Indian Statistical Institute

Bachelor of Statistical Data Science

Statistical Inference

Utkarsh Bharadwaj

Indian Statistical Institute, Delhi

Summer 2026

Contents

1 Basic Concepts

Concept 1.1: Estimator and Estimate

Any function of the random sample which is used to estimate the unknown value of the given parametric function $g(\theta)$ is called an **estimator**.
If $X = (X_1, X_2, \ldots, X_n)$ is a random sample from a population with probability distribution P_θ, a function $d(\underline{x})$ used for estimating $g(\theta)$ is known as an estimator. Let $\underline{x} = (x_1, x_2, \ldots, x_n)$ be a realisation of X. Then $d(\underline{x})$ is called an **estimate**.

Example 1.1

To estimate the average height of adult males in an ethnic group, we may use the sample mean $\bar{x}$ as an estimator. Now if the sample has a mean of 180 cm, then 180 cm is the estimate of the average height.

Concept 1.2: Parameter Space

The parameter space Θ is the set of all possible values of the parameter θ.

2 Desirable Criteria for an Estimator

2.1 Unbiasedness

Concept 2.1: Unbiased Estimator

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with probability distribution P_θ, $\theta \in \Theta$. An estimator $T(X)$, $X = (X_1, X_2, \ldots, X_n)$, is said to be **unbiased** for estimating $g(\theta)$ if

$$E_\theta[T(X)] = g(\theta) \qquad \forall\, \theta \in \Theta.$$

If $E_\theta\, T(X) = g(\theta) + b(\theta)$, then $b(\theta)$ is called the **bias** of T. In particular:

- $b(\theta) > 0 \Rightarrow$ overestimate,
- $b(\theta) < 0 \Rightarrow$ underestimate.

Example 2.1

Let $X \sim \text{Bin}(n, p)$, n known, $0 \le p \le 1$.

$$E\left(\frac{X}{n}\right) = p \qquad \forall\, p.$$

So X/n (the *sample proportion*) is unbiased for p.
Also, $E[X(X-1)] = n(n-1)p^2$, so

$$E\left[\frac{X(X-1)}{n(n-1)}\right] = p^2.$$

Hence $\dfrac{X(X-1)}{n(n-1)}$ is unbiased for p^2.

Now $\text{Var}(X) = np(1-p) = n(p-p^2)$. Define

$$d(X) = n\left[\frac{X}{n} - \frac{X(X-1)}{n(n-1)}\right] = \frac{X(n-X)}{n-1}.$$

This is unbiased for $\text{Var}(X)$.

Example 2.2

Let $X_1, X_2, \ldots, X_n \sim P(\lambda)$. The estimators

$$T_1(X) = \bar{X}, \quad T_2(X) = X_1, \quad T_3(X) = \frac{X_1 + 2X_2}{3}$$

are all unbiased estimators of λ.
Also,

$$T_4(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is unbiased for λ (since $\text{Var}(X_i) = \lambda$ for Poisson).

Any linear combination $d(X) = \sum_{j=1}^{4} c_j T_j(X)$ is also unbiased for λ provided $\sum_{j=1}^{4} c_j = 1$.

Example 2.3

1. If $E(X)$ exists, then the sample mean $\bar{X}$ is an unbiased estimator of the population mean.
2. Let $E(X^2)$ exist, i.e. $\text{Var}(X) = \sigma^2$ exists. Then

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is unbiased for σ^2 (the *sample variance*).

Proof that $E(S^2) = \sigma^2$***:***

$$E(S^2) = \frac{1}{n-1}\sum_{i=1}^{n} E(X_i - \bar{X})^2 = \frac{1}{n-1} E\left(\sum X_i^2 - n\bar{X}^2\right).$$

We know $EX_i^2 = \mu^2 + \sigma^2$ and $E\bar{X}^2 = \mu^2 + \sigma^2/n$. Therefore

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}\left[n(\mu^2+\sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] \\ &= \frac{1}{n-1}\cdot(n-1)\sigma^2 = \sigma^2. \quad \square \end{aligned}$$

Further note. Consider

$$E\left(\bar{X}^2 - \frac{S^2}{n}\right) = \mu^2 + \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = \mu^2.$$

So $d^*(X) = \bar{X}^2 - S^2/n$ is unbiased for μ^2. $\square$

Example 2.4

Let $f_X(x) = \lambda e^{-\lambda x}$ for $x, \lambda > 0$ (exponential distribution). Let $X_1, \ldots, X_n$ be a random sample. Then $E(X_i) = 1/\lambda$, so $\bar{X}$ is an unbiased estimator for $1/\lambda$.
Also, $E(X_i^k) = k!/\lambda^k$, so

$$d_k(X) = \frac{1}{nk!} \sum_{i=1}^{n} X_i^k$$

is unbiased for $1/\lambda^k$.
Now let $Y = \sum X_i \sim \Gamma(n, \lambda)$ with pdf

$$f_Y(y) = \frac{\lambda^n}{\Gamma(n)} e^{-\lambda y} y^{n-1}, \quad y > 0.$$

Then

$$\begin{aligned} E\left(\frac{1}{Y}\right) &= \int_0^\infty \frac{1}{y} \cdot \frac{\lambda^n}{\Gamma(n)} e^{-\lambda y} y^{n-1}\, dy \\ &= \frac{\lambda^n}{\Gamma(n)} \cdot \frac{\Gamma(n-1)}{\lambda^{n-1}} \\ &= \frac{\lambda}{n-1}. \end{aligned}$$

Hence $E\left(\frac{n-1}{Y}\right) = \lambda$, and so we can estimate unbiasedly the reciprocal of the mean (i.e. the rate).
For higher powers: for $n > k$,

$$\begin{aligned} E\left(\frac{1}{Y^k}\right) &= \frac{\lambda^n}{\Gamma(n)} \cdot \frac{\Gamma(n-k)}{\lambda^{n-k}} \\ &= \frac{\Gamma(n-k)}{\Gamma(n)} \lambda^k. \end{aligned}$$

Therefore

$$\frac{\Gamma(n)}{\Gamma(n-k)} \cdot \frac{1}{Y^k}$$

is unbiased for λ^k.

Example 2.5

Let $X \sim P(\lambda)$. We want to estimate the probability of no occurrence, i.e. $P(X = 0) = e^{-\lambda}$.
Define

$$I(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{if } x \neq 0. \end{cases}$$

Then

$$E\{I(x)\} = 1 \cdot P(X = 0) + 0 \cdot P(X \neq 0) = e^{-\lambda}.$$

So $I(X)$ is an unbiased estimator of $e^{-\lambda}$.

A method to find unbiased estimators is to solve directly the equation $E\,T(X) = g(\theta)$

for all $\theta \in \Theta$.

