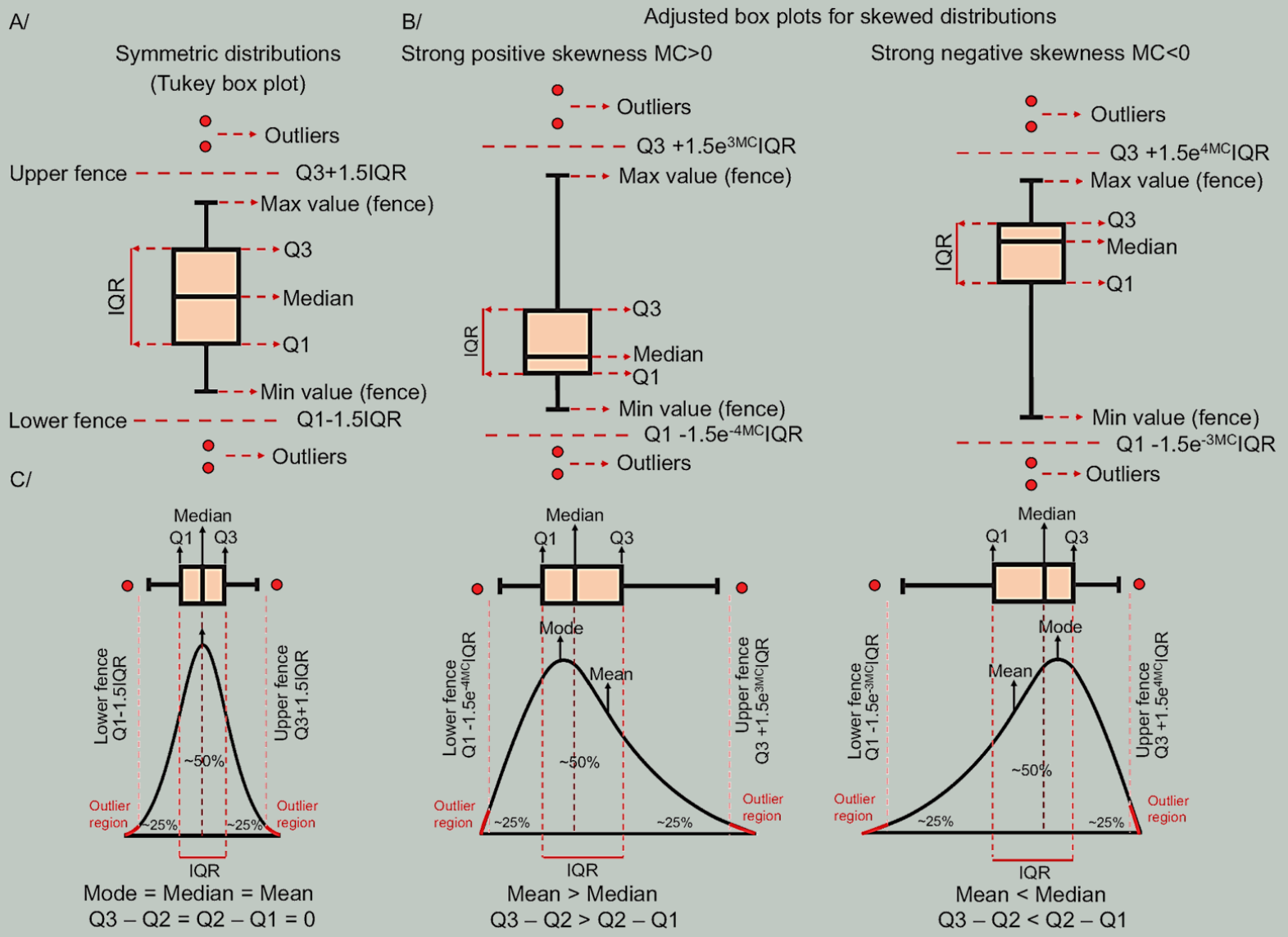
$$\mathrm{E}[X] = \sum_{x \in \mathcal{X}} xP(x)$$

Variance

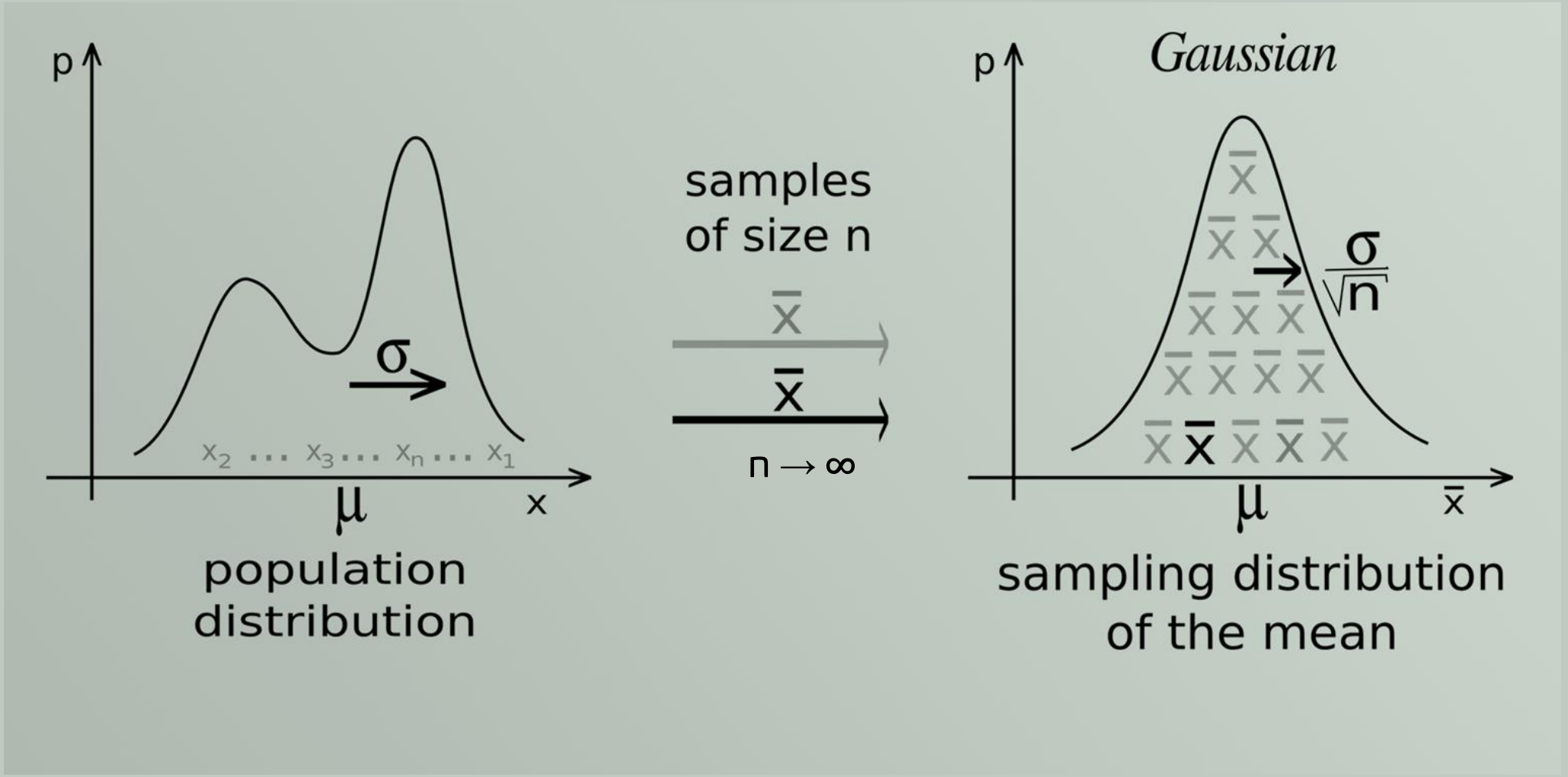
$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

