Fundamentals of Mathematical Statistics

Daniele Cambria

January 9, 2026

1 Notation

We consider a probability space $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ a σ-algebra of events, and P a probability measure.

- Random variables $X_1, \ldots, X_n$ are measurable functions
$$X_i : \Omega \to \mathcal{X},$$
where $\mathcal{X}$ is the observation space (e.g., $\mathbb{R}$ or $\mathbb{R}^d$) equipped with an appropriate σ-algebra.
- A realization or measurement of X_i is denoted by $x_i \in \mathcal{X}$. The observed data vector is
$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n,$$
corresponding to the random vector $\mathbf{X} = (X_1, \ldots, X_n)$.
- The joint distribution of $\mathbf{X}$ is denoted by P, which belongs to a collection (statistical model) $\mathcal{P}$ of possible distributions of $\mathbf{X}$.
- In parametric models, $\mathcal{P}$ is indexed by parameters $\theta \in \Theta$, so that
$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$
- Often, the random variables are assumed independent and identically distributed (iid) with distribution P_θ on $\mathcal{X}$. Then $\mathbf{X}$ has distribution
$$P_\theta \otimes \ldots \otimes P_\theta = P_\theta^{\otimes n} \text{ on } \mathcal{X}^n,$$
where $P_\theta^{\otimes n}$ is the n-fold product distribution.
- A parameter of interest is typically a function $\gamma = g(\theta)$, defined on Θ.

2 Estimation

Definition: Estimator

An estimator (or statistic, or decision) is a known measurable function
$$T : \mathcal{X}^n \to \mathbb{R}$$
evaluated at the random vector $\mathbf{X}$, i.e. the estimator random variable is $T(\mathbf{X})$. The function T itself must not depend on unknown parameters, since we want to compute it using only the data we have observed.

Likelihood function and MLE

For data $\mathbf{X} = (X_1, \ldots, X_n)$, the *likelihood function* is
$$L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} p_\theta(X_i), \quad \theta \in \Theta.$$

A *maximum likelihood estimator* (MLE) is any
$$\hat{\theta} \in \arg\max_{\theta \in \Theta} L_{\mathbf{X}}(\theta).$$

Equivalently, $\hat{\theta}$ maximizes the *log-likelihood*:
$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p_\theta(X_i).$$

3 Intermezzo

3.1 Conditional Distributions

The conditional probability of A given B is defined as
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

Continuous Case

Given two continuous random vectors $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ defined on the same probability space and joint density $f_{X,Y}(x, y)$, the marginal density of X is given by
$$f_X(x) = \int_{\mathbb{R}^m} f_{X,Y}(x, y)\, dy.$$
The conditional density of X given $Y = y$ is
$$f_X(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$
If $f_Y(y) > 0$, and 0 otherwise.
For measurable $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, the conditional expectation given $Y = y$ is
$$\mathbb{E}[g(X, Y) \mid Y = y] = \int_{\mathbb{R}^n} g(x, y) f_X(x|y)\, dx.$$

Discrete Case

Suppose X and Y are discrete random vectors taking values in $\mathcal{X} = \{a_i\}_{i \in \mathbb{N}}$. The joint distribution is described by the probabilities
$$P(X = a_i, Y = a_j), \quad i, j \in \mathbb{N}.$$
The marginal distribution of Y is
$$p_Y(a_j) = P(Y = a_j) = \sum_{i=1}^{\infty} P(X = a_i, Y = a_j).$$
The conditional distribution of X given $Y = a_j$ is
$$p_X(a_i \mid a_j) = P(X = a_i \mid Y = a_j) = \frac{P(X = a_i, Y = a_j)}{P(Y = a_j)},$$
if $P(Y = a_j) > 0$, and 0 otherwise.
For any function $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the conditional expectation given $Y = a_j$ is
$$\mathbb{E}[g(X, Y) \mid Y = a_j] = \sum_{i=1}^{\infty} g(a_i, a_j) p_X(a_i \mid a_j),$$
provided the series converges.