Application: Truncated Poisson

Let X be a truncated Poisson r.v. with zero truncation:

$$P_X(x) = \frac{1}{e^\lambda - 1} \cdot \frac{\lambda^x}{x!}, \quad x = 1, 2, \ldots$$

Suppose we want to estimate λ. Setting $E\,T(X) = \lambda$ for all $\lambda > 0$:

$$\sum_{x=1}^{\infty} T(x) \frac{\lambda^x}{x!\,(e^\lambda - 1)} = \lambda.$$

$$\Rightarrow \quad T(1)\lambda + T(2)\frac{\lambda^2}{2!} + \cdots = \lambda(e^\lambda - 1)$$
$$= \lambda\left(\lambda + \frac{\lambda^2}{2!} + \cdots\right).$$

Since two power series can be identical on an open interval iff their coefficients match, comparing coefficients gives:

$$T(1) = 0, \quad T(2) = 2, \quad T(3) = 3, \quad \ldots, \quad T(x) = x.$$

So the unbiased estimator is

$$T(X) = \begin{cases} 0 & \text{if } x = 1, \\ x & \text{if } x = 2, 3, \ldots \end{cases}$$

Remark 2.1: Remarks on Unbiasedness

1. Unbiased estimators may not exist.
   Example: $X \sim \text{Bin}(n, p)$, $g(p) = p^{n+1}$. Then $E\,T(X) = p^{n+1}$ for all $p \in [0, 1]$. The LHS is a polynomial in p of degree at most n, while the RHS has degree $n + 1$; so LHS $\neq$ RHS and no such $T(X)$ exists.
2. Unbiased estimators may not be reasonable. The estimator must take values in the range of the parameter only.

2.2 Consistency

Concept 2.2: Consistent Estimator

An estimator $T_n = T(X_1, \ldots, X_n)$ is said to be **consistent** for estimating $g(\theta)$ if for each $\varepsilon > 0$,

$$P(|T_n - g(\theta)| \geq \varepsilon) \to 0 \quad \text{as } n \to \infty, \quad \forall\, \theta \in \Theta.$$

Example 2.6

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. By Chebyshev's inequality:

$$P\Big(|\bar{X} - \mu| \geq \varepsilon\Big) \leq \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ as } n \to \infty.$$

Hence $\bar{X}$ is consistent for μ.

Example 2.7

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.'s with mean μ. Then by the **Weak Law of Large Numbers**,

$$\bar{X} \xrightarrow{P} \mu.$$

So whenever the population mean exists, the sample mean is consistent for it. However, consider the **Cauchy distribution**:

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \nu)^2}, \quad x, \nu \in \mathbb{R}.$$

Here $E(X)$ does not exist and

$$P(|\bar{X} - \nu| < \varepsilon) = \frac{2}{\pi} \arctan \varepsilon \not\to 0.$$

Hence here consistency fails.

Theorem 2.1

If $E(T_n) = \theta_n \to \theta$ and $\mathrm{Var}(T_n) = s_n^2 \to 0$, then T_n is consistent for θ.

Theorem 2.2

If T_n is consistent for θ and $a_n \to 1$, $b_n \to 0$, then $a_n T_n + b_n$ is also consistent for θ.

Theorem 2.3

If T_n is consistent for θ and h is a continuous function, then $h(T_n)$ is consistent for $h(\theta)$.

Example 2.8: Exponential with Location Parameter

Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with location parameter ν:

$$f_{X_i}(x) = \begin{cases} e^{\nu - x} & x > \nu, \\ 0 & \text{otherwise.} \end{cases}$$

Then $E(X_i) = \nu + 1$, so $E(\bar{X}) = \nu + 1$. Thus $T_1 = \bar{X} - 1$ is an unbiased and consistent estimator of ν.

Now consider $Y = X_{(1)} = \min\{X_1, \ldots, X_n\}$. The CDF of $X_{(1)}$ is:

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \le x) = 1 - P(X_{(1)} > x) \\ &= 1 - [P(X_1 > x)]^n = 1 - [1 - F_{X_i}(x)]^n. \end{aligned}$$

Since $F_{X_i}(x) = \int_{\nu}^{x} e^{\nu - t}\, dt = 1 - e^{\nu - x}$ for $x > \nu$, we get

$$F_{X_{(1)}}(x) = 1 - e^{n(\nu - x)}, \quad x > \nu.$$

The pdf is:

$$f_{X_{(1)}}(x) = n e^{n(\nu - x)}, \quad x > \nu.$$

$$E(X_{(1)}) = \nu + \frac{1}{n}.$$

So $T_2 = X_{(1)} - \dfrac{1}{n}$ is also unbiased for ν.
Since $\mathrm{Var}(T_2) = \dfrac{1}{n^2} \to 0$ as $n \to \infty$, T_2 is also consistent for ν.
Note: $T_3 = X_n$ (the maximum) is consistent for θ. Since $T_2 = \left(1 + \frac{1}{n}\right) X_{(1)} \to 1 \cdot X_{(1)}$ as $n \to \infty$, and $X_{(1)} \xrightarrow{P} \nu$, T_2 is also consistent.

Example 2.9: Uniform Distribution

Let $X_1, \ldots, X_n$ be a random sample from Unif$[0, \theta]$, $\theta > 0$.

$$f_{X_i}(x) = \begin{cases} 1/\theta & 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$EX_i = \theta/2$, so $E\bar{X} = \theta/2$, hence $T_1 = 2\bar{X}$ is unbiased and consistent for θ.
Let $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Then:

$$F_{X_{(n)}}(x) = [F_{X_i}(x)]^n = \left(\frac{x}{\theta}\right)^n, \quad 0 \le x \le \theta.$$

$$f_{X_{(n)}}(x) = \frac{n x^{n-1}}{\theta^n}, \quad 0 \le x \le \theta.$$

$$E(X_{(n)}) = \int_0^{\theta} x \cdot \frac{n x^{n-1}}{\theta^n}\, dx = \frac{n\theta}{n+1}.$$

$$E\left[\frac{n+1}{n} X_{(n)}\right] = \theta.$$

So $T_2 = \dfrac{n+1}{n} X_{(n)}$ is unbiased for θ.

Also, $\lim\limits_{n \to \infty} F_{X_{(n)}}(x) = \begin{cases} 0 & x < \theta, \\ 1 & x \ge \theta. \end{cases}$

This is the CDF of a degenerate r.v. at θ, so $X_{(n)} \xrightarrow{P} \theta$. Thus $T_3 = X_{(n)}$ is consistent for θ.
Since $T_2 = \left(1 + \frac{1}{n}\right) X_{(n)} \to X_{(n)}$ as $n \to \infty$, T_2 is also consistent for θ.

Example 2.10: General Continuous Population

Let $X_1, \ldots, X_n$ be a random sample from a continuous population with CDF $F(x)$ and the range of variables is the interval $[a, b]$. Let $U = X_{(1)}$ and $V = X_{(n)}$.

$$F_U(x) = \begin{cases} 0 & x < a, \\ 1 - [1 - F(x)]^n & a \le x < b, \\ 1 & x \ge b. \end{cases} \qquad F_V(x) = \begin{cases} 0 & x < a, \\ [F(x)]^n & a \le x < b, \\ 1 & x \ge b. \end{cases}$$

As $n \to \infty$:

$$F_U(x) \to \begin{cases} 0 & x < a, \\ 1 & x \ge a. \end{cases} \quad \Rightarrow \quad U \xrightarrow{d} a, \quad U \xrightarrow{P} a.$$

$$F_V(x) \to \begin{cases} 0 & x < b, \\ 1 & x \ge b. \end{cases} \quad \Rightarrow \quad V \xrightarrow{P} b.$$

As a consequence, the sample range $X_{(n)} - X_{(1)}$ is consistent for the population range $(b - a)$.