Idealized representations of observed data

Components and construction of box plots

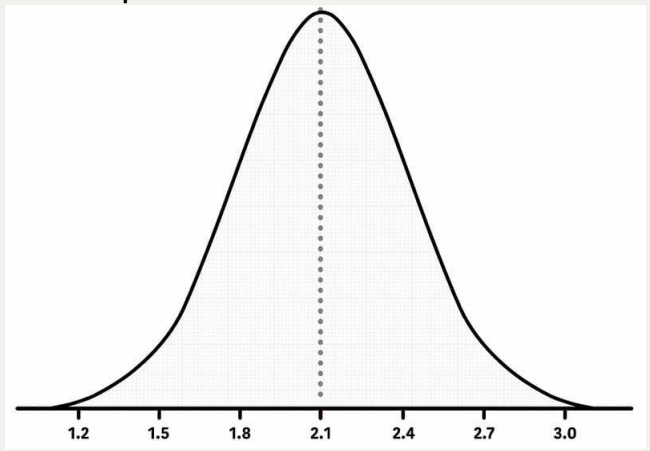


Central Limit Theorem



Hypothesis testing

Population distribution

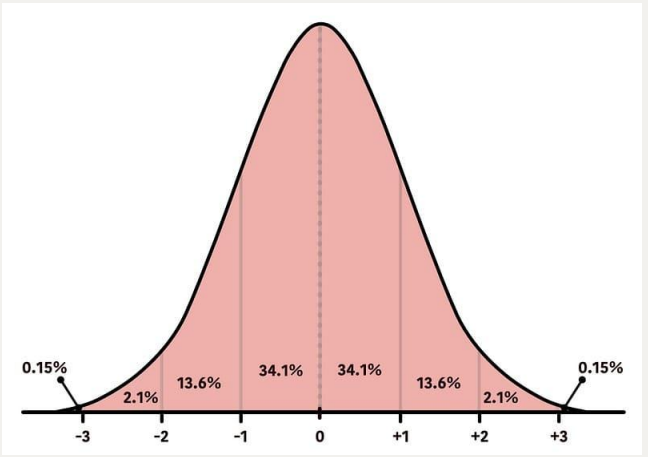


raw score

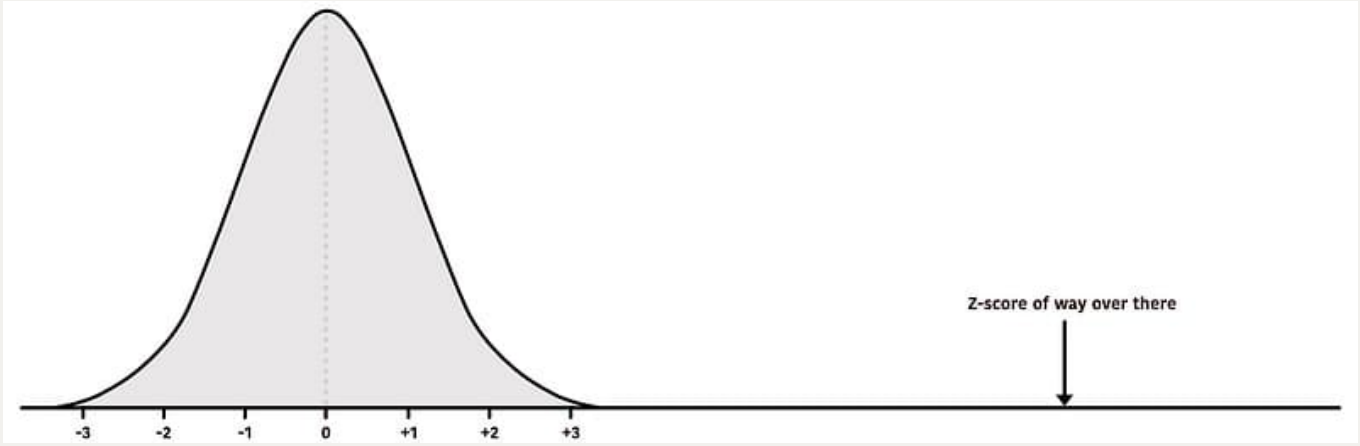
mean

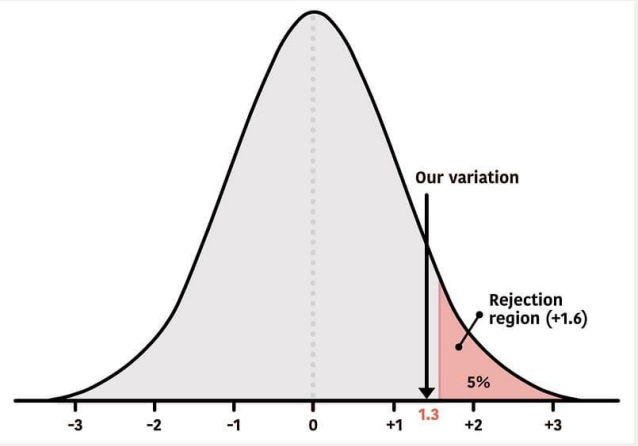
$$z = \frac{x - \mu}{\sigma}$$

standard deviation



How unusual your observed data is?

