Fundamentals of Mathematical Statistics

Daniele Cambria

January 29, 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 typically indexed by parameters $\vartheta \in \Theta$, so that
$$\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}.$$
If the model is *correctly specified*, there exists a true parameter value $\theta \in \Theta$ such that the data are actually generated according to P_θ.
- 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.

Given data $X_1, \ldots, X_n$, the empirical distribution is
$$\widehat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$
(where δ_{X_i} is the Dirac measure at X_i). A plug-in estimator for a parameter $\gamma = Q(P_\theta)$ is obtained by evaluating the functional Q at $\widehat{P}_n$:
$$\hat{\gamma}_n := Q(\widehat{P}_n)$$

Method of Moments Estimators

Let $X \in \mathbb{R}$ and $P = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$. Suppose the first p moments exist:
$$\mu_j(\theta) = \int x^j dF_\theta(x), \quad j = 1, \ldots, p.$$
Assume the moment map $m(\theta) = (\mu_1(\theta), \ldots, \mu_p(\theta))$ is injective with inverse m^{-1} on $M = m(\Theta) \subset \mathbb{R}^p$.
Estimate each moment by its empirical counterpart:
$$\widehat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j = \int x^j d\widehat{F}_n(x), \quad j = 1, \ldots, p.$$
If $(\widehat{\mu}_1, \ldots, \widehat{\mu}_p) \in M$, define the method of moments estimator
$$\widehat{\theta} = m^{-1}(\widehat{\mu}_1, \ldots, \widehat{\mu}_p).$$
In plain words, match the sample moments to the theoretical moments to solve for θ.

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 The mean and covariance matrix of a random vector

Mean and covariance of a random vector

Let $Z = (Z_1, \ldots, Z_k)'$ with $E|Z_i| < \infty$. **Mean:** $EZ = (EZ_1, \ldots, EZ_k)'$ (componentwise).
For a random matrix $M = (M_{ij})$ with $E|M_{ij}| < \infty$, **expectation:** $EM = (EM_{ij})$ (componentwise).
Covariance matrix:

$$\Sigma = \mathrm{Cov}(Z) = E[ZZ'] - (EZ)(EZ)' = (\mathrm{Cov}(Z_i, Z_j))_{ij}$$

Σ is symmetric and positive semi-definite: for any $a \in \mathbb{R}^k$, $a'\Sigma a = \mathrm{Var}(a'Z) \geq 0$.
For deterministic $A \in \mathbb{R}^{\ell\times k}$, $b \in \mathbb{R}^{\ell}$:

$$\mathrm{Cov}(AZ + b) = A\,\mathrm{Cov}(Z)A'$$

3.3 Exercises

Definition: Characteristic function

The characteristic function of a random variable X is

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

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

Density of $Z = X + Y$ (E1): convolution formula

Discrete case: If X and Y are independent discrete random variables with PMFs $p_X(k)$ and $p_Y(l)$, then the PMF of $Z = X + Y$ is

$$p_Z(z) = \sum_k p_X(k) \cdot p_Y(z - k).$$

Continuous case: If X and Y are independent continuous random variables with PDFs $f_X(x)$ and $f_Y(y)$, then the PDF of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(z - x)\, dx.$$

Density of $Z = X + Y$ (E1): characteristic functions

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

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

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

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

Fiber reweighting (tool for proofs)

Let $X = (X_1,\dots,X_n)$ be a discrete random vector and $S = S(X)$ a function of X (e.g. the sum, number of successes, or maximum). For any s in the range of S, the conditional probability on the "fiber" over s is

$$P(X = x \mid S = s) = \begin{cases} \dfrac{P(X=x)}{P(S=s)}, & \text{if } S(x) = s \\ 0, & \text{otherwise} \end{cases}$$

That is, conditioning on $S = s$ restricts to outcomes x with $S(x) = s$.

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

- $S(x) = (T_1(X),\dots,T_k(X))$ 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 $\mathbf{X} = (X_1,\dots,X_n)$

If $X_1,\dots,X_n$ 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 $T(x) = (T_1(x),\dots,T_k(x))'$, $x \in \mathcal{X}$.

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

The *bias* of an estimator $T = T(X)$ is $\mathrm{bias}_\vartheta(T) := \mathbb{E}_\vartheta[T] - g(\vartheta)$. An estimator T of $g(\vartheta)$ is unbiased if

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

The *mean squared error* of T is

$$\mathrm{MSE}_\vartheta(T) = \mathbb{E}_\vartheta\left[(T - g(\vartheta))^2\right] = \left(\mathrm{bias}_\vartheta(T)\right)^2 + \mathrm{Var}_\vartheta(T).$$

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

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

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

Basu's Lemma

Suppose that S is sufficient and complete, and that $Y = Y(X) \in \mathcal{Z}$ has a distribution that does not depend on ϑ, where $\mathcal{Z}$ is the (measurable) outcome space of Y. Then, for all $\vartheta \in \Theta$, S and Y are independent under P_ϑ

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

Cramér-Rao lower bound

The Cramér-Rao lower bound provides a lower bound on the variance of any unbiased estimator T of a parameter $g(\theta)$. Define $q(\vartheta) = \mathbb{E}_\vartheta[T]$, then q is differentiable with derivative

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

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

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

Note, if T is unbiased, then $g(\vartheta) = q(\vartheta)$.