3 Efficiency of Estimators

Concept 3.1: Mean Squared Error

For an estimator T of $g(\theta)$:

$$\text{MSE}(T) = E[T - g(\theta)]^2 = \text{Var}(T) + B^2(T),$$

where $B(T) = E(T) - g(\theta)$ is the bias. (The cross-product term vanishes.)
We say estimator T_1 is **better (more efficient)** than T_2 if

$$\text{MSE}(T_1) \le \text{MSE}(T_2) \qquad \forall\, \theta \in \Theta.$$

If T is unbiased for $g(\theta)$, then $\text{MSE}(T) = \text{Var}(T)$.

4 Uniformly Minimum Variance Unbiased Estimator (UMVUE)

Concept 4.1: UMVUE

An estimator w is said to be the **Uniformly Minimum Variance Unbiased Estimator (UMVUE)** of $g(\theta)$ if w is unbiased and for any other unbiased estimator w^* of $g(\theta)$,

$$\text{Var}(w) \le \text{Var}(w^*) \qquad \forall\, \theta \in \Theta.$$

Theorem 4.1: Uniqueness of UMVUE

If w is the UMVUE of $g(\theta)$, then w is unique almost everywhere.

Proof:

Let w^* be another UMVUE of $g(\theta)$. Then $E(w) = E(w^*) = g(\theta)$ and $\text{Var}(w) = \text{Var}(w^*) = \sigma^2$ (say). Define $w_1 = \frac{1}{2}(w + w^*)$. Then

$$\begin{aligned}\text{Var}(w_1) &= \frac{1}{4}\left[\text{Var}(w) + \text{Var}(w^*) + 2\text{Cov}(w, w^*)\right] \\ &= \frac{1}{4}\left[2\sigma^2 + 2\text{Cov}(w, w^*)\right].\end{aligned}$$

By Cauchy–Schwarz:

$$\text{Cov}(w, w^*) \leq \sqrt{\text{Var}(w)\,\text{Var}(w^*)} = \sigma^2.$$

So $\text{Var}(w_1) \leq \sigma^2 = \text{Var}(w)$. Since w is UMVUE, equality must hold in (1): $\text{Var}(w_1) = \text{Var}(w)$, which requires $\text{Cov}(w, w^*) = \sigma^2$.
For equality to hold, $w^* = a(\theta)\, w + b(\theta)$ with probability 1. Taking expectations:

$$g(\theta) = a(\theta)\, g(\theta) + b(\theta) \quad \Rightarrow \quad a(\theta) = 1,\ b(\theta) = 0 \quad \Rightarrow \quad w^* = w.$$

So w is unique. □

Condition for T to be UMVUE of $g(\theta)$

Let U_g be the class of all unbiased estimators of $g(\theta)$ and U_0 be the class of all unbiased estimators of 0.

Theorem 4.2

$T \in U_g$ with $\text{Var}_{\theta_0}(T) < \infty$ has minimum variance at $\theta = \theta_0$ iff $\text{Cov}(T, f) = 0$ for all $f \in U_0$ for which $\text{Var}_{\theta_0}(f) < \infty$.

Proof:

Let $T \in U_g$ with $\text{Var}_{\theta_0}(T)$ be minimum. Now let $f \in U_0 \ni \text{Cov}(T, f) \neq 0$. Consider $T + \lambda f$.

$$E(T + \lambda f) = g(\theta) + \lambda \cdot 0 = g(\theta), \quad \text{so } T + \lambda f \in U_g.$$

$$\text{Var}_{\theta_0}(T + \lambda f) = \text{Var}_{\theta_0}(T) + \lambda^2 \text{Var}_{\theta_0}(f) + 2\lambda \text{Cov}_{\theta_0}(T, f) \leq \text{Var}_{\theta_0}(T).$$

$$\Rightarrow \quad \lambda(\lambda \text{Var}_{\theta_0}(f) + 2\text{Cov}_{\theta_0}(T, f)) \leq 0 \quad \cdots (*)$$

Condition $(*)$ is satisfied for

$$0 < \lambda \leq -\frac{2\text{Cov}_{\theta_0}(T, f)}{\text{Var}_{\theta_0}(f)} \quad \text{if } \text{Cov}_{\theta_0}(T, f) < 0,$$

$$-\frac{2\text{Cov}_{\theta_0}(T, f)}{\text{Var}_{\theta_0}(f)} \leq \lambda < 0 \quad \text{if } \text{Cov}_{\theta_0}(T, f) > 0.$$

Since we assumed $\text{Var}_{\theta_0}(T)$ to be minimum, the above relation contradicts our assumption. Hence $\text{Cov}_{\theta_0}(T, f) = 0$.
Conversely, let $\text{Cov}_{\theta_0}(T, f) = 0$ for all $f \in U_0$. Let $T' \in U_g$; then $T - T' \in U_0$.

$$\text{Cov}(T, T-T') = 0 \;\Rightarrow\; \text{Var}(T) = \text{Cov}(T, T') \le \sqrt{\text{Var}(T) \cdot \text{Var}(T')} \;\Rightarrow\; \text{Var}(T) \le \text{Var}(T').$$

This proves T has minimum variance at θ_0. □

Remark 4.1: Remarks on UMVUE

1. The covariance between the UMVUE and any other unbiased estimator is always positive.
2. If T_1 and T_2 are UMVUEs of $g_1(\theta)$ and $g_2(\theta)$ respectively, then $a_1T_1 + a_2T_2$ is UMVUE for $a_1g_1(\theta) + a_2g_2(\theta)$.
   Proof: $\text{Cov}(a_1T_1 + a_2T_2, f) = a_1\text{Cov}(T_1, f) + a_2\text{Cov}(T_2, f) = 0$ for all $f \in U_0$.

Example 4.1

$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I)$, $n \times 1$.
Let $h(Y)$ be a real valued function with $E\,h(Y) = 0$ for all $\beta \in \mathbb{R}^p$.

$$K \int h(y)\, e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)}\, dy = 0 \quad \forall\, \beta \in \mathbb{R}^p.$$

$$\Rightarrow \quad K \int h(y)\, e^{-\frac{1}{2\sigma^2}(y'y-2y'X\beta)}\, dy = 0 \quad \forall\, \beta.$$

Differentiating with respect to β:

$$K \int h(y)\, X'y\, e^{-\frac{1}{2\sigma^2}(y'y-2y'X\beta)}\, dy = 0 \quad \forall\, \beta.$$

$$\Rightarrow \quad E(h(Y) \cdot X'y) = 0 \;\Rightarrow\; E(h(Y) - X'Xy) = 0.$$

So $X'Xy$ is UMVUE of $E(X'Xy) = X'X\beta$.