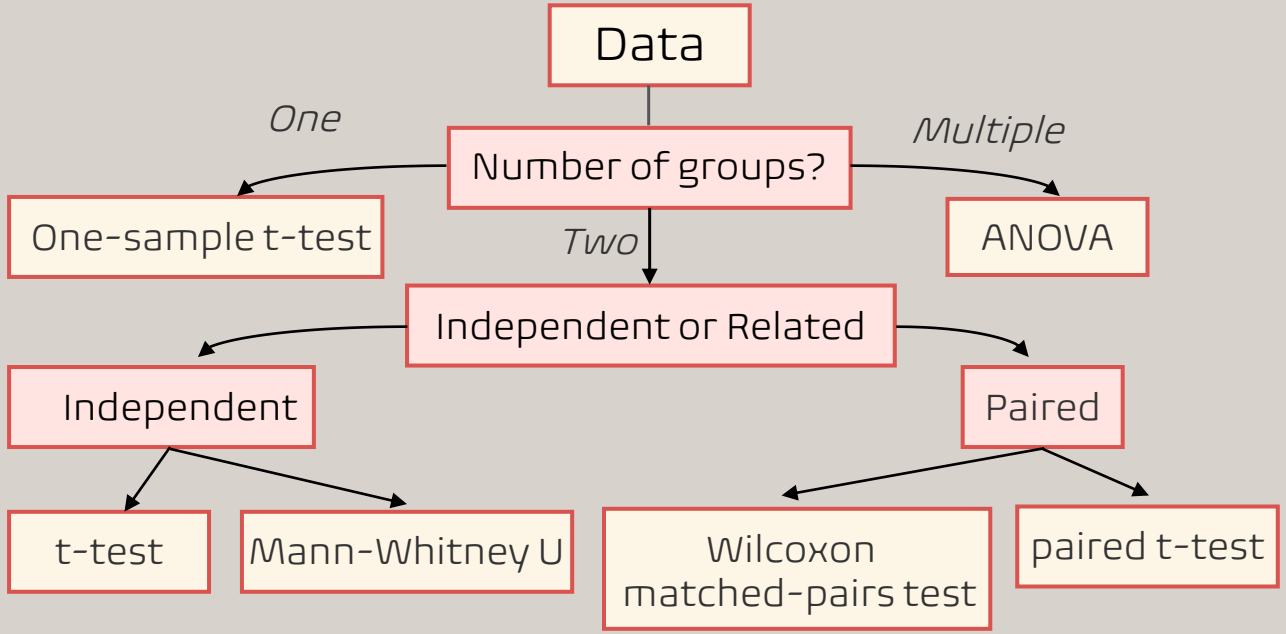




What is the problem with the "Cookbook" approach to statistics?

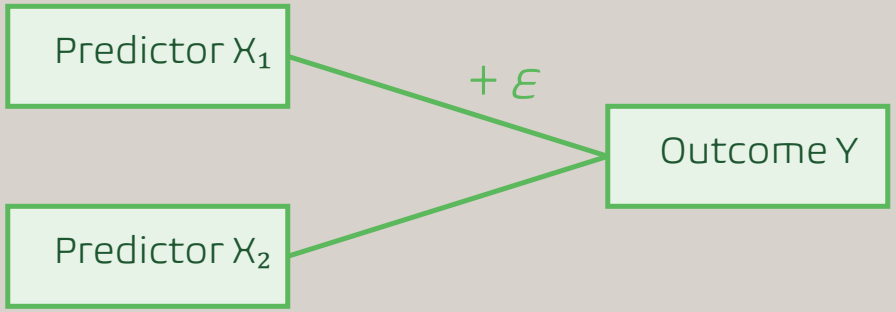
The "Cookbook" Approach

Focus: *Which test should I run?*



The Principled (Modeling) Approach

What process generated my data?

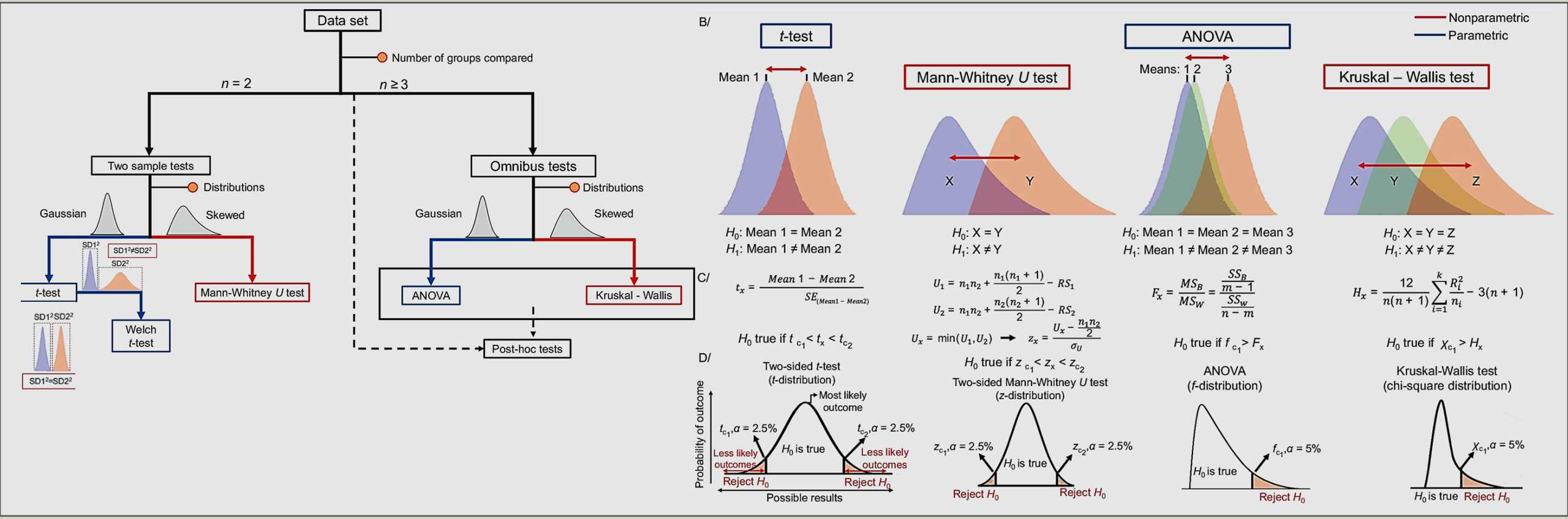


Consequence:

Shaky conclusions, risk of p-hacking, and a disconnect from the scientific question

Flexible insights, robust understanding of uncertainty, direct link to the scientific hypothesis.

