Module 2 — Statistical Models

Statistics

hbar.university

Statistical Models

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.1

Formal Definition

A **statistical model** is a family of probability distributions:

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}.$$

- Θ is the parameter space.
- Each θ indexes a probability distribution.
- Data is assumed to arise from one unknown member of this family.

Intuition

Probability theory assumes the distribution is known.

Statistics assumes:

> The distribution belongs to a structured family, but we do not know which one.

The parameter θ encodes the unknown structure of reality.

Inference means learning θ from data.

Structural Role

Statistical models:

- Bridge probability and data.
- Introduce parameters.
- Define the object of estimation.

- Enable likelihood construction.

Without models, data cannot inform unknown structure.

Mathematical Development

Parametric Model

Finite-dimensional parameter space:

$$\theta \in \mathbb{R}^k.$$

Example: Normal model

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

$$\theta = (\mu, \sigma^2).$$

Nonparametric Model

Infinite-dimensional parameter space.

Example: All continuous distributions on $\mathbb{R}$.

Identifiability

A model is **identifiable** if:

$$\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Rightarrow \theta_1 = \theta_2.$$

Otherwise, parameters cannot be uniquely recovered.

Worked Example

Bernoulli Model

$$X_i \sim \text{Bernoulli}(p), \quad p \in [0, 1].$$

Parameter p fully determines the distribution.

Data informs p.

Cross-Links

- **Linear Algebra:** Parameter vectors as elements of $\mathbb{R}^k$.
- **AI:** Model class in machine learning.
- **Economics:** Structural models of agents.
- **Quantum:** Quantum states parameterize measurement distributions.

Structural Compression

Invariant:
Model = structured family of distributions.

Mechanism:
Unknown parameter indexes probability measure.

Cross-System Echo:
In ML, hypothesis class defines learning space.

Exercises

1. Define a Poisson statistical model.
2. Give an example of a non-identifiable model.
3. Explain difference between parametric and nonparametric models.

Sampling and the Data-Generating Process

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.

Likelihood Function

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.3

Formal Definition

Let:

$$X_1, \ldots, X_n$$

be observed data from a statistical model:

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}.$$

The **likelihood function** is defined as:

$$L(\theta; x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n),$$

viewed as a function of θ, with the data fixed.

Under the i.i.d. assumption:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

Intuition

Probability asks:

> Given θ, how likely is the data?

Likelihood asks:

> Given the data, how plausible is θ?

The same formula. Different perspective.

Likelihood reverses the direction of reasoning.

Structural Role

Likelihood is the engine of:

- Maximum likelihood estimation (MLE)
- Likelihood ratio tests
- Asymptotic theory
- Information measures
- Bayesian updating (via Bayes' rule)

It is the bridge between model and inference.

Mathematical Development

Key Distinction

Probability:

$$f_\theta(x)$$

is a function of data x.

Likelihood:

$$L(\theta; x)$$

is a function of parameter θ.

The functional role changes.

Example: Bernoulli Model

If $X_i \sim \text{Bernoulli}(p)$,

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

This simplifies to:

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

Likelihood depends only on the number of successes.

Worked Example

Let:

$$\sum x_i = k.$$

Then:

$$L(p) = p^k(1-p)^{n-k}.$$

The parameter p that maximizes this will be the MLE.

Cross-Links

- **Linear Algebra:** Log-likelihood often leads to quadratic forms.
- **AI:** Empirical risk minimization equals negative log-likelihood minimization.
- **Economics:** Structural estimation uses likelihood maximization.
- **Quantum:** Likelihood used in quantum state tomography.

Structural Compression

Invariant:
Likelihood is the joint density viewed as function of parameter.

Mechanism:
Fix data, vary parameter.

Cross-System Echo:
In ML, model parameters are optimized against fixed dataset.

Exercises

1. Derive likelihood for Poisson model.
2. Show that likelihood depends only on sufficient statistics in Bernoulli case.
3. Explain difference between probability and likelihood.

Log-Likelihood and the Score Function

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.4

Formal Definition

Given the likelihood function:

$$L(\theta; x_1, \ldots, x_n),$$

the **log-likelihood** is defined as:

$$\ell(\theta) = \log L(\theta).$$

Under i.i.d. sampling:

$$\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i).$$

Score Function

The **score function** is the gradient of the log-likelihood:

$$S(\theta) = \frac{\partial}{\partial \theta} \ell(\theta).$$

For vector parameters:

$$S(\theta) = \nabla_\theta \ell(\theta).$$

Intuition

Taking logs turns products into sums.

Optimization becomes tractable.

The score measures:

> How sensitive the likelihood is to changes in the parameter.

If the score is zero, we are at a stationary point.

Structural Role

Log-likelihood enables:

- Maximum Likelihood Estimation (MLE)
- Derivation of estimating equations
- Asymptotic normality
- Fisher Information
- Numerical optimization

The score is the foundation of modern inference.

Mathematical Development

Additivity

Because of independence:

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta).$$

The score becomes:

$$S(\theta) = \sum_{i=1}^{n} S_i(\theta).$$

This additive structure drives asymptotic theory.

Example: Bernoulli Model

If:

$$L(p) = p^k (1-p)^{n-k},$$

then:

$$\ell(p) = k \log p + (n-k) \log(1-p).$$

Score:

$$\frac{d}{dp}\ell(p) = \frac{k}{p} - \frac{n-k}{1-p}.$$

Setting score to zero gives:

$$\hat{p} = \frac{k}{n}.$$

Cross-Links

- **Linear Algebra:** Score is gradient; Hessian gives curvature.
- **AI:** Training neural networks minimizes negative log-likelihood.
- **Economics:** Structural models solved via likelihood maximization.
- **Quantum:** Log-likelihood used in state reconstruction algorithms.