5 Method of Moments Estimation

5.1 Basic Concept

Concept 5.1: Method of Moments Estimation (MME)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution P_θ, $\theta \in \Theta$, where $\theta = (\theta_1, \ldots, \theta_k)$.
Consider the first k non-central moments:

$$\mu_1' = E(X_1) = g_1(\theta), \quad \mu_2' = E(X_1^2) = g_2(\theta), \quad \ldots, \quad \mu_k' = E(X_1^k) = g_k(\theta). \quad (1)$$

Assuming the system of equations (1) has solutions:

$$\theta_1 = h_1(\mu_1', \ldots, \mu_k'), \quad \ldots, \quad \theta_k = h_k(\mu_1', \ldots, \mu_k').$$

Define the first k non-central sample moments:

$$d_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad d_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \ldots, \quad d_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

In the method of moments, we estimate the k-th population moment by the k-th sample moment:

$$\hat{\mu}'_j = d_j, \quad j = 1, 2, \ldots, k.$$

Thus the method of moments estimators are:

$$\hat{\theta}_1 = h_1(d_1, d_2, \ldots, d_k), \quad \ldots, \quad \hat{\theta}_k = h_k(d_1, d_2, \ldots, d_k).$$

Example 5.1

1. $X_1, \ldots, X_n \sim P(\lambda)$. $\mu'_1 = \lambda$, so $\hat{\lambda}_{\text{MME}} = \bar{X}$. This MME is unbiased and consistent.
2. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. $\mu'_1 = \mu$, $\mu'_2 = \mu^2 + \sigma^2$, so $\mu = \mu'_1$ and $\sigma^2 = \mu'_2 - (\mu'_1)^2$.

$$\hat{\mu}_{\text{MME}} = \bar{X}, \qquad \hat{\sigma}^2_{\text{MME}} = d_2 - d_1^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2.$$

   $\hat{\mu}_{\text{MME}}$ is unbiased and consistent; $\hat{\sigma}^2_{\text{MME}}$ is biased but consistent.
3. $X_1, \ldots, X_n \sim \text{Bin}(n, p)$.
   Case I: n known. $\mu'_1 = np \Rightarrow p = \mu'_1/n \Rightarrow \hat{p} = \bar{X}/n$.
   Case II: Both n, p unknown.

$$\mu'_1 = np, \quad \mu'_2 = n^2p^2 + np(1-p).$$

$$\Rightarrow \quad 1 - p = \frac{\mu'_2 - (\mu'_1)^2}{\mu'_1}, \quad p = 1 - \frac{\mu'_2 - \mu'^2_1}{\mu'_1}, \quad n = \frac{\mu'_1}{p}.$$

$$\hat{p}_{\text{MME}} = 1 - \frac{d_2 - d_1^2}{d_1} = \bar{X} - \frac{\frac{1}{k}\sum (X_i - \bar{X})^2}{\bar{X}},$$
$$\hat{n}_{\text{MME}} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{k}\sum (X_i - \bar{X})^2}.$$

   $\hat{p}_{\text{MME}}$ is unbiased and consistent when n was known; when n is unknown, $\bar{X} \xrightarrow{P} np$ and $\hat{n} \xrightarrow{P} n$.

Remark 5.1: Remarks on MME

1. The MME need not be unbiased.
2. If the functions g_i's are continuous and one-to-one, then the MMEs will be consistent.

Example 5.2

1. $X_1, \ldots, X_n \sim \text{Unif}[a, b]$, $a < b$.

$$\mu_1' = \frac{a+b}{2}, \qquad \mu_2' = \frac{a^2+b^2+ab}{3}.$$

$$a = \mu_1' - \sqrt{3(\mu_2' - \mu_1'^2)}, \qquad b = \mu_1' + \sqrt{3(\mu_2' - \mu_1'^2)}.$$

$$\hat{a}_{\text{MME}} = \bar{X} - \sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2},$$
$$\hat{b}_{\text{MME}} = \bar{X} + \sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Here $\hat{a}_{\text{MME}}$ and $\hat{b}_{\text{MME}}$ are consistent but biased.

2. $X_1, \ldots, X_n \sim \text{Unif}[-\theta, \theta]$, $\theta > 0$. $\mu_1' = 0$, $\mu_2' = \theta^2/3$, so $\theta = \sqrt{3\mu_2'}$.

$$\hat{\theta} = \sqrt{3\mu_2'} = \sqrt{3d_2} = \sqrt{\frac{3}{n}\sum X_i^2}.$$

3. $X_1, \ldots, X_n \sim T(P, \lambda)$ (Gamma):

$$f(x) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda x} x^{P-1}, \quad x \geq 0,\ P, \lambda > 0.$$

$$\mu_1' = \frac{P}{\lambda}, \qquad \mu_2' = \frac{P(P+1)}{\lambda^2}.$$
$$P = \frac{\mu_1'^2}{\mu_2' - \mu_1'^2}, \qquad \lambda = \frac{\mu_1'}{\mu_2' - \mu_1'^2}.$$

$$\hat{P}_{\text{MME}} = \frac{\bar{X}^2}{\frac{1}{n}\sum(X_i - \bar{X})^2},$$
$$\hat{\lambda}_{\text{MME}} = \frac{\bar{X}}{\frac{1}{n}\sum(X_i - \bar{X})^2}.$$

These MMEs are consistent but biased.

4. $X_1, \ldots, X_n \sim \beta(d, \beta)$ (Beta):

$$f(x) = \frac{1}{B(d, \beta)} x^{d-1}(1-x)^{\beta-1}, \quad 0 < x < 1,\ d, \beta > 0.$$

$$\mu_1' = \frac{d}{d+\beta}, \qquad \mu_2' = \frac{d(d+1)}{(d+\beta)(d+\beta+1)}.$$
$$d = \frac{\mu_1'(\mu_1' - \mu_2')}{\mu_2' - \mu_1'^2}, \qquad \beta = \frac{(1-\mu_1')(\mu_1' - \mu_2')}{\mu_2' - \mu_1'^2}.$$

$$\hat{d}_{\text{MME}} = \frac{\bar{X}\left(\bar{X} - \frac{1}{n}\sum X_i^2\right)}{\frac{1}{n}\sum (X_i - \bar{X})^2},$$
$$\hat{\beta}_{\text{MME}} = \frac{(1-\bar{X})\left(\bar{X} - \frac{1}{n}\sum X_i^2\right)}{\frac{1}{n}\sum (X_i - \bar{X})^2}.$$

These MMEs are consistent but biased.
Note: For a k-dimensional parameter, we may have to consider more than k equations. Sometimes $\mu_1' = 0$ (e.g. for Unif$[-\theta, \theta]$), so we must use μ_2'.

6 Method of Maximum Likelihood Estimation (MLE)

This method was developed by **Sir R.A. Fisher** in 1922.

Example 6.1: Motivating Example

Consider a box containing black and white marbles. We know the ratio of one colour marble to the other is 1 : 3, but we don't know which marble is more numerous. A statistician draws a random sample of size 3 from the box and decides to base his estimate on the reported outcome.
Let x = number of black marbles in sample, p = unknown proportion of black marbles in the box. Then $x \sim \text{Bin}(3, p)$ with $p = 1/4$ or $3/4$.

$$P_x(x) = \binom{3}{x} p^x (1-p)^{3-x}, \quad x = 0, 1, 2, 3.$$

	$x=0$	$x=1$	$x=2$	$x=3$
$p = 1/4$	27/64	27/64	9/64	1/64
$p = 3/4$	1/64	9/64	27/64	27/64

We observe that the likelihood of $p = 1/4$ is higher when $x = 0$ or 1, and the likelihood of $p = 3/4$ is higher when $x = 2$ or 3. So:

$$\hat{p}_{\text{MLE}} = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } 1, \\ 3/4 & \text{if } x = 2 \text{ or } 3. \end{cases}$$