Fisher information for independent samples

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

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

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

If T attains the CRLB (in *any* model):

For ANY 1-dimensional exponential family, the following holds:

$$\mathbb{E}_\vartheta[T(X)] = \frac{\dot{d}(\vartheta)}{\dot{c}(\vartheta)}.$$

Score function:

$$s_\vartheta(x) = \dot{c}(\vartheta)T(x) - \dot{d}(\vartheta).$$

Fisher information:

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

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

Then: The model $\mathcal{P}$ must be a 1-dimensional exponential family, and the following further identities hold:

$$I(\vartheta) = \dot{c}(\vartheta)\dot{g}(\vartheta)$$

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

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

If ϕ is non-randomized, then $E_\theta \phi(X) = P_\theta(\phi(X) = 1)$, and we reject H_0 when $\phi(X) = 1$.

For a randomized test with $\phi(X) = q \in [0,1]$, we reject H_0 by flipping a coin with success probability q.

Definition 6.3: p-value

A statistic $p = p(X) \in [0,1]$ is a *p-value* for H_0 if, for all $\theta \in \Theta_0$,

$$P_\theta(p \le u) \le u, \quad \text{for all } u \in (0,1).$$

Intuition: Under any parameter in H_0, small p-values are rare; seeing a small p suggests evidence against H_0. Given a p-value p for H_0, we can construct a test at level $\alpha \in (0,1)$ by defining

$$\phi(X) = \mathbb{1}\{p \le \alpha\}.$$

Indeed, for any $\theta \in \Theta_0$,

$$P_\theta(\phi = 1) = P_\theta(p \le \alpha) \le \alpha.$$

Let $\Theta_1 \subseteq \Theta$ with $\Theta_0 \cap \Theta_1 = \varnothing$, and consider the alternative hypothesis $H_1 : \theta \in \Theta_1$, that is,

$$H_1 = \{P_\theta : \theta \in \Theta_1\}.$$

If we believe that the true parameter θ can only be in Θ_1 if it is not in Θ_0, we call H_1 our alternative hypothesis.

Definition 6.4: Power of a test

Let ϕ be a test. The power of ϕ against the alternative $\theta \in \Theta_1$ is $E_\theta \phi(X)$.

Definition 6.5: Pivot

A pivot is a function $Z : \mathcal{X} \times \Gamma \to \mathbb{R}$ such that for all $\vartheta \in \Theta$, the distribution of $Z(X, g(\vartheta)) = Z(X, \gamma)$ under P_ϑ does not depend on ϑ:

$$P_\vartheta(Z(X, g(\vartheta)) \le z) = G(z)$$

for some cdf G independent of ϑ.

To test $H_0 : \gamma = \gamma_0$ using a pivot $Z(X, \gamma_0)$, reject H_0 if $Z(X, \gamma_0)$ is outside the interval $[q_L, q_R]$, where q_L and q_R are the appropriate $\alpha/2$ and $1 - \alpha/2$ quantiles of G:

$$\phi(X, \gamma_0) := \mathbb{1}\{Z(X, \gamma_0) \notin [q_L, q_R]\}$$

This gives a level α test.

Two-sided p-value based on the pivot:

$$p = 2\min\{G(Z(X, \gamma_0)),\ 1 - G(Z(X, \gamma_0)^-)\}$$

For a one-sided alternative, reject for large values of Z:

$$\phi(X, \gamma_0) := \mathbb{1}\{Z(X, \gamma_0) > q_G(1 - \alpha)\}$$

One-sided p-value:

$$p = 1 - G(Z(X, \gamma_0)^-)$$

6.2 Duality of confidence sets and tests

Definition 6.9: Confidence set

A random set $I = I(X) \subseteq \Gamma$, depending only on the data X, is a confidence set for γ at level $1 - \alpha$ if, for all $\vartheta \in \Theta$ and $\gamma = g(\vartheta)$,

$$P_\vartheta(\gamma \in I(X)) \ge 1 - \alpha.$$

If $I = [\underline{\gamma}(X), \overline{\gamma}(X)]$, it is a confidence interval.

Duality: A confidence set $I(X)$ induces a family of tests—one for each $\gamma_0 \in \Gamma$:

$$\phi(X, \gamma_0) := \mathbb{1}\{\gamma_0 \notin I(X)\}$$

is a level-α test for $H_0 : \gamma = \gamma_0$.

Conversely, given a family of level-α tests $\{\phi(X, \gamma_0)\}_{\gamma_0 \in \Gamma}$ for $H_0 : \gamma = \gamma_0$, define

$$I(X) := \{\gamma \in \Gamma : \phi(X, \gamma) = 0\}$$

which is a $(1 - \alpha)$ confidence set for γ.

Good confidence sets have short intervals for fixed coverage. *Good tests* have high power for alternatives.

6.3 The two-sample problem

Consider independent samples $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$ from F_X, F_Y. Test

$$H_0 : F_X = F_Y$$

against a one- or two-sided alternative.

6.3.1 Student's t-test

Assume F_X and F_Y are normal with equal variance σ^2:

$$F_X = \Phi\left(\frac{\cdot - \mu}{\sigma}\right), \quad F_Y = \Phi\left(\frac{\cdot - (\mu + \gamma)}{\sigma}\right)$$

γ is the difference in means (*effect*). Parameter of interest: $g(\theta) = \gamma$.

For $H_0 : \gamma = 0$, the test statistic is given by

$$S^2 = \frac{1}{n + m - 2}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right]$$

$$T = \sqrt{\frac{nm}{n+m}}\,\frac{\bar{Y} - \bar{X}}{S}$$

Under H_0, $T \sim t_{n+m-2}$.