Tower Property

If the expectation of $g(X,Y)$ exists, then

$$\mathbb{E}[\mathbb{E}[g(X,Y)|Y]] = \mathbb{E}[g(X,Y)].$$

3.2 Exercises

Definition: Characteristic function

The characteristic function of a random variable X is

$$\varphi_X(t) = \mathbb{E}[e^{itX}]$$

Distribution	Characteristic function
$\mathcal{N}(\mu,\sigma^2)$	$\varphi_X(t) = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right)$
$\text{Exp}(\lambda)$	$\varphi_X(t) = \frac{\lambda}{\lambda - it}$
$\text{Bern}(p)$	$\varphi_X(t) = 1 - p + pe^{it}$
$\text{Poisson}(\lambda)$	$\varphi_X(t) = \exp\left(\lambda(e^{it}-1)\right)$
$\text{Binomial}(n,p)$	$\varphi_X(t) = (1-p+pe^{it})^n$
$\text{Gamma}(\alpha,\beta)$	$\varphi_X(t) = \left(1 - \frac{it}{\beta}\right)^{-\alpha}$

Density of $Z = X + Y$ (E1)

For independent X, Y and $Z = X + Y$

$$\varphi_Z(t) = \varphi_X(t) \cdot \varphi_Y(t)$$

Transformation of one-dimensional random variables

Let $Y = g(X)$ be a differentiable, strictly monotone transformation of a continuous random variable X. Then the density of Y is

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|.$$

Multivariate transformation of random variables (E3)

Let $(U,V) = g(X,Y)$ with inverse $g^{-1}(u,v) = (x(u,v), y(u,v))$. The joint density of (U,V) is

$$f_{U,V}(u,v) = f_{X,Y}(x(u,v), y(u,v))\left|\det J_{g^{-1}}(u,v)\right|$$

where

$$J_{g^{-1}}(u,v) = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$

CDF of $g(X,Y)$ (E3)

Let X and Y have joint pdf $f_{X,Y}(x,y)$. For a random variable $Z = g(X,Y)$, the cumulative distribution function (CDF) at t is

$$F_Z(t) = \mathbb{P}(Z \le t) = \iint\limits_{\{(x,y):g(x,y)\le t\}} f_{X,Y}(x,y)\,dx\,dy.$$

In other words, the CDF of Z is found by integrating the joint pdf over the region where the function $g(X,Y)$ is at most t.

4 Sufficiency and exponential families

4.1 Sufficiency

Definition: Sufficient statistic

Let $S : \mathcal{X} \to \mathcal{Y}$ be some map. We consider the statistic $S = S(X)$ sufficient if, for all $\theta \in \Theta$, all possible $s = S(X) \in \mathcal{Y}$ and every measurable set $A \subseteq \mathcal{X}^n$ the conditional distribution

$$P_\theta\left(X \in A | S(X) = s\right).$$

does not depend on θ.

Intuitively, we can reduce the data X to S without losing any information about the model parameter θ.

4.2 Factorisation Theorem of Neyman

Factorisation Theorem of Neyman

Suppose that each element of $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure ν. Let $p_\theta = \frac{dP_\theta}{d\nu}$ denote the densities. Then, S is sufficient if and only if one can write

$$p_\theta(x) = g_\theta(S(x))\,h(x) \quad \text{for all } x \text{ and } \theta,$$

for some functions $g_\theta(\cdot)$ on $\mathcal{Y}$ and $h(\cdot)$ on $\mathcal{X}$. The functions g_θ and h can be chosen to be non-negative.

Moreover, if there is a sufficient statistic S for θ, and the MLE exists, it only depends on the sufficient statistic $S = S(X)$ and is given by

$$\hat{\theta} \in \arg\max_\theta L_X(\theta) = \arg\max_\theta g_\theta(S).$$

4.3 Exponential families

k-dimensional exponential family

A *k-dimensional exponential family* (where k is the dimensionality of Θ) is a class of probability distributions whose densities can be expressed in the form

$$p_\theta(x) = \exp\left[\sum_{j=1}^{k} c_j(\theta)T_j(x) - d(\theta)\right]h(x).$$

Where

- $S(x) = (T_1(X), \ldots, T_k(X))$ is a *k-dimensional sufficient statistic*,
- $c_j(\theta)$ is a *natural parameter*,
- $d(\theta)$ is the *log-partition function* ensuring that the density integrates (or sums) to 1,
- and $h(x)$ is the *base measure*, independent of θ.

Distribution of $\mathbf{X} = (X_1, \ldots, X_n)$

If $X_1, \ldots, X_n$ is an iid sample of a k-dimensional exponential family, then the density of $\mathbf{X}$ is

$$\prod_{i=1}^{n} p_\theta(x_i) = \exp\left[\sum_{j=1}^{k} c_j(\theta)\bar{T}_j - nd(\theta)\right] \prod_{i=1}^{n} h(x_i),$$

Where $\bar{T}_j(x) = \sum_{i=1}^{n} T_j(x_i)$

4.4 Canonical form of an exponential family

A k-dimensional exponential family is in canonical form if $c_j(\theta) = \theta_j$. Assuming necessary derivatives exist, denote

$$\dot{d}(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} d(\theta), \quad \ddot{d}(\theta) = \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} d(\theta)\right)_{ij},$$

and write $T(x) = (T_1(x), \ldots, T_k(x))'$, $x \in \mathcal{X}$.