More generally, for $P_x(x) = \binom{3}{x} p^x (1-p)^{3-x}$ with $0 \le p \le 1$:

$$\log P(x) = \log \binom{3}{x} + x \log p + (3-x)\log(1-p).$$

$$\frac{d \log P(x)}{dp} = \frac{x}{p} - \frac{3-x}{1-p} = \frac{x - 3p}{p(1-p)}.$$

This is > 0 if $p < x/3$ and < 0 if $p > x/3$. So the maximum is attained at $p = x/3$:

$$\hat{p}_{\text{MLE}} = \frac{x}{3}.$$

Concept 6.1: Likelihood Function and MLE

To formalise the procedure, consider the **likelihood function**. If random sample $x_1, x_2, \ldots, x_n$ is observed:

$$L(\theta, \underline{x}) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta), \qquad \underline{x} = (x_1, x_2, \ldots, x_n).$$

The value of θ, say $\hat{\theta}(\underline{x})$, such that

$$L(\hat{\theta}, \underline{x}) \geq L(\theta, \underline{x}) \qquad \forall\, \theta \in \Theta$$

is called the **Maximum Likelihood Estimator (MLE)** of θ.
Remark: In practice, we often maximise $\log L(\theta, \underline{x}) = \ell(\theta, \underline{x})$ w.r.t. θ, as log is an increasing function of θ.
Remark: A useful approach is to find solutions of the **likelihood equations**:

$$\frac{\partial \ell}{\partial \theta} = 0, \qquad \left(\frac{\partial \ell}{\partial \theta_1} = 0, \ \ldots, \ \frac{\partial \ell}{\partial \theta_k} = 0 \right).$$

Example 6.2: MLE Examples

1. $X \sim \text{Bin}(n, p)$, $0 \leq p \leq 1$, n known.

$$L(p, x) = f(x, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

$$\ell(p) = \log \binom{n}{x} + x \log p + (n - x) \log(1 - p).$$

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

This is > 0 if $p < x/n$ and < 0 if $p > x/n$. So the maximum is attained at $\hat{p}_{\text{MLE}} = x/n$.

2. $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$, $\lambda > 0$.

$$L(\lambda, \underline{x}) = \prod_{i=1}^{n} f(x_i; \lambda) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod (x_i!)}$$

$$\ell(\lambda, \underline{x}) = \log L = -n\lambda + \sum x_i \log \lambda - \log \prod (x_i!)$$

$$\frac{d\ell}{d\lambda} = -n + \frac{\sum x_i}{\lambda} = \frac{\sum x_i - n\lambda}{\lambda} \begin{cases} > 0 & \lambda < \bar{x} \\ < 0 & \lambda > \bar{x} \end{cases}$$

So λ is maximised at $\lambda = \bar{x}$, i.e. $\hat{\lambda}_{\text{MLE}} = \bar{x}$.
Restriction $\lambda \leq \lambda_0$*:* If $\lambda_0 > \bar{x}$, MLE is still $\bar{x}$. If $\lambda_0 < \bar{x}$, MLE is λ_0.

$$\hat{\lambda}_{\text{MLE}} = \begin{cases} \bar{x} & \lambda_0 > \bar{x} \\ \lambda_0 & \lambda_0 < \bar{x} \end{cases}$$

3. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, μ unknown, $\sigma^2 = 1$ known (WLOG).

$$L(\mu, \underline{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum(x_i - \mu)^2\right\}$$

$$\ell(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum(x_i - \mu)^2$$

$$\frac{d\ell}{d\mu} = \sum 2(x_i - \mu)\cdot\frac{1}{2} = \sum x_i - n\mu = 0 \;\Rightarrow\; \hat{\mu}_{\text{MLE}} = \frac{\sum x_i}{n} = \bar{x}.$$

If $\mu \geq \mu_0$: MLE is $\max\{\bar{x}, \mu_0\}$ (clipped at boundary).

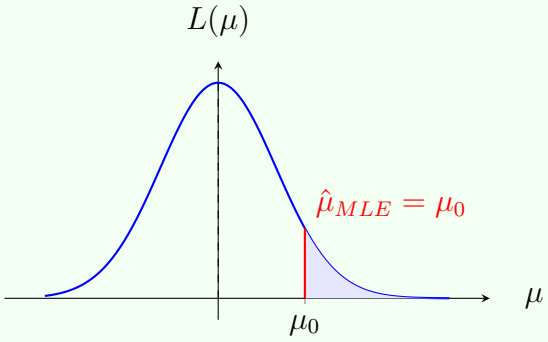

Figure: Likelihood function when $\mu \geq \mu_0$ but $\bar{x} < \mu_0$. The maximum in the valid region occurs at the boundary.

If $a \leq \mu \leq b$: MLE is $\bar{x}$ if $a \leq \bar{x} \leq b$, b if $\bar{x} > b$, a if $\bar{x} < a$.

4. $X_1, \ldots, X_n \sim \text{Unif}[0, \theta]$, $\theta > 0$.

$$L(\theta, \underline{x}) = \frac{1}{\theta^n}, \quad 0 \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \leq \theta.$$

We may write $L(\theta, \underline{x}) = \dfrac{1}{\theta^n}\,\mathbf{1}_{[x_{(n)}, \infty)}(\theta)$.

The minimum possible value of θ is $x_{(n)}$, so to maximise L, we minimise θ.

$$\hat{\theta}_{\text{MLE}} = x_{(n)}.$$

5. $X_1, \ldots, X_n \sim \text{Unif}[\theta_1, \theta_2]$, $\theta_1 < \theta_2$.

$$L(\theta_1, \theta_2, \underline{x}) = \frac{1}{(\theta_2 - \theta_1)^n}, \quad \theta_1 \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \leq \theta_2.$$

To maximise L, θ_2 should be maximum and θ_1 should be minimum.

$$\hat{\theta}_{1,\text{MLE}} = x_{(1)}, \qquad \hat{\theta}_{2,\text{MLE}} = x_{(n)}.$$

6. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, both μ and σ^2 unknown.

$$L(\mu, \sigma^2, \underline{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2\right\}$$

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(x_i - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = 0 \;\Rightarrow\; \frac{\sum(x_i-\mu)}{\sigma^2} = 0 \;\Rightarrow\; \hat{\mu}_{\text{MLE}} = \bar{x}.$$

$$\frac{\partial \ell}{\partial \sigma^2} = 0 \;\Rightarrow\; -\frac{n}{2\sigma^2} + \frac{\sum(x_i-\mu)^2}{2\sigma^4} = 0 \;\Rightarrow\; \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum(x_i-\hat{\mu})^2 = \frac{1}{n}\sum(x_i-\bar{x})^2.$$

Prior information $\mu \geq 0$:

$$\hat{\mu}_{\text{MLE}} = \max\{\bar{x}, 0\}, \qquad \hat{\sigma}^2_{\text{MLE}} = \begin{cases} \frac{1}{n}\sum(x_i-\bar{x})^2 & \text{if } \bar{x} > 0, \\ \frac{1}{n}\sum x_i^2 & \text{if } \bar{x} < 0. \end{cases}$$

7. $X_1, \ldots, X_n \sim$ Discrete Uniform with pmf $P_X(k) = 1/N$, $k = 1, \ldots, N$, $N \in \mathbb{Z}^+$.

$$L(N, k_1, \ldots, k_n) = \frac{1}{N^n}, \quad 1 \leq k_{(1)} \leq k_{(2)} \leq \cdots \leq k_{(n)} \leq N.$$

L is maximised when N is minimum, attained at $N = k_{(n)}$.

$$\hat{N}_{\text{MLE}} = k_{(n)}.$$

8. **Hypergeometric Distribution.** Population of size N: category A (size M) and category B (size $N-M$). A random sample of size n is taken; let $X =$ number of items of type A.