For a one-sided test, $H_1 : \gamma < 0$

$$\phi(\mathbf{X}, \mathbf{Y}) = \mathbb{1}\{T < t_{n+m-2;\alpha}\}$$

Critical value $t_{n+m-2;\alpha}$: α-quantile of t-distribution with $n + m - 2$ df, and p-value:

$$p = t_{n+m-2}(T)$$

where $t_{n+m-2}(\cdot)$ is t CDF with $n + m - 2$ df.

6.3.2 Wilcoxon's rank-sum test

Test $H_0 : F_X = F_Y$ for F_X, F_Y continuous (no ties).

Combine both samples: $N = n + m$,

$$Z_i = \begin{cases} X_i, & i = 1, \dots, n \\ Y_{i-n}, & i = n+1, \dots, N \end{cases}$$

Let $R_i = \text{rank}(Z_i)$ among all N values. Wilcoxon rank-sum statistic:

$$T = \sum_{i=1}^{n} R_i$$

Alternatively,

$$T = \#\{(i,j) : Y_j < X_i\} + \frac{n(n+1)}{2}$$

Under H_0, the distribution of T is $P_{H_0}(T = t) =$

$$\frac{1}{N!} \# \left\{ r : r \text{ is a permutation of } \{1, \dots, N\}, \sum_{i=1}^{n} r_i = t \right\}$$

Distribution does not depend on F_X or F_Y.

6.3.3 Comparison: t-test vs. Wilcoxon's test

T-test assumes normality, Wilcoxon test only continuity (for ranks).

Wilcoxon's test is nearly as efficient as t-test under normality: relative asymptotic efficiency ≈ 0.955 for Normal distribution, but, if the normality assumption fails, Wilcoxon's test can be much more efficient.

6.4 The lemma of Neyman and Pearson

UMP Tests

A test ϕ is called UMP (uniformly most powerful) at level α if
- ϕ has level α for H_0,
- For any other level α test $\tilde{\phi}$ and all $\theta \in \Theta_1$, $E_\theta \tilde{\phi}(X) \le E_\theta \phi(X)$.

For *simple* hypotheses: $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, let p_i be the density of P_{θ_i}, $i = 0, 1$, with respect to dominating measure ν. The Neyman–Pearson (NP) test is

$$\phi_{NP}(x) = \begin{cases} 1, & \text{if } \frac{p_1(x)}{p_0(x)} > c \\ q, & \text{if } \frac{p_1(x)}{p_0(x)} = c \\ 0, & \text{if } \frac{p_1(x)}{p_0(x)} < c \end{cases} \quad q \in [0,1],\ c \ge 0$$

Neyman-Pearson Lemma:

For any test ϕ and the NP test ϕ_{NP},

$$E_{\theta_1}\phi - E_{\theta_1}\phi_{NP} \le c\,[E_{\theta_0}\phi - E_{\theta_0}\phi_{NP}]$$

with equality $\iff$
$\nu(p_1 > cp_0,\ \phi < 1) = 0 = \nu(p_1 < cp_0,\ \phi > 0)$.

Implications:

- For every level-α test, there exist constants c_α, q_α such that the NP test ϕ^*_{NP} with these constants attains $E_{\theta_0}\phi^*_{NP} = \alpha$ and maximizes power $E_{\theta_1}\phi$.
- If $P_{\theta_1} \ne P_{\theta_0}$, then $E_{\theta_1}\phi^*_{NP} > \alpha$.
- The UMP NP test is unique up to ν-null sets, except possibly on $\{p_1 = c_\alpha p_0\}$.

6.5 UMP tests for exponential families

Consider the one-sided testing problem for a **one-dimensional** exponential family:

$$H_0 : \theta \le \theta_0 \qquad \text{vs.} \qquad H_1 : \theta > \theta_0$$

with $c(\theta)$ **strictly increasing.**

UMP Test Structure

For level $\alpha \in (0,1)$, the UMP test is

$$\phi^*(T(X)) = \begin{cases} 1, & T(X) > t_\alpha \\ q_\alpha, & T(X) = t_\alpha \\ 0, & T(X) < t_\alpha \end{cases}$$

where t_α is the $(1-\alpha)$-quantile of $T(X)$ under P_{θ_0}, and

$$q_\alpha = \begin{cases} \frac{\alpha - P_{\theta_0}(T(X) > t_\alpha)}{P_{\theta_0}(T(X) = t_\alpha)}, & P_{\theta_0}(T(X) = t_\alpha) > 0 \\ 1, & P_{\theta_0}(T(X) = t_\alpha) = 0 \end{cases}$$

This test has exact level α.

No UMP test exists for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \ne \theta_0$ in general.

6.6 Unbiased tests

Unbiased and UMPU tests

- A test ϕ is *unbiased* if for all $\theta \in \Theta_0$ and all $\theta' \in \Theta_1$:
$$E_\theta \phi(X) \le E_{\theta'} \phi(X)$$
- ϕ is *uniformly most powerful unbiased (UMPU)* at level α if:
  - Level α for H_0,
  - Unbiased,
  - For all unbiased level α tests $\tilde{\phi}$ and all $\theta \in \Theta_1$:
$$E_\theta \tilde{\phi}(X) \le E_\theta \phi(X)$$

UMPU test in exponential families (two-sided case)