Commonly used statistical tests

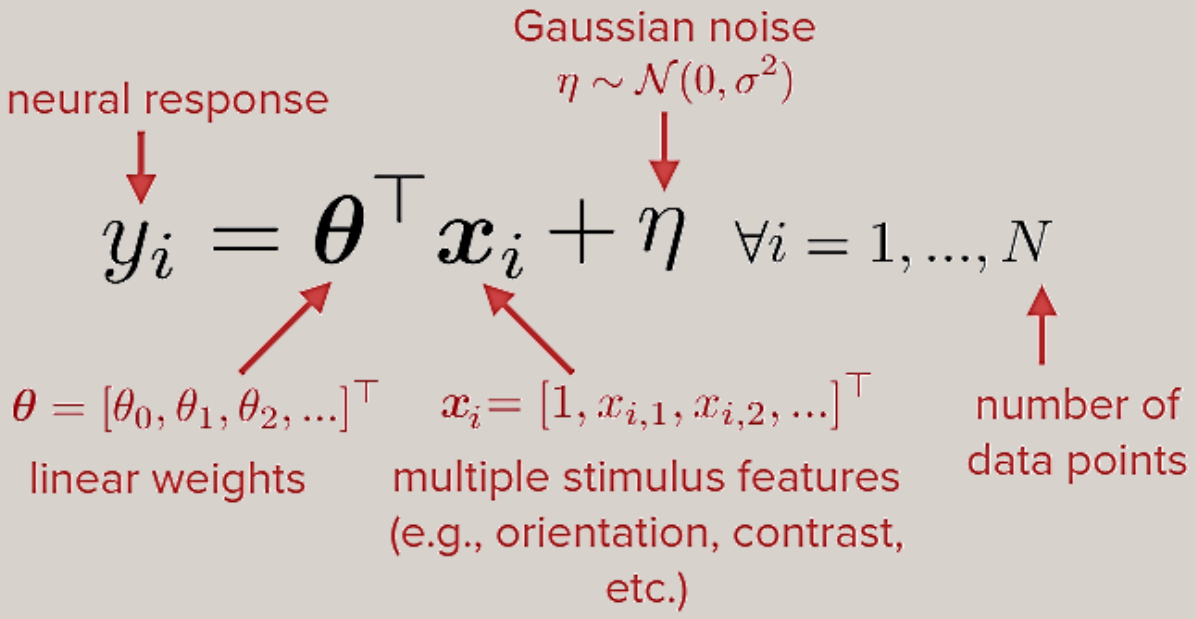


Idkowiak, J., et al. Nat Commun 2025

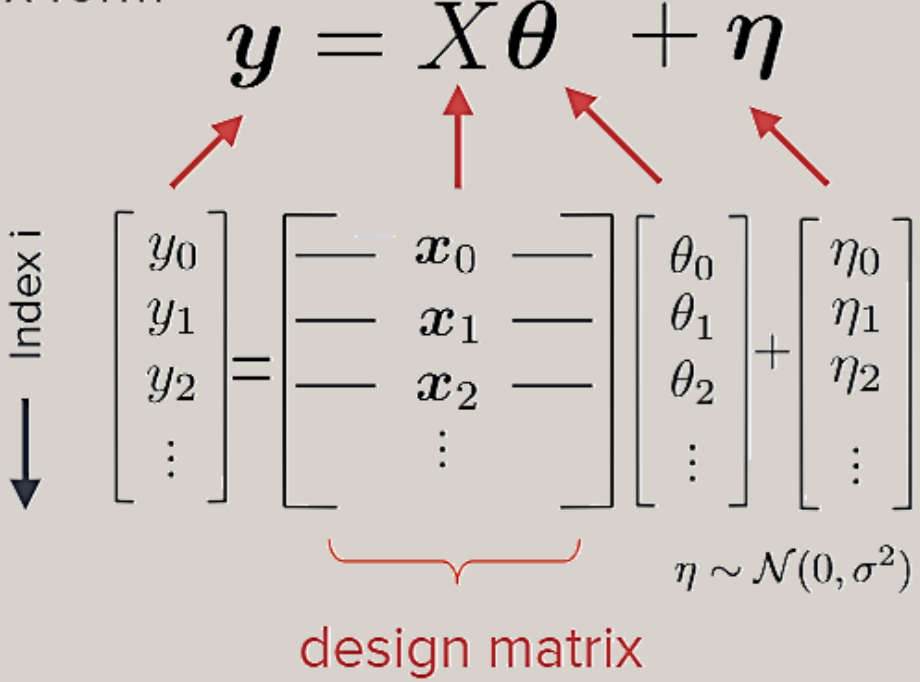
02 A Unified Framework: The General Linear Model

The GLM explains variation in Y as a linear combination of predictors

Vector form

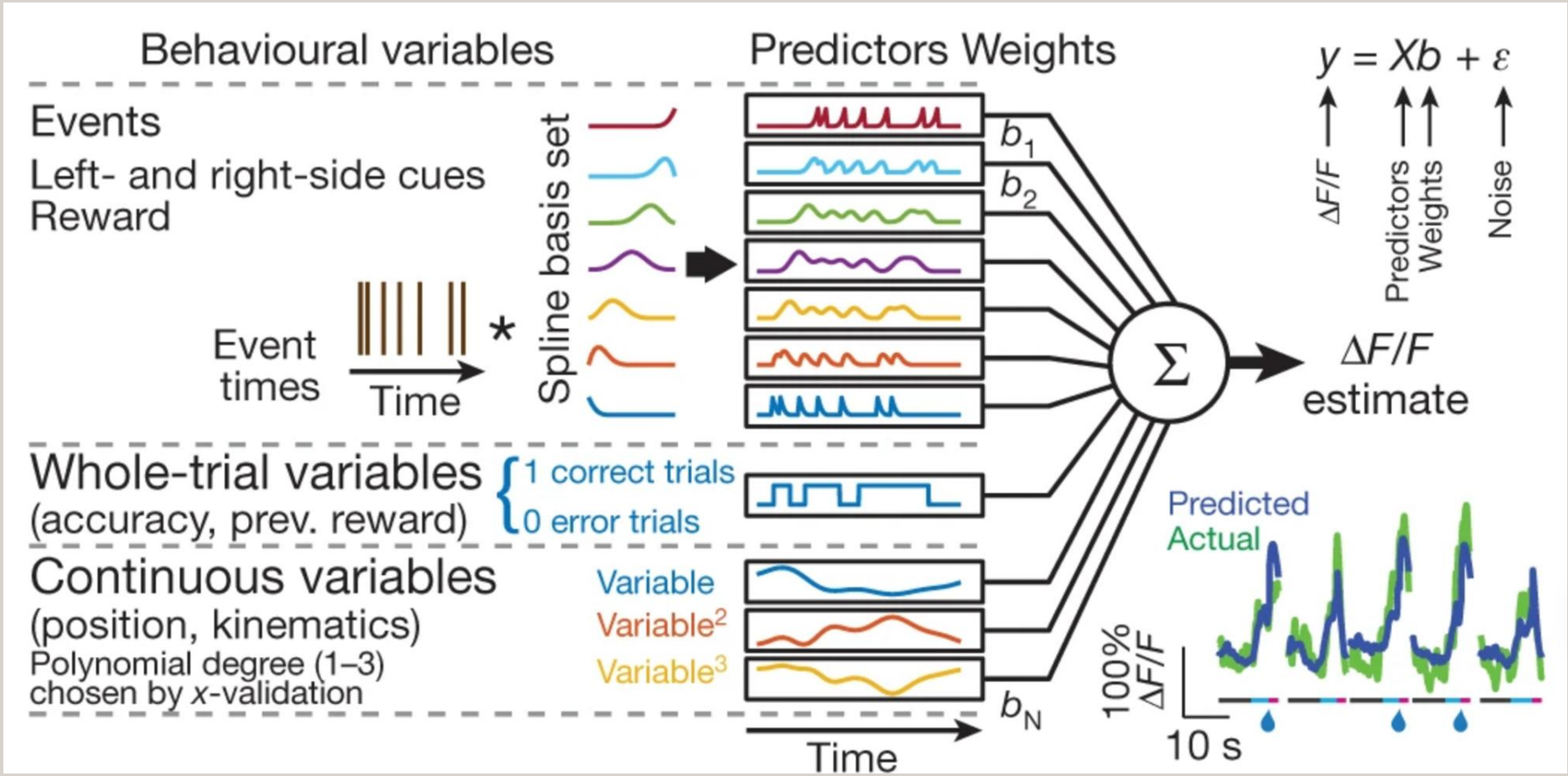


Matrix form



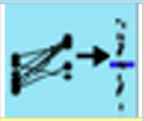
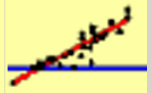
GLM for quantification responses to specific behavioral variables

encoding model to quantify the relationship between behavioral variables and the activity of neurons