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad \max(0, n-N+M) \leq x \leq \min(n, M).$$

Case I: M known, N unknown.

$$L(N, x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

$$\frac{L(N,x)}{L(N-1,x)} = \frac{(N-n)(N-M)}{N(N-M-n+x)} > 1 \;\text{ iff }\; N < \frac{nM}{x}.$$

So L achieves its maximum at $N = \dfrac{nM}{x}$ (as nM/x need not be an integer, we take):

$$\hat{N}_{\text{MLE}} = \left\lfloor \frac{nM}{x} \right\rfloor.$$

Case II: M unknown, N known.

$$L^*(M, x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{x}}.$$

$$\frac{L^*(M,x)}{L^*(M-1,x)} = \frac{M}{M-x} \cdot \frac{N+x-M-n+1}{N-M+1}.$$

$$\frac{L^*(M,x)}{L^*(M-1,x)} > 1 \;\text{ iff }\; M < \frac{N+1}{n}\,x.$$

$$\hat{M}_{\text{MLE}} = \left\lfloor \frac{N+1}{n}\,x \right\rfloor.$$

Example 6.3: Exp(μ, τ) distribution

Let $X_1, \ldots, X_n \sim \text{Exp}(\mu, \tau)$ with pdf

$$f(x) = \frac{1}{\tau} e^{-(x-\mu)/\tau}, \quad x > \mu,\ \mu \in \mathbb{R},\ \tau > 0.$$

$$L(\mu, \tau, \underline{x}) = \frac{1}{\tau^n} \exp\Big\{-\frac{1}{\tau}\sum(x_i - \mu)\Big\}, \quad x_i > \mu.$$

Case I: μ known, say $\mu = 0$.

$$L(\tau, \underline{x}) = \frac{1}{\tau^n} \exp\Big\{-\frac{1}{\tau}\sum x_i\Big\}, \quad x_i > 0.$$

$$\ell(\tau) = -n\log\tau - \frac{1}{\tau}\sum x_i.$$

$$\frac{d\ell}{d\tau} = -\frac{n}{\tau} + \frac{\sum x_i}{\tau^2} = \frac{\sum x_i - n\tau}{\tau^2} \begin{cases} > 0 & \tau < \bar{x} \\ < 0 & \tau > \bar{x} \end{cases} \quad \Rightarrow \quad \hat{\tau}_{\text{MLE}} = \bar{x}.$$

Case II: τ known, $\tau = 1$.

$$L(\mu, \underline{x}) = \exp\Big\{-\sum(x_i - \mu)\Big\} = \frac{(e^{\mu})^n}{e^{\sum x_i}}, \quad \mu \le x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}.$$

$L(\mu)$ is maximised when μ takes its maximum value, which is $x_{(1)}$. So $\hat{\mu}_{\text{MLE}} = x_{(1)}$.

Case III: Both μ and τ unknown.

$$L(\mu, \tau, \underline{x}) = \frac{1}{\tau^n} \exp\Big\{-\frac{1}{\tau}\Big[\sum x_i - n\mu\Big]\Big\}, \quad \mu \le x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}.$$

For fixed τ, L is maximised w.r.t. μ when μ is maximum, i.e. $\hat{\mu}_{\text{MLE}} = x_{(1)}$.

$$\frac{\partial\ell}{\partial\tau} = -\frac{n}{\tau} + \frac{1}{\tau^2}\Big(\sum x_i - n\mu\Big) = 0 \ \Rightarrow\ \tau = \frac{\sum x_i - n\mu}{n} = \bar{x} - \mu.$$

$$\hat{\tau}_{\text{MLE}} = \bar{x} - x_{(1)}.$$

Example 6.4: Laplace / Double Exponential Distribution

Let $X_1, \ldots, X_n$ be a random sample from the double exponential distribution with pdf

$$f(x; \mu, \tau) = \frac{1}{2\tau} e^{-|x-\mu|/\tau}, \quad x, \mu \in \mathbb{R},\ \tau > 0.$$

Case I: $\mu = 0$.

$$L(\tau, \underline{x}) = \frac{1}{(2\tau)^n} \exp\Big\{-\frac{1}{\tau}\sum|x_i|\Big\}.$$

$$\ell(\tau) = -n\log 2 - n\log\tau - \frac{1}{\tau}\sum|x_i|.$$

$$\frac{d\ell}{d\tau} = -\frac{n}{\tau} + \frac{\sum|x_i|}{\tau^2} = \frac{\sum|x_i| - n\tau}{\tau^2} = 0 \ \Rightarrow\ \hat{\tau}_{\text{MLE}} = \frac{\sum|x_i|}{n}.$$

Case II: $\tau = 1$.

$$L(\mu, \underline{x}) = \frac{1}{2^n} \exp\Big\{ - \sum |x_i - \mu| \Big\}.$$

L is maximised when $\sum |x_i - \mu|$ is minimised. It can be shown that $S = \sum |x_i - \mu|$ is minimised when μ is the median of $x_1, \ldots, x_n$.
Write $S = \sum_{i=1}^{n} |x_{(i)} - \mu|$ where $x_{(i)}$'s are ordered values.
Case (a): n odd, $n = 2R + 1$.

$$\begin{aligned} S &= |x_{(1)} - \mu| + \cdots + |x_{(2R+1)} - \mu| \\ &= \Big(|x_{(1)} - \mu| + |x_{(2R+1)} - \mu| \Big) + \cdots + \Big(|x_{(R)} - \mu| + |x_{(R+2)} - \mu| \Big) + |x_{(R+1)} - \mu|. \end{aligned}$$

The term $|x_{(j)} - \mu| + |x_{(2R+1-j)} - \mu|$ is minimum (equals $x_{(2R+1-j)} - x_{(j)}$) whenever $x_{(j)} \le \mu \le x_{(2R+1-j)}$.
We can clearly see that $\mu = x_{(R+1)}$ simultaneously satisfies all conditions. So $\hat{\mu}_{\text{MLE}} = x_{(R+1)} =$ median of $x_1, \ldots, x_{2R+1}$.
Case (b): n even, $n = 2R$.

$$S = \Big(|x_{(1)} - \mu| + |x_{(2R)} - \mu| \Big) + \cdots + \Big(|x_{(R)} - \mu| + |x_{(R+1)} - \mu| \Big).$$

Arguing as before, S is minimum when $x_{(R)} \le \mu \le x_{(R+1)}$. So any value between $x_{(R)}$ and $x_{(R+1)}$ satisfies all conditions; we take $\mu = \dfrac{x_{(R)} + x_{(R+1)}}{2} =$ median.
Case III: Both μ and τ unknown.

$$L = \frac{1}{(2\tau)^n} \exp\Big\{ -\frac{1}{\tau} \sum |x_i - \mu| \Big\}, \qquad \ell(\mu, \tau) = -n \log 2 - n \log \tau - \frac{1}{\tau} \sum |x_i - \mu|.$$

ℓ is maximised w.r.t. μ when $\sum |x_i - \mu|$ is minimum, i.e. at $\mu = \text{med}(x_1, \ldots, x_n) = m$.
So $\hat{\mu}_{\text{MLE}} = m$.

$$\frac{\partial \ell}{\partial \tau} = -\frac{n}{\tau} + \frac{\sum |x_i - \mu|}{\tau^2} = 0 \;\Rightarrow\; \hat{\tau}_{\text{MLE}} = \frac{1}{n} \sum |x_i - m|$$

(mean deviation about the median).