For a one-dimensional exponential family with strictly increasing $c(\theta)$, to test

$$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \ne \theta_0$$

the UMPU test has the form

$$\phi(T(x)) = \begin{cases} 1, & T(x) < t_L \text{ or } T(x) > t_R \\ q_L, & T(x) = t_L \\ q_R, & T(x) = t_R \\ 0, & t_L < T(x) < t_R \end{cases}$$

with $t_L, t_R \in \mathbb{R}$ and $q_L, q_R \in [0,1]$ chosen such that

$$E_{\theta_0}\phi(T(X)) = \alpha \qquad \text{and} \qquad \frac{d}{d\theta} E_\theta \phi(T(X))\Big|_{\theta=\theta_0} = 0$$

7 Equivariant estimators

Given an estimator $T(X)$ for parameter $\gamma = g(\theta)$, a *loss function* $L : \Theta \times \Gamma \to \mathbb{R}$ quantifies the quality of an estimate $a \in \Gamma$ when the true parameter is θ. The *risk* of an estimator $T(X)$ is the expected loss:

$$R(\theta, T) = E_\theta\big[\, L(\theta, T(X))\,\big]$$

In this chapter, we focus on the location model:

$$X_i = \theta + \epsilon_i, \quad i = 1, \dots, n$$

where θ is unknown and $(\epsilon_1, \ldots, \epsilon_n)$ are iid with known distribution.

Equivariance and Invariance

Location equivariant: T satisfies $T(x_1 + c, \ldots, x_n + c) = T(x_1, \ldots, x_n) + c$ for all $c \in \mathbb{R}, x \in \mathbb{R}^n$.
Location invariant loss: $L(\theta + c, a + c) = L(\theta, a)$ for all $c \in \mathbb{R}$.

Risk for Equivariant Estimator and Invariant Loss

If T is equivariant and L is invariant, then

$$R(\theta, T) = E_0 L(0, T(\epsilon)) \qquad \text{for all } \theta,$$

where $\epsilon = X - \theta$; the risk does not depend on θ.

UMRE Estimator (Uniform Minimum Risk Equivariant)

T is UMRE if it is equivariant and

$$R(\theta, T) = \min_{\text{equivariant } d} R(\theta, d) \quad \text{for all } \theta.$$

For invariant loss, this is equivalent to minimizing $R(0, T)$ over all equivariant estimators.

Construction of UMRE estimator

Given equivariant d, define

$$\tilde{T}(z) = \arg\min_{v \in \mathbb{R}} E_0[L(0, v + d(\epsilon)) \mid \epsilon - d(\epsilon) = z]$$

Then $T^*(X) = \tilde{T}(X - d(X)) + d(X)$ is UMRE.

Pitman estimator (quadratic loss)

If $L(\theta, a) = (\theta - a)^2$ and $d(X) = X_n$, then

$$T^*(X) = X_n - E_0[\epsilon_n \mid X - X_n].$$

8 Decision Theory

8.1 Decisions and their risk

Let A be the action space. A decision is a function $d : \mathcal{X} \to A$; after observing X, we choose $d(X)$. The loss function: $L : \Theta \times A \to \mathbb{R}$ gives the loss $L(\theta, a)$ for parameter θ and action a. The risk of d is

$$R(\theta, d) := E_\theta L(\theta, d(X)).$$

If $A = \mathbb{R}$, d is often called an estimator (T); if $A = \{0, 1\}$ or $A = [0, 1]$, a test (ϕ).

Risk decomposition for quadratic loss

For any estimator $d(X)$ of the parameter θ under quadratic loss, where $L(\theta, a) = (\theta - a)^2$ and assuming $\mathbb{E}_\theta[d(X)^2] < \infty$, the risk decomposes as

$$R(\theta, d) = \mathbb{E}_\theta\left[\left(\theta - d(X)\right)^2\right] = \mathrm{Var}_\theta\left(d(X)\right) + \left[\mathrm{Bias}_\theta\left(d(X)\right)\right]^2,$$

where

$$\mathrm{Var}_\theta\left(d(X)\right) = \mathbb{E}_\theta\left[\left(d(X) - \mathbb{E}_\theta d(X)\right)^2\right],$$
$$\mathrm{Bias}_\theta\left(d(X)\right) = \mathbb{E}_\theta d(X) - \theta.$$

8.2 Risk and sufficiency

Suppose $S = S(X)$ is a sufficient statistic.

A *randomised decision* is a map $\delta : X \to M_1(A)$, where $M_1(A)$ denotes the probability measures on the action space A. The risk of a randomised decision is

$$R(\theta, \delta) = E_\theta\left[\int_A L(\theta, a)\, d\delta(X)(a)\right]$$

Upon observing X, a random sample is drawn from $\delta(X)$ and used as the decision.

Sufficiency preserves risk

If $d : X \to A$ is any decision, there exists a randomised decision $\delta(S)$, depending only on the sufficient statistic S, such that

$$R(\theta, \delta) = R(\theta, d) \qquad \text{for all } \theta$$

If the risk is convex in the action, further improvement is possible:

Rao-Blackwell theorem (risk reduction by conditioning)

Let A be convex and $a \mapsto L(\theta, a)$ convex for all θ. For any decision $d(X)$, define $d'(S) = E_\theta\big[d(X) \mid S\big]$. Then,

$$R(\theta, d') \le R(\theta, d) \qquad \text{for all } \theta$$

and d' does not depend on θ.

Special case (mean squared error): For any estimator $T(X)$ with sufficient statistic S

$$R(\theta, T') = \mathrm{MSE}_\theta(T') \le \mathrm{MSE}_\theta(T), \quad \text{where } T'(S) = E_\theta[T \mid S]$$

Variance is reduced by Rao-Blackwellisation, while the bias is preserved.