Under regularity assumptions it holds that

$$\mathbb{E}_\theta[T(X)] = \dot{d}(\theta), \quad \mathrm{Cov}_\theta(T(X)) = \ddot{d}(\theta),$$

and in the one-dimensional case,

$$\mathrm{Var}_\theta(T(X)) = \ddot{d}(\theta).$$

5 Bias, variance, and the Cramér-Rao lower bound

5.1 Unbiased estimators

Unbiased Estimator

An estimator T of $g(\theta)$ is unbiased if

$$\mathbb{E}_\theta[T] = g(\theta) \quad \forall\theta \in \Theta.$$

5.2 UMVU estimators

Definition: UMVU estimator

An unbiased estimator T^* is *Uniformly Minimum Variance Unbiased (UMVU)* if

$$\mathrm{Var}_\vartheta(T^*) \leq \mathrm{Var}_\vartheta(T) \quad \forall\vartheta \in \Theta$$

for any other unbiased estimator T.

If a UMVU estimator exists, then it is unique.

Complete statistics

A statistics S is called complete if, for any measurable function h (such that $h(S)$ is integrable with respect to all P_θ)

$$\mathbb{E}_\theta[h(S)] = 0 \quad \forall\theta \in \Theta \implies h(S) = 0, \quad P_\theta\text{-a.s. } \forall\theta \in \Theta.$$

Theorem: Lehmann-Scheffé

Let T be an unbiased estimator of $g(\vartheta)$ with finite variance, and let S be a sufficient and complete statistic. Then

$$T^* := \mathbb{E}[T \mid S]$$

is UMVU.

A consequence of the Lehmann-Scheffé theorem is the following: Let S be a complete and sufficient statistic for ϑ. Then any estimator of the form

$$T^* = c \cdot S,$$

where c is a non-random constant chosen such that T^* is unbiased for $g(\vartheta)$, is UMVU.

Completeness in exponential families

Suppose we have a k-dimensional exponential family. Define

$$\mathcal{C} := \{(c_1(\theta), \ldots, c_k(\theta)) : \theta \in \Theta\} \subseteq R^k.$$

If $\mathcal{C}$ contains an open ball in $\mathbb{R}^k$, then $S := (T_1, \ldots, T_k)$ is complete.

5.3 The Cramér-Rao lower bound

We define the score function as

$$s_\theta(x) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(x) = \frac{\dot{p}_\theta(x)}{p_\theta(x)},$$

And the Fisher information as

$$I(\theta) = \mathbb{E}_\theta\left[s_\theta(X)^2\right] = \mathrm{Var}(s_\theta(X)),$$

as $\mathbb{E}_\theta\left[s_\theta(X)\right] = 0$.

Cramér-Rao lower bound

The Cramér-Rao lower bound provides a lower bound on the variance of any unbiased estimator T of a parameter $g(\theta)$. is differentiable with derivative

$$\dot{q}(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} q(\theta) = \mathrm{Cov}(T, s_\theta(X)).$$

Moreover, if $I(\theta) > 0$,

$$\mathrm{Var}_\theta(T) \geq \frac{(\dot{q}(\theta))^2}{I(\theta)}.$$

Fisher information for independent samples

Suppose $X_1, \ldots, X_n$ are i.i.d. with density p_θ, differentiable in θ, and let s_θ denote the score function. The joint density of $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{n} p_\theta(x_i), \quad \mathbf{x} = (x_1, \ldots, x_n).$$

The joint score function satisfies

$$s_\theta(\mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(\mathbf{x}) = \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(x_i) = \sum_{i=1}^{n} s_\theta(x_i),$$

and the Fisher information of $\mathbf{X}$ is additive:

$$I_n(\theta) = \mathrm{Var}_\theta\big(s_\theta(\mathbf{X})\big) = \sum_{i=1}^{n} \mathrm{Var}_\theta\big(s_\theta(X_i)\big) = nI(\theta),$$

where $I(\theta)$ is the Fisher information of a single observation.