7 Properties of MLE

7.1 Non-Uniqueness of MLE

Example 7.1

Let $X_1, \ldots, X_n \sim \text{Unif}(\theta - a, \theta + a)$, $\theta \in \mathbb{R}$, $a > 0$ a constant.

$$L(\theta, \underline{x}) = \frac{1}{(2a)^n}, \quad \theta - a < x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)} < \theta + a.$$

Equivalently: $x_{(n)} - a < \theta < x_{(1)} + a$ (and $x_{(n)} - a > 0 > x_{(1)} - a$ automatically when $2a > x_{(n)} - x_{(1)}$).

L is maximised when $\theta - a \le x_{(1)}$ or $\theta \ge x_{(n)} + a$, and $\theta + a \ge x_{(n)}$, i.e. $x_{(n)} - a \le \theta \le x_{(1)} + a$.
So any value between $x_{(n)} - a$ and $x_{(1)} + a$ is an MLE of θ. We may choose the midpoint:

$$\hat{\theta}_{\text{MLE}} = \frac{x_{(1)} + x_{(n)}}{2}.$$

7.2 MLE May Not Be in a Nice Analytic Form

Example 7.2

$X_1, \ldots, X_n \sim \mathcal{N}(\theta, \theta^2)$, $\theta > 0$.

$$L(\theta, \underline{x}) = \frac{1}{(\theta\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2\theta^2}\sum (x_i - \theta)^2\right\}.$$

$$\ell(\theta) = -n\log\theta - \frac{n}{2}\log(2\pi) - \frac{1}{2\theta^2}\sum (x_i - \theta)^2.$$

$$\frac{d\ell}{d\theta} = -\frac{n}{\theta} + \frac{\sum(x_i - \theta)}{\theta^2} + \frac{\sum(x_i - \theta)^2}{\theta^3} = \frac{\sum x_i^2 - n\theta\bar{x} - n\theta^2}{\theta^3}.$$

Setting $d\ell/d\theta = 0$ gives a quadratic in θ:

$$\frac{d\ell}{d\theta} = -\frac{n}{\theta^3}\left[\theta^2 + \bar{x}\,\theta - d_2\right] = 0,$$

with solutions

$$\hat{\theta}_{1,2} = \frac{-\bar{x} \pm \sqrt{(\bar{x})^2 + 4d_2}}{2}.$$

Since $\theta > 0$ and $\ell(\theta)$ is maximised at $\theta = \hat{\theta}_2$, the negative root $\hat{\theta}_1$ is automatically rejected:

$$\hat{\theta}_{\text{MLE}} = \frac{-\bar{x} + \sqrt{\bar{x}^2 + 4d_2}}{2}.$$

7.3 MLE May Not Be in a Closed Form

Example 7.3

$X_1, \ldots, X_n \sim T(\lambda, r)$ (Gamma).
Case I: λ unknown, $r \ge 1$ known.

$$L(\lambda, \underline{x}) = \frac{\lambda^{nr}}{[T(\lambda)]^n}\, e^{-\lambda\sum x_i} \left(\prod x_i\right)^{r-1}.$$

$$\ell(\lambda) = nr\log\lambda - n\log\Gamma(\lambda) - \lambda\sum x_i + (r-1)\sum \log x_i.$$

$$\frac{d\ell}{d\lambda} = \frac{nr}{\lambda} - n\frac{\Gamma'(\lambda)}{\Gamma(\lambda)} + \sum \log x_i = 0 \quad \Rightarrow \quad -\frac{n\,\Gamma'(\lambda)}{\Gamma(\lambda)} + \sum \log x_i = 0.$$

This involves the **Euler digamma function** and cannot be solved in closed form.

The **method of scoring** is used to find solutions of likelihood equations numerically.

Method of Scoring

We want to solve the likelihood equation $\dfrac{\partial \ell}{\partial \theta} = 0$.

Let the exact solution lie in the neighbourhood of a value θ_0. Expanding $\dfrac{\partial \ell}{\partial \theta}$ in a Taylor series around θ_0 up to second order terms:

$$\begin{aligned}\frac{\partial \ell}{\partial \theta} &\approx \frac{\partial \ell}{\partial \theta}\bigg|_{\theta_0} + (\theta - \theta_0)\,\frac{\partial^2 \ell}{\partial \theta^2}\bigg|_{\theta_0} \\ &\approx \frac{\partial \ell}{\partial \theta_0} - (\theta - \theta_0)\, E\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] \\ &= \frac{\partial \ell}{\partial \theta_0} - (\theta - \theta_0)\, I(\theta_0).\end{aligned}$$

The equation leads to the **scoring step**:

$$S(\theta) = \frac{\partial \ell / \partial \theta_0}{I(\theta_0)},$$

and we update $\theta_1 = \theta_0 + S(\theta)$, continuing until the desired level of accuracy.

Example 7.4: Cauchy Distribution

$f(x;\theta) = \dfrac{1}{\pi\,[1 + (x-\theta)^2]}$, $x, \theta \in \mathbb{R}$.

Suppose observations $210, 195, 190, 199, 198, 202, 185, 215$ are available. We want the MLE of θ.

$$L(\theta, \underline{x}) = \prod_{i=1}^{n} \frac{1}{\pi[1 + (x_i - \theta)^2]}$$

$$\ell(\theta) = -n \log \pi - \sum_{i=1}^{n} \log\left[1 + (x_i - \theta)^2\right].$$

The likelihood equation is $d\ell/d\theta = 0$:

$$2 \sum_{i=1}^{n} \frac{x_i - \theta}{1 + (x_i - \theta)^2} = 0.$$

For $n = 1$, we get $\hat{\theta} = x_1$. For $n \geq 2$, this is a non-linear equation of degree $2n - 1$. For the method of scoring, we compute $I(\theta)$:

$$\begin{aligned}I(\theta) &= -E\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] = E\left[\frac{\partial \log f}{\partial \theta}\right]^2 \\ &= 4E\left[\frac{(x-\theta)^2}{\{1 + (x-\theta)^2\}^2}\right].\end{aligned}$$

$$\begin{aligned}I(\theta) &= \frac{4}{\pi} \int_{-\infty}^{\infty} \frac{(x-\theta)^2}{\{1 + (x-\theta)^2\}^3}\, dx \\ &= \frac{8}{\pi} \int_{0}^{\infty} \frac{y^2}{(1+y^2)^3}\, dy.\end{aligned}$$

Using the substitution $y = \tan\phi$:

$$I(\theta) = \frac{8}{\pi}\int_0^{\pi/2} \frac{\tan^2\phi \sec^2\phi}{(\sec^2\phi)^3}\, d\phi$$
$$= \frac{8}{\pi}\int_0^{\pi/2} \sin^2\phi \cos^2\phi\, d\phi = \frac{1}{2}.$$

So $I(\theta_0) = 1/2$.
The scoring function:

$$S(\theta) = \frac{4}{n}\sum_{i=1}^{n} \frac{x_i - \theta}{1 + (x_i - \theta)^2}.$$

We take $\theta_0 = 198.5$ (sample median) as initial approximation. The successive approximations are:

$$\theta_1 = 198.4784887, \quad \theta_2 = 198.4656064, \quad \theta_3 = 198.4570831, \quad \ldots$$

We may take $\hat{\theta} \approx 198.4\overline{9}$.