Common statistical tests are linear models

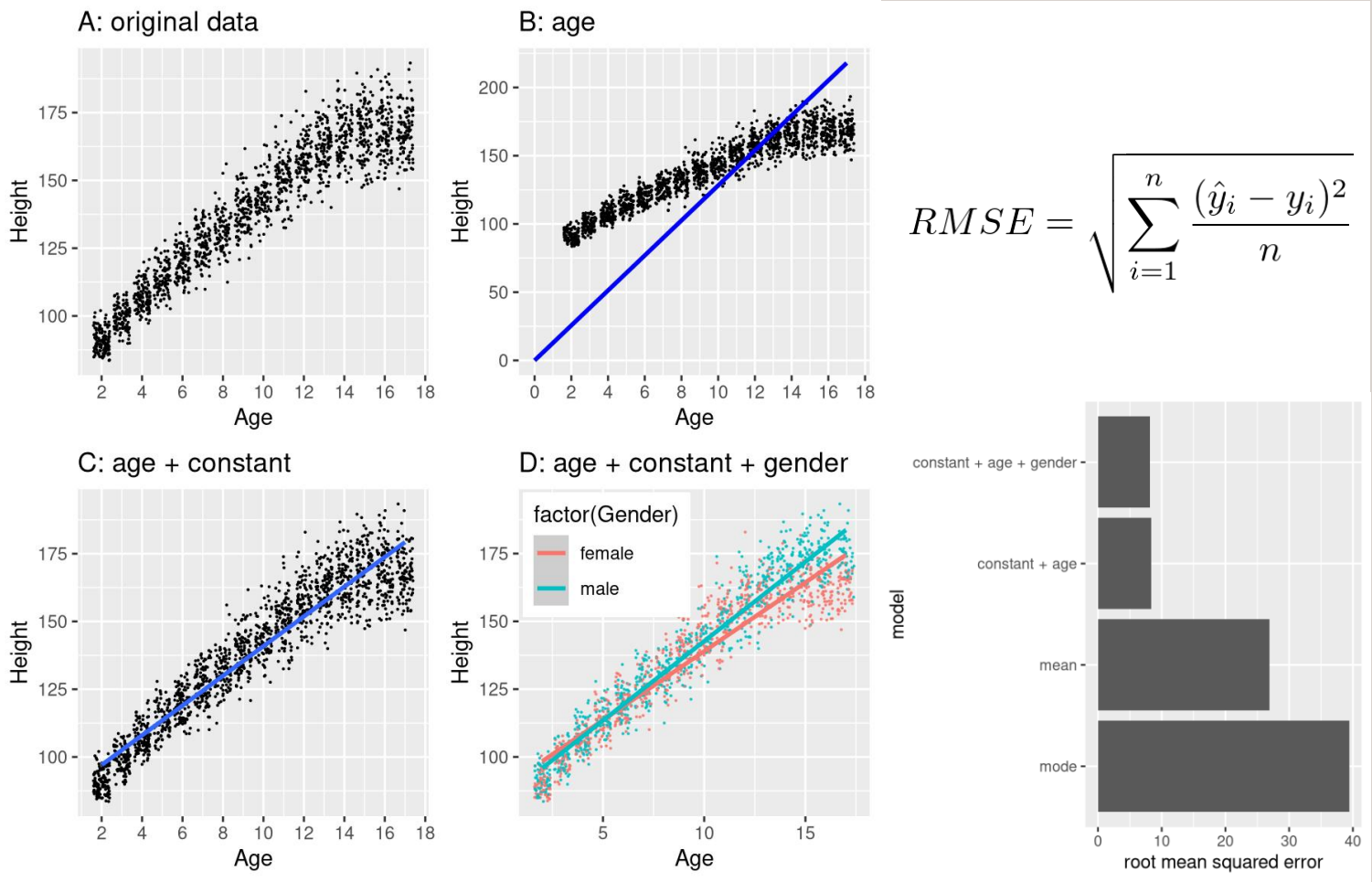
t-tests, ANOVA, regression – are all special cases of one model: $Y=X\beta+\varepsilon$

Common name	Function in scipy.stats	Equivalent linear model in smf.ols	Exact?	The linear model in words	Icon
y is independent of x P: One-sample t-test N: Wilcoxon signed-rank	scipy.stats.ttest_1samp(y) scipy.stats.wilcoxon(y)	smf.ols("y ~ 1", data) smf.ols("y ~ 1", signed_rank(data))	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
P: Paired-sample t-test N: Wilcoxon matched pairs	scipy.stats.ttest_rel(y1, y2) scipy.stats.wilcoxon(y1, y2)	smf.ols("y2_sub_y1 ~ 1", data) smf.ols("y2_sub_y1 ~ 1", signed_rank(data))	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
y ~ continuous x P: Pearson correlation N: Spearman correlation	scipy.stats.pearsonr(x, y) scipy.stats.spearmanr(x, y)	smf.ols("y ~ 1 + x", data) smf.ols("y ~ 1 + x", rank(data))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	scipy.stats.ttest_ind(y1, y2) N/A in Python, but see R version scipy.stats.mannwhitneyu(y1, y1)	smf.ols("y ~ 1 + group", data)[A] N/A in Python, but see R version smf.ols("y ~ 1 + group", signed_rank(data))[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
P: One-way ANOVA N: Kruskal-Wallis	scipy.stats.f_oneway(a, b, c) scipy.stats.kruskal(a, b, c)	smf.ols(y ~ 1 + G_2 + G_3 +…+ G_N)[A] smf.ols(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
P: One-way ANCOVA	N/A in Python, but see R version	smf.ols("y ~ 1 + G_2 + G_3 +…+ G_N + x", data)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	

Different GLMs can solve different problems

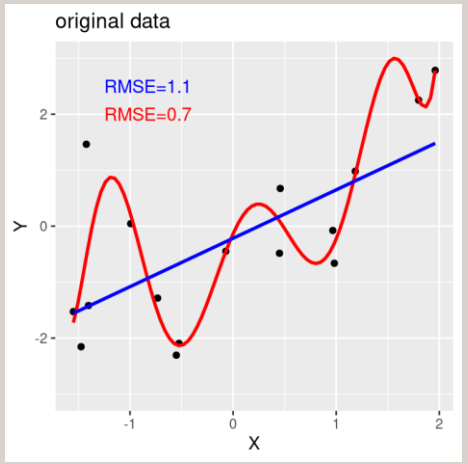
Output type	Likelihood	Nonlinearity		
real values	Gaussian $\mathrm{P}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(y-\mu)^2}{2\sigma^2}$	identity $\mu = \boldsymbol{\theta}^\top \mathbf{x}$		Linear regression
discrete counts 0,1,2,3...	Poisson $\mathrm{P}(y) = \frac{\lambda^y \exp(-\lambda)}{y!}$	exponential $\lambda = \exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)$		Poisson GLM
binary 0,1	Bernoulli $\mathrm{P}(y) = p^y (1-p)^{1-y}$	logistic $p = \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}\right)$		Logistic regression (Bernoulli GLM)

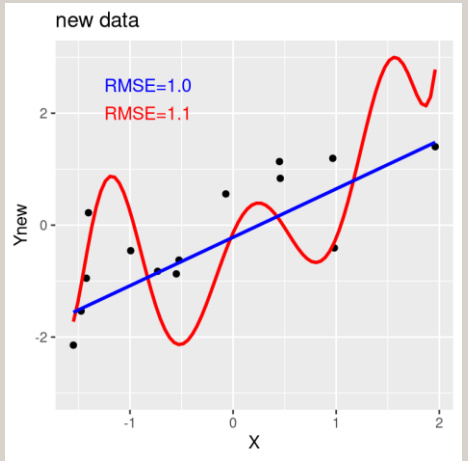
Improving GLM



$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

overfitting





Model Comparison Metrics AIC/BIC, adjusted R^2, likelihood ratio tests, cross-validation

