

# Introduction to Statistical Modeling

## Introduction to Statistical Modeling

Key concepts, workflow, and common techniques

---

## 1. What Is Statistical Modeling?

| Aspect     | Description                                                                                                                                                                                                                                                                                                                                          |
|------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Definition | A mathematical representation of a data-generating process that uses probability theory to describe uncertainty.                                                                                                                                                                                                                                     |
| Purpose    | <ul style="list-style-type: none"><li>• Summarize observed data</li><li>• Infer relationships between variables</li><li>• Predict future or unseen outcomes</li><li>• Quantify uncertainty and test hypotheses</li></ul>                                                                                                                             |
| Components | <ol style="list-style-type: none"><li>1. <b>Random variables</b> (observed &amp; latent)</li><li>2. <b>Probability distribution</b> that governs the variables</li><li>3. <b>Parameters</b> that characterize the distribution (e.g., mean, variance)</li><li>4. <b>Model structure</b> (functional form, covariates, hierarchical levels)</li></ol> |

---

## 2. Modeling Workflow

### 1. Problem Definition

- Identify the scientific or business question.
- Decide whether the goal is *description*, *inference*, or *prediction*.

### 2. Data Exploration & Pre-processing

- Summary statistics, visualizations, missing-value analysis.
- Transformations (log, sqrt), scaling, encoding categorical variables.

### 3. Model Specification

- Choose a family of probability distributions (e.g., Normal, Binomial).
- Define the systematic component (linear predictor, link function).
- Add random effects or hierarchical levels if needed.

# Introduction to Statistical Modeling

## 4. Parameter Estimation

- **Frequentist:** Maximum Likelihood (ML), Restricted ML (REML).
- **Bayesian:** Posterior inference via Markov Chain Monte Carlo (MCMC) or Variational Inference.

## 5. Model Checking & Validation

- Goodness-of-fit (residual plots, QQ-plots).
- Predictive performance (cross-validation, out-of-sample  $R^2$ , AUC).
- Diagnostic tests (e.g., overdispersion, multicollinearity).

## 6. Interpretation & Communication

- Translate parameter estimates into substantive statements.
- Visualize effects (effect plots, coefficient plots).
- Report uncertainty (confidence intervals, credible intervals).

## 7. Deployment (optional)

- Export model objects, create APIs, monitor performance over time.

---

## 3. Core Model Families

| Model                | Typical Outcome      | Link Function            | Example Use                   |
|----------------------|----------------------|--------------------------|-------------------------------|
| Linear Regression    | Continuous           | Identity                 | Predict house price from size |
| Logistic Regression  | Binary (0/1)         | Logit                    | Disease presence/absence      |
| Poisson Regression   | Counts               | Log                      | Number of calls per hour      |
| Negative Binomial    | Overdispersed counts | Log                      | Accident counts               |
| Ordinal Regression   | Ordered categories   | Logit / Probit           | Customer satisfaction rating  |
| Survival (Cox) Model | Time-to-event        | N/A (partial likelihood) | Time until equipment failure  |

# Introduction to Statistical Modeling

| Model                                         | Typical Outcome                   | Link Function | Example Use                   |
|-----------------------------------------------|-----------------------------------|---------------|-------------------------------|
| Generalized Linear Mixed Models (GLMMs)       | Any of the above + random effects | Same as GLM   | Repeated measures on patients |
| Hierarchical Bayesian Models                  | Any                               | Flexible      | Multi-level education data    |
| Time Series Models (ARIMA, State Space)       | Continuous/Counts over time       | N/A           | Forecast sales                |
| Latent Variable Models (Factor Analysis, SEM) | Continuous/ordinal                | N/A           | Psychological constructs      |

---

## 4. Key Statistical Concepts

| Concept    | Why It Matters      | Quick Reminder |
|------------|---------------------|----------------|
| Likelihood | Basis for parameter |                |