8.3 Admissible decisions

Admissible decision

A decision d is *inadmissible* if there exists another d' such that $R(\theta, d') \le R(\theta, d)$ for all θ and strict inequality for some θ. Otherwise, d is *admissible*.

Neyman-Pearson test admissibility

A Neyman-Pearson test φ_{NP} is admissible *if and only if* one of the following holds:

- its power is strictly less than one ($E_{\theta_1}\varphi_{NP}(X) < 1$), or
- it has power one *and* minimal level among all tests of power one.

8.4 Minimax decisions

Minimax decision

A decision d is *minimax* if

$$\sup_{\theta \in \Theta} R(\theta, d) = \inf_{d' \text{ decision}} \sup_{\theta \in \Theta} R(\theta, d')$$

That is, d achieves the lowest possible worst-case risk over θ among all decisions.

Neyman-Pearson test minimaxity

A Neyman-Pearson test φ_{NP} is minimax *if and only if*

$$R(\theta_0, \varphi_{NP}) = R(\theta_1, \varphi_{NP})$$

8.5 Bayes decisions

Given a prior Π on Θ, the *Bayes risk* of a decision d is

$$r(\Pi, d) = \int_\Theta R(\vartheta, d)\, d\Pi(\vartheta).$$

A *Bayes decision* (w.r.t. Π) minimizes this risk among all decisions.

If Π has density w, then

$$r(\Pi, d) = \int_\Theta R(\vartheta, d)\, w(\vartheta)\, d\mu(\vartheta).$$

Intuition: The Bayes risk is a weighted average of risks over Θ under the prior belief about θ.

Bayes test (simple vs. simple): For $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, with $w_0 = \Pi(\{\theta_0\})$, $w_1 = \Pi(\{\theta_1\}) = 1 - w_0$, the Bayes test φ_{Bayes} is:

$$\varphi_{\text{Bayes}}(x) = \begin{cases} 1, & \text{if } \dfrac{p_1(x)}{p_0(x)} > \dfrac{w_0}{w_1} \\ q, & \text{if } \dfrac{p_1(x)}{p_0(x)} = \dfrac{w_0}{w_1} \\ 0, & \text{if } \dfrac{p_1(x)}{p_0(x)} < \dfrac{w_0}{w_1} \end{cases}$$

Intuition: Take the prior odds into account in the likelihood ratio test; only reject H_0 if the evidence in the data outweighs your prior preference for H_0 over H_1.

Minimum Bayes risk: For $w_0 = w_1 = 1/2$:

$$2r(w, \varphi_{\text{Bayes}}) = 1 - \frac{1}{2} \int |p_0(x) - p_1(x)|\, d\nu(x)$$

Interpretation: The risk is high if the two distributions are similar (hard to distinguish); it is minimized by maximizing the weighted difference between p_0 and p_1.

8.6 Constructing Bayes estimators

Suppose P_θ ($\theta \in \Theta$) are dominated by a measure ν, with densities $p(x \mid \theta)$. The prior Π has density w with respect to μ on Θ.

Posterior density:

$$w(\vartheta \mid x) = \frac{p(x \mid \vartheta) w(\vartheta)}{p(x)} \qquad \text{where } p(x) = \int p(x \mid \vartheta) w(\vartheta) d\mu(\vartheta)$$

Given observed x, define the *posterior expected loss:*

$$\ell(x, a) = \int L(\vartheta, a) w(\vartheta \mid x) d\mu(\vartheta)$$

Bayes estimator

The Bayes estimator is any measurable

$$d_{\text{Bayes}}(x) = \arg\min_a \ell(x, a)$$

If the minimum is unique, this d_{Bayes} minimizes Bayes risk among all estimators.

Interpretation: Minimize expected loss, treating θ as a random variable with posterior $w(\vartheta \mid x)$.

Quadratic loss: If $L(\theta, a) = (\theta - a)^2$,

$$d_{\text{Bayes}}(x) = E[\theta \mid X = x] = \int \vartheta\, w(\vartheta \mid x) d\mu(\vartheta)$$

(*posterior mean*)

MAP estimator: For loss $L(\theta, a) = 1\{|\theta - a| > c\}$ with c small,

$$d_{\text{MAP}}(x) = \arg\max_a w(a \mid x)$$

(*posterior mode*)

If the prior is uniform over Θ (i.e., w constant), then the MAP estimator coincides with the MLE.

Role of sufficiency

If $S = S(X)$ is sufficient, the Bayes estimator $d_{\text{Bayes}}(X)$ depends only on $S(X)$.

Summary: Bayesian estimation reduces to computing the posterior, then minimizing posterior expected loss—giving the posterior mean for quadratic loss, and the posterior mode (MAP) for small error-threshold loss.

Bayes estimator for quadratic loss

Consider the case where the loss is quadratic: $L(\theta, a) = (\theta - a)^2$. Then,

$$d_{\text{Bayes}}(x) = \mathbb{E}[\theta \mid X = x] = \int_\Theta \vartheta\, w(\vartheta \mid x) d\mu(\vartheta).$$

This estimator $T = d_{\text{Bayes}}$ is also called the *posterior mean* (the mean of θ given x).

