# Sampling and the Data-Generating Process

Statistics · Module 2 — Statistical Models

Node 2.2

## Position in Knowledge Graph

**System:**
Statistics

**Structural Domain:**
Statistical Models and Likelihood

**Node:**
2.2

---

## Formal Definition

Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a statistical model.

A dataset:

$$X_1, \ldots, X_n$$

is assumed to be generated according to:

$$X_i \sim \mathbb{P}_\theta,$$

often with the additional assumption that the observations are **independent and identically distributed (i.i.d.)**.

---

## Intuition

A statistical model alone is not enough.

We also assume a mechanism:

> Reality produces data from one unknown distribution in the model family.

This mechanism is called the **data-generating process (DGP).**

Statistics attempts to reverse-engineer the DGP.

---

## Structural Role

Sampling assumptions determine:

- Joint distribution of the data
- Likelihood structure
- Variance scaling
- Asymptotic behavior

For i.i.d. sampling:

$$\mathbb{P}_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i).$$

This factorization is the backbone of likelihood theory.

---

## Mathematical Development

### Joint Density (Continuous Case)

If density exists:

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

### Non-i.i.d. Case

More generally:

$$f_\theta(x_1, \ldots, x_n) \neq \prod_{i=1}^{n} f_\theta(x_i).$$

Examples include:

- Time series
- Markov chains
- Dependent sampling

---

## Worked Example

### Normal Model

If:

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

then joint density:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(x_i - \mu)^2}{2\sigma^2} \right).$$

---

## Cross-Links

- **AI:** Training data assumed i.i.d. in ERM.
- **Economics:** Structural models assume sampling from populations.
- **Quantum:** Repeated measurement outcomes under identical preparation.
- **Linear Algebra:** Vector of observations as random vector.

---

## Structural Compression

**Invariant:**
Joint distribution determined by sampling structure.

**Mechanism:**
Independence factorizes probability.

**Cross-System Echo:**
In ML, mini-batch sampling approximates full distribution.

---

## Exercises

1. Derive joint density for Bernoulli model.
2. Explain how dependent sampling changes likelihood.
3. Give example where i.i.d. assumption fails.

---

*Statistics · Module 2 — Statistical Models · Node 2.2*
**Concepts:** random-variables, probability-space

*hbar.university — own.your.cognition*