5.4 The CRLB and exponential families

Suppose T is an unbiased estimator of $g(\theta)$ with finite positive variance and attains the CRLB. Then $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ is a 1-dimensional exponential family. Moreover, $c(\cdot)$ and $d(\cdot)$ are differentiable and

$$g(\theta) = \frac{\dot{d}(\theta)}{\dot{c}(\theta)} = \mathbb{E}_\theta\left[T\right], \quad \theta \in \Theta.$$

The score function is

$$s_\theta(x) = \dot{c}(\theta)T(x) - \dot{d}(\theta),$$

The Fisher information can be written as

$$I(\theta) = \mathrm{Var}_\theta(s_\theta(X)) = \dot{c}(\theta)^2 \mathrm{Var}_\theta(T(X)) = \dot{c}(\theta)\dot{g}(\theta).$$

Since T is unbiased for $g(\theta)$, its variance attains the Cramér-Rao lower bound:

$$\mathrm{Var}_\theta(T(X)) = \frac{\dot{g}(\theta)^2}{I(\theta)} = \frac{\dot{g}(\theta)}{\dot{c}(\theta)} = \frac{I(\theta)}{(\dot{c}(\theta))^2}$$

6 Tests and confidence intervals

6.1 Constructing tests

Definition 6.1: Randomized test

A *(randomized) test* is a statistic $\phi : \mathcal{X} \to [0,1]$. If $\phi = \phi(X) \in \{0,1\}$, the test is *non-randomized*.
Let $\Theta_0 \subseteq \Theta$ and $\alpha \in (0,1)$. The test ϕ is a *test at level* α for the *(null) hypothesis*

$$H_0 : \theta \in \Theta_0$$

if

$$\sup_{\theta \in \Theta_0} E_\theta \phi(X) \leq \alpha.$$

If ϕ is non-randomized, then $E_\theta \phi(X) = P_\theta(\phi(X) = 1)$. We reject H_0 when $\phi(X) = 1$, controlling the error probability at level α.

For a randomized test with $\phi(X) = q \in [0,1]$, we reject H_0 by flipping a coin with success probability q. Thus, controlling the test level ensures the average rejection probability under H_0 is at most α.

7 Useful

7.1 Formulas

Convolution of independent discrete random variables

CHANGE TO CONTINUOUS CASE For independent discrete random variables X, Y with pmfs p_X, p_Y

$$p_{X+Y}(z) = \sum_y p_X(z-y) p_Y(y).$$

7.2 Tables

Distribution	Density / PMF
$\mathcal{N}(\mu, \sigma^2)$	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$
$\mathrm{Exp}(\lambda)$	$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{x \geq 0\}}$
$\mathrm{Bern}(p)$	$f_X(x) = p^x (1-p)^{1-x}, \quad x \in \{0,1\}$
$\mathrm{Poisson}(\lambda)$	$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$
$\mathrm{Binomial}(n, p)$	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$
$\mathrm{Gamma}(\alpha, \beta)$	$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{\{x \geq 0\}}$

7.3 Distributions

Poisson distribution $\mathrm{Pois}(\lambda)$

Intuition: Models the number of events occurring in a fixed interval of time or space, assuming events happen independently at a constant average rate.
Parameters: $\lambda \in (0, \infty)$ (rate)
Support: $k \in \mathbb{N}_0$
PMF:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

CDF: Not used, too complicated
Mean: $\mathbb{E}[X] = \lambda$
Variance: $\mathrm{Var}(X) = \lambda$
Exponential family form:

$$p_\lambda(k) = \exp\left(k \log \lambda - \lambda - \log(k!)\right), \quad k \in \mathbb{N}_0.$$

Characteristic function:

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \exp\left[\lambda(e^{it} - 1)\right].$$

Fisher information:

$$I(\lambda) = \frac{1}{\lambda}.$$

Properties: If $X \sim \mathrm{Pois}(\theta)$ and $Y \sim \mathrm{Pois}(\lambda)$ are independent, then

$$X + Y \sim \mathrm{Pois}(\theta + \lambda).$$

Exponential distribution $\mathrm{Exp}(\lambda)$

Intuition: Models the waiting time until the first event in a Poisson process with rate λ, i.e., time between independent events occurring at a constant average rate.
Parameters: $\lambda \in (0, \infty)$ (rate)
Support: $x \in [0, \infty)$
PDF:

$$f(x) = \lambda e^{-\lambda x}.$$

CDF:

$$F(x) = 1 - e^{-\lambda x}.$$

Mean: $\mathbb{E}[X] = \frac{1}{\lambda}$
Variance: $\mathrm{Var}(X) = \frac{1}{\lambda^2}$
Exponential family form:

$$f_\lambda(x) = \exp\left(\log \lambda - \lambda x\right), \quad x \geq 0.$$

Characteristic function:

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \frac{\lambda}{\lambda - it}, \quad t \in \mathbb{R}.$$

Fisher information:

$$I(\lambda) = \frac{1}{\lambda^2}.$$

Properties:

- Memoryless property: For $s, t \geq 0$,

$$P(X > s + t \mid X > s) = P(X > t).$$

- The sum of n independent $\mathrm{Exp}(\lambda)$ variables is $\mathrm{Gamma}(n, \lambda)$ distributed.

ADD OTHER DISTRIBUTIONS