8.7 Conjugate distributions

Poisson likelihood with Gamma prior

Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(θ) with unknown $\theta > 0$. Assume a Gamma prior $\theta \sim \Gamma(k, \lambda)$, $k, \lambda > 0$, with density

$$w(\vartheta) = \frac{\lambda^k}{\Gamma(k)} \vartheta^{k-1} e^{-\lambda\vartheta}, \quad \vartheta > 0.$$

Given observed sample $x = (x_1, \ldots, x_n)$ with total count $s = \sum_{i=1}^n x_i$, the posterior distribution is

$$\theta \mid x \sim \Gamma\left(k + s,\, \lambda + n\right).$$

Binomial model with Beta prior

Suppose $X \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$, with $\alpha, \beta > 0$. The prior density is

$$w(\vartheta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \vartheta^{\alpha-1}(1 - \vartheta)^{\beta-1}, \quad \vartheta \in (0, 1).$$

Given data $X = x$, the posterior is

$$\theta \mid x \sim \text{Beta}(x + \alpha,\, n - x + \beta).$$

The Bayes estimator under quadratic loss (posterior mean) is

$$\mathbb{E}[\theta \mid X = x] = \frac{x + \alpha}{n + \alpha + \beta}.$$

Normal likelihood with Normal prior

Suppose $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\theta, \sigma^2)$ with known variance σ^2. Assume prior $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Given data $x = (x_1, \dots, x_n)$ and sample mean $\bar{x} = \frac{1}{n}\sum x_i$, the posterior distribution is

$$\theta \mid x \sim \mathcal{N}\left(\frac{\frac{n}{\sigma^2}\bar{x} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}, \quad \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \right).$$

Posterior distribution for exponential family with conjugate prior

Suppose the data $X = (X_1, \dots, X_n)$ are i.i.d. with density in an exponential family form

$$p(x \mid \theta) = h(x) \exp\left(\eta(\theta) \cdot T(x) - A(\theta)\right),$$

where $\eta(\theta)$ is the natural parameter, $T(x)$ is the sufficient statistic, and $A(\theta)$ the log-partition function. Assume the prior on θ is conjugate, having density

$$w(\theta) \propto \exp\left(\eta(\theta) \cdot \nu - \tau A(\theta)\right),$$

with hyperparameters ν and τ.
Given sample x, the posterior distribution is also conjugate with updated parameters

$$w(\theta \mid x) \propto \exp\left(\eta(\theta) \cdot (\nu + T(x)) - (\tau + n)A(\theta)\right).$$

Thus, the posterior belongs to the same family with parameters $\nu + T(x)$ and $\tau + n$.

9 Proving admissibility and minimaxity

In unbounded parameter spaces, the following extended notion is useful:

Definition: extended Bayes estimator

T is called *extended Bayes* if there exists a sequence of priors $(w_m)_{m\in\mathbb{N}}$ such that

$$r(w_m, T) - \inf_{T'} r(w_m, T') \to 0 \quad \text{as } m \to \infty,$$

where $r(w_m, T)$ is the Bayes risk under prior w_m.

9.1 Minimaxity

Proposition: Criteria for minimaxity

Suppose T is a statistic with constant risk $R(\theta, T) = R(T)$ independent of θ. Then:

1. Admissible $T \implies T$ is minimax.
2. Bayes $T \implies T$ is minimax.
3. Extended Bayes $T \implies T$ is minimax.

The Pitman estimator is extended Bayes for quadratic loss, thus is minimax.

9.2 Admissibility

Bayes admissibility conditions

Suppose T is Bayes for the prior w on Θ (an open set), with $R(\theta, T) < \infty$ for all T. Either of the following is sufficient for admissibility of T:

1. T is the unique Bayes estimator: $r(w, T) = r(w, T')$ implies $T = T'$ P_θ-almost surely for all $\theta \in \Theta$.
2. For all estimators T', $R(\theta, T')$ is continuous in θ, and for every open set $U \subseteq \Theta$, $\Pi(U) = \int_U w(\vartheta)\, d\mu(\vartheta) > 0$.

Extended Bayes admissibility

Suppose T is extended Bayes, and assume for all estimators T', $R(\theta, T')$ is continuous in θ, and every open $U \subseteq \Theta$ satisfies $\Pi_m(U) := \int_U w_m(\vartheta) d\mu_m(\vartheta) > 0$ for all m. If also

$$\frac{r(w_m, T) - \inf_{T'} r(w_m, T')}{\Pi_m(U)} \to 0$$

as $m \to \infty$ for all open U, then T is admissible.

Admissible estimators for the normal mean

Let $X \sim N(\theta, 1)$, $\theta \in \mathbb{R}$, quadratic loss. For estimators of the form $T = aX + b$ with $a > 0$, $b \in \mathbb{R}$:

1. T is admissible if and only if $a < 1$,
2. or $a = 1$ and $b = 0$.

9.3 Admissible estimators in exponential families

Admissibility in canonical exponential families

Let $\Theta = \mathbb{R}$, and suppose $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ is an exponential family in **canonical** form with d twice differentiable. For quadratic loss $L(a, \vartheta) = |a - \dot{d}(\vartheta)|^2$, the estimator $T = T(X)$ is admissible for $g(\vartheta) = \dot{d}(\vartheta)$.

Caution: parameter space matters!

If the canonical parameter space is not $\mathbb{R}$, admissibility can fail.
Example: $X \sim N(0, \sigma^2)$, $\sigma^2 \in (0, \infty)$ unknown, written as $\theta = 1/\sigma^2$. Here, $T(x) = -x^2/2$ is NOT admissible; the risk is minimized for $T_a(x) = -ax^2$ with $a = 1/6$.

10 Asymptotic Theory

10.1 Notions of convergence

Convergence in probability

Z_n converges in probability to Z if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\big(\|Z_n - Z\| > \epsilon\big) = 0.$$

Notation: $Z_n \xrightarrow{P} Z$.