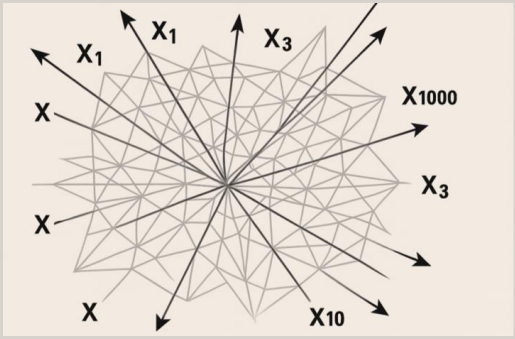
Diagnostic tools to confirm you used the right statistical test

Test Type	Key Assumptions to Check	Diagnostic Tools
t-test / ANOVA	Normality, equal variances	• Q–Q plot • Shapiro–Wilk test • Levene's test
Regression / GLM	Linearity, independence, homoscedasticity, normal residuals	• Residual plots • Durbin–Watson test • Variance Inflation Factor (VIF) • Breusch–Pagan test
Logistic regression	Linearity of logit, absence of separation	• Box–Tidwell test • ROC curve • Hosmer–Lemeshow test
Chi-square	Expected counts ≥ 5	• Inspect contingency table

Use bootstrapping or permutation tests to verify that your inference is not dependent on assumptions

03 The Challenge of High-Dimensional Data

Datasets often have many more features (f) than observations (n).
Traditional statistics break down when f ≫ n.



Curse of Dimensionality

Data become sparse in high-dimensional space.
Distances and similarities lose meaning.

Overfitting

Models capture noise, not signal. Perfect fit ≠ generalizable insight.

Multicollinearity

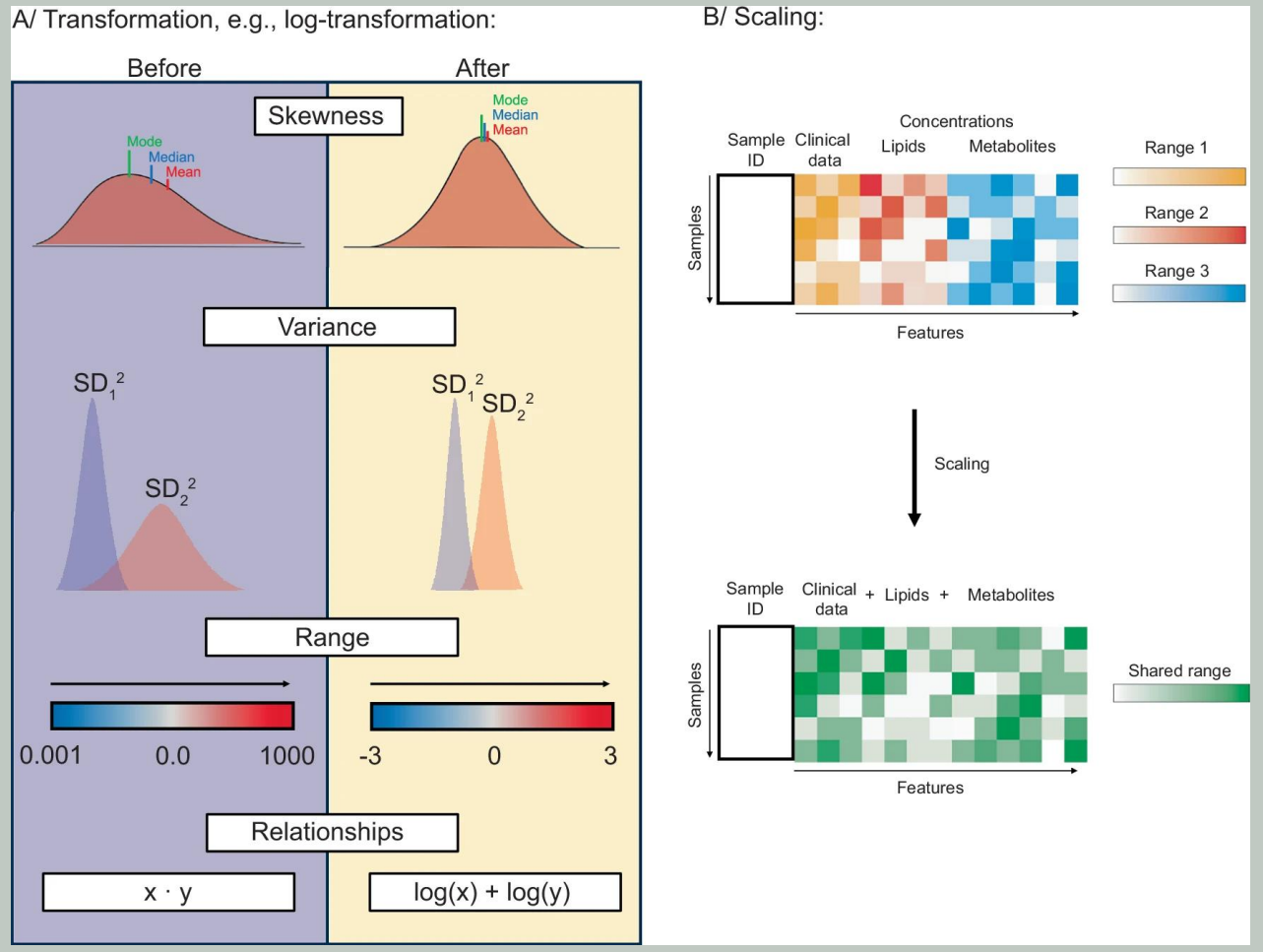
Predictors overlap → unstable estimates of effects.