Structural Compression

Invariant:
Log preserves maxima of likelihood.

Mechanism:
Product $\rightarrow$ sum $\rightarrow$ gradient optimization.

Cross-System Echo:
In ML, loss gradients guide parameter updates.

Exercises

1. Derive log-likelihood for Poisson model.
2. Compute score for Normal mean with known variance.
3. Show that maximizing likelihood equals maximizing log-likelihood.
4. Explain why logs simplify asymptotic analysis.

Fisher Information

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.5

Formal Definition

Let $\ell(\theta)$ be the log-likelihood for a single observation.

The **Fisher Information** is defined as:

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right].$$

Equivalently, under regularity conditions:

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right].$$

For n i.i.d. observations:

$$\mathcal{I}_n(\theta) = n\mathcal{I}(\theta).$$

Intuition

Fisher Information measures how sharply peaked the likelihood is.

- Flat likelihood → low information
- Sharp likelihood → high information

It quantifies how much the data tells us about the parameter.

It is curvature of the log-likelihood in expectation.

Structural Role

Fisher Information underlies:

- Cramér–Rao lower bound
- Asymptotic normality of MLE
- Efficiency
- Information geometry
- Hypothesis testing

It introduces metric structure on parameter space.

Mathematical Development

Score Variance Form

Since:

$$S(\theta) = \frac{\partial}{\partial \theta} \log f_\theta(X),$$

and

$$\mathbb{E}[S(\theta)] = 0,$$

we have:

$$\mathcal{I}(\theta) = \mathrm{Var}(S(\theta)).$$

Thus, information equals variance of the score.

Example: Bernoulli(p)

$$\log f_p(x) = x \log p + (1 - x) \log(1 - p).$$

Score:

$$S(p) = \frac{x}{p} - \frac{1 - x}{1 - p}.$$

Then:

$$\mathcal{I}(p) = \frac{1}{p(1 - p)}.$$

Information is highest near boundaries.

Cross-Links

- **Linear Algebra:** Information matrix is positive semidefinite.
- **AI:** Natural gradient uses Fisher Information metric.
- **Economics:** Efficiency bounds in estimation.
- **Quantum:** Quantum Fisher Information generalizes classical case.

Structural Compression

Invariant:
Information ≥ 0.

Mechanism:
Curvature of expected log-likelihood.

Cross-System Echo:
In optimization, curvature determines step size and stability.

Exercises

1. Prove equivalence of the two definitions of Fisher Information.
2. Compute Fisher Information for Normal mean with known variance.
3. Explain why information scales linearly with sample size.
4. Show that score has zero expectation under regularity conditions.

Exponential Families

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.6

Formal Definition

A statistical model belongs to the **exponential family** if its density can be written as:

$$f_\theta(x) = h(x) \exp\left(\eta(\theta)^\top T(x) - A(\theta)\right),$$

where:

- $T(x)$ is the sufficient statistic,
- $\eta(\theta)$ is the natural parameter,
- $A(\theta)$ is the log-partition function,
- $h(x)$ is the base measure.

Intuition

Exponential families separate:

- Data structure $T(x)$
- Parameter structure $\eta(\theta)$

All parameter dependence flows through a low-dimensional summary $T(x)$.

They are maximally structured statistical models.

Structural Role

Exponential families:

- Guarantee existence of sufficient statistics
- Simplify likelihood analysis
- Enable conjugate priors in Bayesian inference
- Underlie generalized linear models
- Connect to information geometry

They are the algebraic backbone of classical statistics.

Mathematical Development

Log-Likelihood Form

For i.i.d. data:

$$\ell(\theta) = \eta(\theta)^\top \sum_{i=1}^{n} T(x_i) - nA(\theta) + \sum_{i=1}^{n} \log h(x_i).$$

Parameter dependence appears only through:

$$\sum_{i=1}^{n} T(x_i).$$

This implies sufficiency.

Mean and Variance Structure

The log-partition function satisfies:

$$\nabla_\theta A(\theta) = \mathbb{E}_\theta[T(X)].$$

Second derivative:

$$\nabla^2_\theta A(\theta) = \mathrm{Var}_\theta(T(X)).$$

Thus, curvature encodes variance.

Examples

Bernoulli

$$f_p(x) = \exp\left(x \log \frac{p}{1-p} + \log(1-p)\right).$$

Natural parameter:

$$\eta = \log \frac{p}{1-p}.$$

Normal (Known Variance)

$$f_\mu(x) = \exp\left(\frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) h(x).$$

Cross-Links

- **Linear Algebra:** Natural parameter space as vector space.
- **AI:** Softmax, logistic regression are exponential family models.
- **Economics:** Logit models in discrete choice.
- **Quantum:** Gibbs states resemble exponential family structure.

Structural Compression

Invariant:
Parameter enters linearly in sufficient statistic.

Mechanism:
Log-partition function controls normalization and variance.

Cross-System Echo:
In physics, Boltzmann distributions are exponential family models.

Exercises

1. Show Poisson distribution is exponential family.
2. Derive sufficient statistic for Normal mean.
3. Prove that expectation of sufficient statistic equals gradient of $A(\theta)$.
4. Explain why exponential families are closed under i.i.d. sampling.