Chebyshev's inequality

For any increasing function $\psi : [0,\infty) \to [0,\infty)$, $\epsilon > 0$, and random variable Z,

$$P(\|Z\| > \epsilon) \le \frac{\mathbb{E}\psi(\|Z\|)}{\psi(\epsilon)}.$$

Convergence in distribution

Z_n converges in distribution to Z if for all continuous and bounded functions f,

$$\lim_{n\to\infty} \mathbb{E}f(Z_n) = \mathbb{E}f(Z).$$

Notation: $Z_n \xrightarrow{D} Z$.

Convergence in distribution only depends on the distributions of Z_n and Z, often written as $Q_n \xrightarrow{w} Q$ (where w stands

Equivalence with cdf convergence

$Z_n \xrightarrow{D} Z \iff F_n(x) \to F(x)$ at all x where F is continuous, where F_n and F are the cdfs of Z_n and Z.

Convergence in probability implies convergence in distribution, but not vice versa.

Central limit theorem (CLT)

Let $(X_n)_{n\in\mathbb{N}}$ be iid random variables with mean μ and variance $\sigma^2 < \infty$. Then,

$$\sqrt{n}\Big(\overline{X}_n - \mu\Big) \xrightarrow{D} N(0,\sigma^2), \quad n \to \infty,$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Cramér–Wold device

For Z_n, Z $\mathbb{R}^p$-valued random variables,

$$Z_n \xrightarrow{D} Z \iff a'Z_n \xrightarrow{D} a'Z, \text{ for all } a \in \mathbb{R}^p.$$

Portmanteau Theorem

$Z_n \xrightarrow{D} Z$ (with Z distribution Q and cdf G) if and only if the following equivalent conditions hold:

1. $Z_n \xrightarrow{D} Z$
2. $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z)$ for all bounded Lipschitz functions f
3. $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z)$ for all bounded, Q-a.e. continuous f
4. $P(Z_n \le z) \to G(z)$ for all z at which G is continuous

Continuous mapping theorem

If $Z_n \xrightarrow{D} Z$ and $f : \mathbb{R}^p \to \mathbb{R}^q$ is Q-a.e. continuous, then $f(Z_n) \xrightarrow{D} f(Z)$.

The continuous mapping theorem also holds for convergence in probability: If $Z_n \xrightarrow{P} Z$ then $f(Z_n) \xrightarrow{P} f(Z)$ for Q-a.e. continuous f.

Stochastic order notation

- $Z_n = O_P(1)$ if $\lim_{M\to\infty} \limsup_{n\to\infty} P(\|Z_n\| > M) = 0$ (bounded in probability).
- $Z_n = O_P(r_n)$ if $Z_n/r_n = O_P(1)$.
- $Z_n = o_P(1)$ if $Z_n \xrightarrow{P} 0$.
- $Z_n = o_P(r_n)$ if $Z_n/r_n \xrightarrow{P} 0$.

- If $Z_n \xrightarrow{D} Z$, then $Z_n = O_P(1)$ (bounded in probability).
- If $Z_n = O_P(f_n)$ and $f_n = o(g_n)$, then $Z_n = o_P(g_n)$.
- If $\|Z_n\|^\alpha = O_P(r_n)$ and $g(x) = o(\|x\|^\alpha)$ as $x \to 0$, then $g(Z_n) = o_P(r_n)$.

Slutsky's theorem

If $Z_n \xrightarrow{D} Z$ and $A_n \xrightarrow{P} a$, then $A_n'Z_n \xrightarrow{D} a'Z$.

10.2 Consistency and asymptotic normality

Symmetry of estimators

For iid data, it is natural to assume each T_n is symmetric in its arguments:

$$T_n(X_1,\dots,X_n) = T_n(X_{\pi(1)},\dots,X_{\pi(n)})$$

for any permutation π.

Consistency

A sequence of estimators $(T_n)_{n\in\mathbb{N}} \subset \mathbb{R}^p$ is called *consistent* for $\gamma = g(\vartheta)$ if, for all $\vartheta \in \Theta$,

$$T_n \xrightarrow{P_\vartheta} \gamma = g(\vartheta), \quad n \to \infty.$$

Asymptotic normality

$(T_n)_{n\in\mathbb{N}} \subset \mathbb{R}^p$ is *asymptotically normal* for $\gamma = g(\vartheta)$ with covariance $V_\vartheta \in \mathbb{R}^{p\times p}$ if, for all $\vartheta \in \Theta$,

$$\sqrt{n}(T_n - \gamma) \xrightarrow{D_\vartheta} N_p(0, V_\vartheta), \quad n \to \infty,$$

where $\xrightarrow{D_\vartheta}$ denotes convergence in distribution under P_ϑ.

10.3 Asymptotic linearity

Asymptotic linearity

A sequence of estimators $(T_n)_{n\in\mathbb{N}} \subset \mathbb{R}^p$ for $\gamma = g(\vartheta)$ is *asymptotically linear* if, for all $\vartheta \in \Theta$, there exists a function $\ell_\vartheta : \mathcal{X} \to \mathbb{R}^p$ with $\mathbb{E}_\vartheta \ell_\vartheta(X_1) = 0$ and $\mathbb{E}_\vartheta \|\ell_\vartheta(X_1)\|^2 < \infty$, such that

$$T_n - \gamma = \frac{1}{n}\sum_{i=1}^{n} \ell_\vartheta(X_i) + o_{P_\vartheta}\left(\frac{1}{\sqrt{n}}\right).$$

The function ℓ_ϑ is the *influence function*. The matrix $V_\vartheta = \mathbb{E}_\vartheta \ell_\vartheta(X_1)\ell_\vartheta(X_1)'$ is the *asymptotic covariance*.