Interpretability vs. Complexity Trade-off

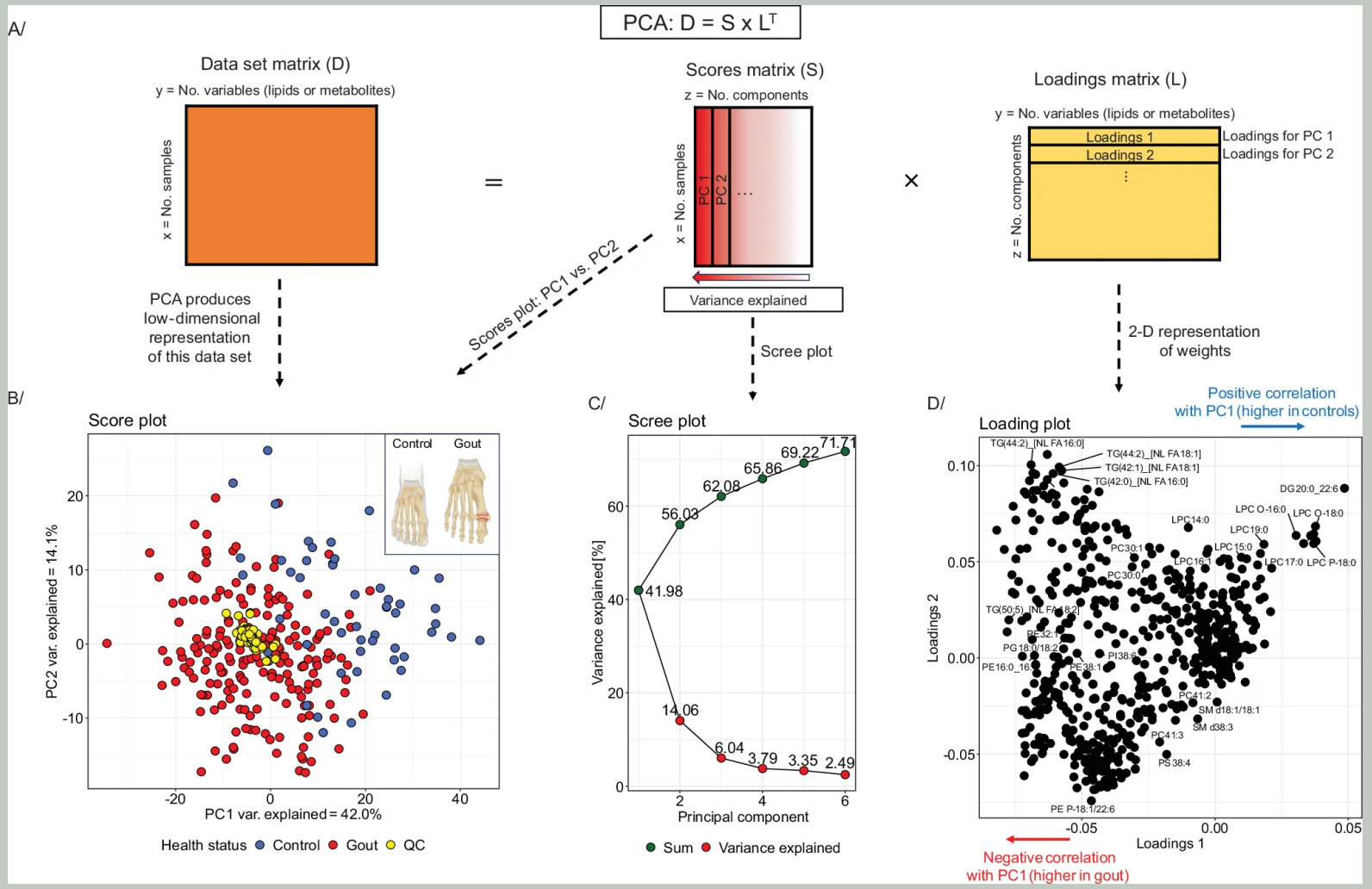
Simple models = understandable, limited power
Complex models = powerful, hard to interpret