Asymptotic linearity implies asymptotic normality.

10.4 The delta-technique

Delta-method (-technique)

Let (T_n), Z be $\mathbb{R}^p$-valued random variables, $c \in \mathbb{R}^p$, (r_n) with $r_n > 0$, $r_n \downarrow 0$, and $h : \mathbb{R}^p \to \mathbb{R}$ differentiable at c with gradient $\dot{h}(c) \in \mathbb{R}^p$. If

$$\frac{T_n - c}{r_n} \xrightarrow{D} Z,$$

then

$$h(T_n) - h(c) = \dot{h}(c)'(T_n - c) + o_P(r_n),$$

and in particular,

$$\frac{h(T_n) - h(c)}{r_n} \xrightarrow{D} \dot{h}(c)'Z.$$

Corollary: Asymptotic normality

If (T_n) is a sequence of asymptotically normal estimators for $\gamma = g(\vartheta) \in \mathbb{R}^p$ with asymptotic covariance V_ϑ and h is differentiable at γ, then $(h(T_n))$ is asymptotically normal for $h(\gamma)$ with asymptotic variance

$$\dot{h}(\gamma)' V_\vartheta \dot{h}(\gamma).$$

If (T_n) is asymptotically linear with influence function ℓ_ϑ, then $(h(T_n))$ is asymptotically linear for $h(\gamma)$, with influence function $\dot{h}(\gamma)'\ell_\vartheta$.

11 Useful

11.1 Formulas

Convolution of independent discrete random variables

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

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

Moments and Moment Generating Functions (MGF)

For a random variable X, the moment generating function is defined as

$$M_X(t) := \mathbb{E}[e^{tX}],$$

provided the expectation exists in a neighborhood of $t = 0$.
Moments of X can be extracted from the MGF by differentiation:

- The n-th moment is

$$\mathbb{E}[X^n] = M_X^{(n)}(0) := \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}.$$

- In particular,

$$\mathbb{E}[X] = M_X'(0), \quad \mathbb{E}[X^2] = M_X''(0).$$

11.2 Tables

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

11.3 Distributions

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

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: $\text{Var}(X) = \lambda$
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 $X \sim \text{Pois}(\theta)$ and $Y \sim \text{Pois}(\lambda)$ are independent, then

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

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.

Binomial distribution $\mathrm{Binomial}(n, p)$

Intuition: Models the number of successes in n independent trials, each with success probability p.
Parameters: $n \in \mathbb{N}$ (number of trials), $p \in [0, 1]$ (success probability)
Support: $k \in \{0, 1, \ldots, n\}$
PMF:

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

CDF: Not used, too complicated
Mean: $\mathbb{E}[X] = np$
Variance: $\mathrm{Var}(X) = np(1-p)$
Exponential family form:

$$p_p(k) = \exp\left(k \log p + (n-k)\log(1-p) + \log \binom{n}{k}\right),$$

for $k \in \{0, 1, \ldots, n\}$.
Characteristic function:

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \left(1 - p + pe^{it}\right)^n.$$

Fisher information:

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

Properties:

- If $X \sim \mathrm{Binomial}(n, p)$ and $Y \sim \mathrm{Binomial}(m, p)$ are independent with the same probability, then

$$X + Y \sim \mathrm{Binomial}(n + m, p).$$

- The binomial distribution is a special case of the multinomial distribution with $k = 2$ categories.

Multinomial distribution $\mathrm{Mult}(n, \mathbf{p})$

Intuition: Models the counts of outcomes when performing n independent trials, each resulting in one of k mutually exclusive categories with fixed probabilities.
Parameters: $n \in \mathbb{N}$ (number of trials), $\mathbf{p} = (p_1, \ldots, p_k)$ with $p_i \in [0, 1]$ and $\sum_{i=1}^k p_i = 1$ (category probabilities)
Support: $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{N}_0^k$ with $\sum_{i=1}^k x_i = n$
PMF:

$$p(\mathbf{x}) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

CDF: Not used, too complicated
Mean: $\mathbb{E}[X_i] = np_i$ for $i = 1, \ldots, k$
Variance: $\mathrm{Var}(X_i) = np_i(1-p_i)$ for $i = 1, \ldots, k$
Covariance: $\mathrm{Cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$
Exponential family form:

$$p_{\mathbf{p}}(\mathbf{x}) = \exp\left(\sum_{i=1}^k x_i \log p_i + \log \binom{n}{x_1, \ldots, x_k}\right),$$

with $\sum_{i=1}^k x_i = n$.
Characteristic function:

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}^T\mathbf{X}}] = \left(\sum_{i=1}^k p_i e^{it_i}\right)^n.$$

Fisher information: The Fisher information matrix is

$$I(\mathbf{p})_{ij} = \frac{n\delta_{ij}}{p_i} + \frac{n}{p_k}, \quad i, j = 1, \ldots, k-1,$$

where δ_{ij} is the Kronecker delta.
Properties:

- Each marginal $X_i \sim \mathrm{Binomial}(n, p_i)$.
- If $\mathbf{X} \sim \mathrm{Mult}(n, \mathbf{p})$ and $\mathbf{Y} \sim \mathrm{Mult}(m, \mathbf{p})$ are independent with the same probabilities, then

$$\mathbf{X} + \mathbf{Y} \sim \mathrm{Mult}(n + m, \mathbf{p}).$$