Data Preprocessing and Exploration

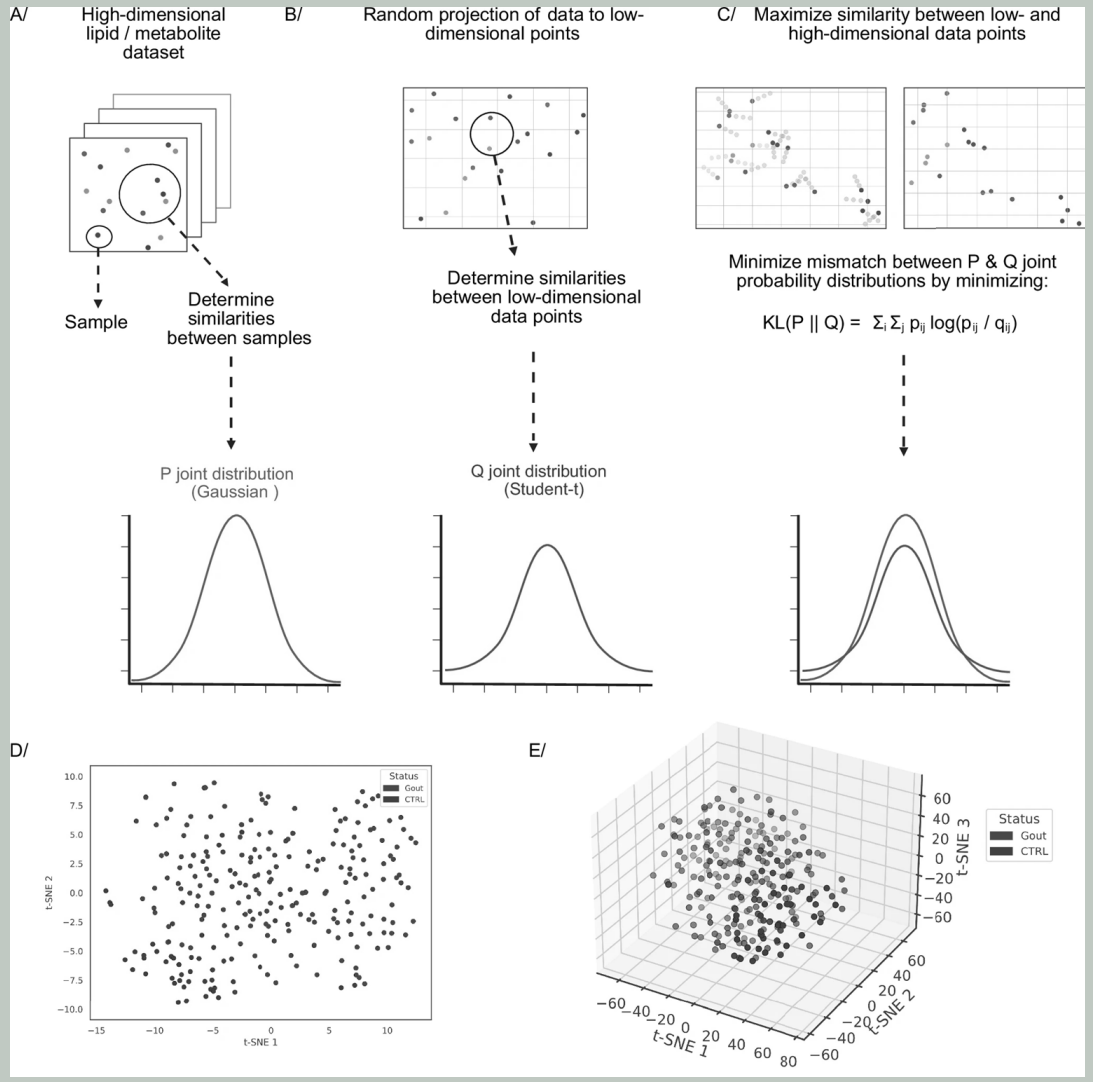
Data transformation and scaling



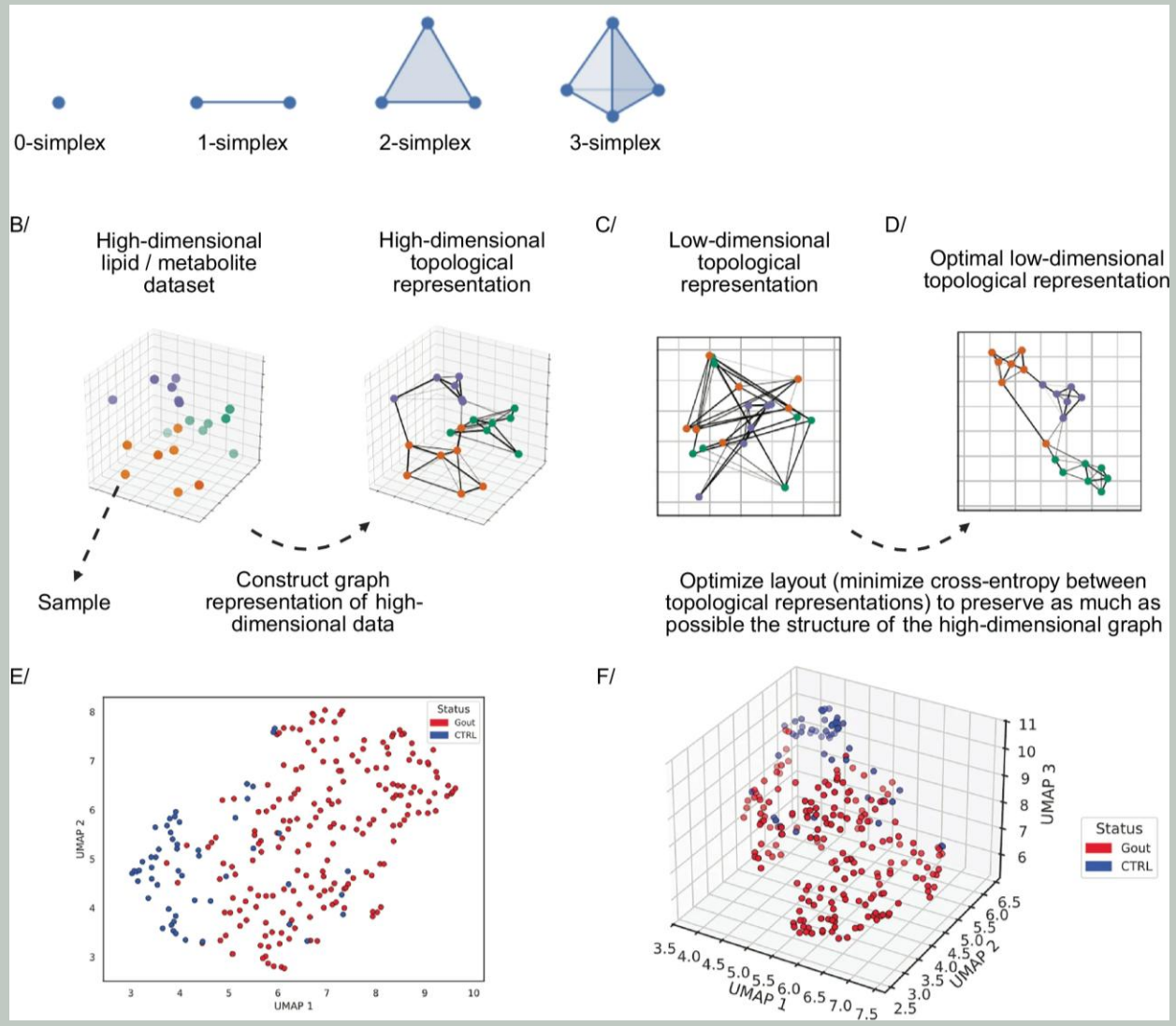
Overview of principal component analysis



t-Distributed Stochastic Neighbor Embedding

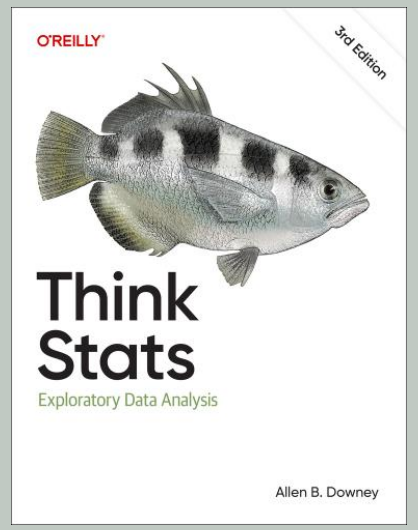


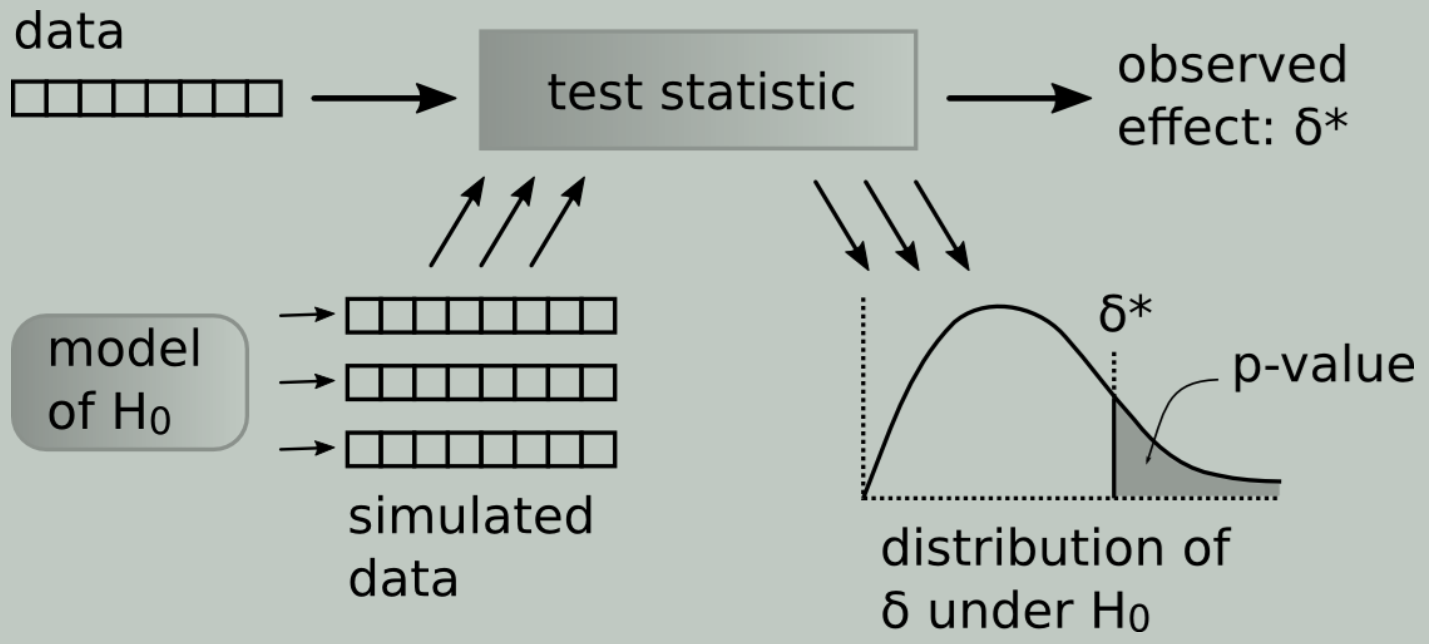
Uniform manifold approximation and projection



There is only one test!

Allen Downey





Libraries for statistical analysis

Time to practice
Open the Google Colaboratory (Colab)