7.4 Invariance of MLE

Theorem 7.1: Invariance of MLE

Let $\phi = g(\theta)$, where g is a one-to-one function. Then $\hat{\phi}_{\text{MLE}} = g(\hat{\theta}_{\text{MLE}})$.

Proof:

$\theta \in \Theta$, $\phi \in \Xi = g(\Theta) = \{g(\theta) : \theta \in \Theta\}$.
Let $\ell^*(\phi)$ and $\ell(\theta)$ denote the likelihood functions corresponding to ϕ and θ respectively.
Any MLE of ϕ should satisfy $L^*(\hat{\phi}_{\text{MLE}}) \geq L^*(\phi)$ for all $\phi \in \Xi$.
Now $L(\hat{\theta}_{\text{MLE}}) \geq L(\theta)$ for all $\theta \in \Theta$. For $\hat{\phi}_{\text{MLE}} = g(\hat{\theta}_{\text{MLE}})$:

$$L^*(\hat{\phi}_{\text{MLE}}) = L^*(g(\hat{\theta}_{\text{MLE}})) = L(g^{-1}(g(\hat{\theta}_{\text{MLE}}))) = L(\hat{\theta}_{\text{MLE}}) \geq L(\theta) = L^*(\phi)$$

for all $\phi \in \Xi$. So $\hat{\phi}_{\text{MLE}} = g(\hat{\theta}_{\text{MLE}})$ is the MLE of $\phi = g(\theta)$. □

Theorem 7.2: Zehna (1967)

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability distributions and let $\ell(\theta)$ be the likelihood function. Suppose $\Theta \subset \mathbb{R}^K$ and $\theta^* = g(\theta)$ be a function from $\mathbb{R}^K \to \mathbb{R}^P$, $1 \leq P \leq K$. Also $\theta^* \in \Theta^* \to$ an interval in $\mathbb{R}^P$. If $\hat{\theta}$ is an MLE of θ, then $g(\hat{\theta})$ is the MLE of $\theta^* = g(\theta)$.

7.5 Asymptotic Properties of MLE

Concept 7.1: Regularity Assumptions for MLE

Let $x \sim f(x,\theta)$, $\theta \in \Theta$ an open interval in $\mathbb{R}$.

Regularity Assumptions

1. $\dfrac{\partial^2 \log f}{\partial \theta^2}$ exists for almost all x in $|\theta - \theta_0| < \delta$ for some $\delta > 0$.
2. $E_{\theta_0}\left[\dfrac{\partial \log f(x,\theta)}{\partial \theta}\bigg|_{\theta=\theta_0}\right] = 0.$
3. $E_{\theta_0}\left[\dfrac{\partial \log f(x,\theta)}{\partial \theta}\bigg|_{\theta=\theta_0}\right]^2 > 0.$
4. $\left|\dfrac{\partial^3 \log f}{\partial \theta^3}\right| \le m(x)$ for all $\theta \in (\theta_0 - \delta, \theta_0 + \delta)$, and $E_{\theta_0}\, m(x) < K$.

Under these assumptions, we have the following large sample results for the MLE.

Theorem 7.3

The likelihood equation has a root with probability 1 as $n \to \infty$, which converges to θ_0 with probability 1 under θ_0.

Theorem 7.4

Let $\bar{\theta}$ be a consistent root of the likelihood equation. Then

$$\sqrt{n}\left[(\bar{\theta} - \theta_0)\, I(\theta_0) - \frac{1}{n}\frac{d\ell}{d\theta_0}\right] \to 0 \text{ with prob. 1.}$$

The asymptotic distribution of $\sqrt{n}(\bar{\theta} - \theta_0)$ is $\mathcal{N}\left(0, \dfrac{1}{I(\theta_0)}\right)$.

8 Lower Bounds for Variance

8.1 Wolfowitz's Regularity Condition (Fréchet–Rao–Cramér Inequality)

Concept 8.1: Wolfowitz's Regularity Condition

Let $X_1, \ldots, X_n$ be a random sample from a distribution having pdf (pmf) $f(x,\theta)$ w.r.t. measure ν. An estimate $S(\underline{x})$ is to be considered where:

1. θ lies in an open interval Θ of the real line.
2. $\dfrac{\partial f(z,\theta)}{\partial \theta}$ exists for all $\theta \in \Theta$ and a.e. x.
3. $\displaystyle\int S(\underline{x}) \prod_{i=1}^{n} f(x_i,\theta)\, d\nu(x_i)$ can be differentiated under the integral sign for any S such that the above integral exists.

4. $E_\theta\left[\dfrac{\partial \log f(x,\theta)}{\partial\theta}\right]^2 > 0$ for all $\theta \in \Theta$.

Theorem 8.1: Fréchet–Rao–Cramér (FRC) Inequality

Under assumptions (i)–(iv), if $E_\theta\, S(\underline{x}) = \theta + b(\theta)$, then

$$\mathrm{Var}_\theta(S) \geq \frac{[1+b'(\theta)]^2}{n\, E_\theta\left[\dfrac{\partial \log f(x,\theta)}{\partial\theta}\right]^2}. \qquad (*)$$

Proof:

$E_\theta\, S(\underline{x}) = \theta + b(\theta)$.

$$\Rightarrow \quad \int S(\underline{x}) \prod_{i=1}^n f(x_i,\theta)\, d\nu(\underline{x}) = \theta + b(\theta) \quad \forall\, \theta \in \Theta.$$

Differentiating under the integral sign w.r.t. θ:

$$\int S(\underline{x}) \left\{\sum_{i=1}^n \frac{\partial \log f(x_i,\theta)}{\partial\theta}\right\} \prod_{i=1}^n f(x_i,\theta)\, d\nu(\underline{x}) = 1 + b'(\theta).$$

Let $s(\underline{x},\theta) = \displaystyle\sum_{i=1}^n \frac{\partial \log f(x_i,\theta)}{\partial\theta}$.

$$E_\theta[S(\underline{x})\, s(\underline{x},\theta)] = 1 + b'(\theta). \qquad (2)$$

Now $\displaystyle\int \prod_{i=1}^n f(x_i,\theta)\, d\nu(\underline{x}) = 1$, so differentiating again:

$$\int \left\{\sum_{i=1}^n \frac{\partial \log f(x_i,\theta)}{\partial\theta}\right\} \prod_{i=1}^n f(x_i,\theta)\, d\nu(\underline{x}) = 0 \quad \Rightarrow \quad E_\theta\, s(\underline{x},\theta) = 0.$$

Using this in (2): $\mathrm{Cov}(S(\underline{x}), s(\underline{x},\theta)) = 1 + b'(\theta)$.
Squaring and using Cauchy–Schwarz:

$$\{1+b'(\theta)\}^2 = \mathrm{Cov}^2(S,s) \leq \mathrm{Var}(S)\cdot\mathrm{Var}(s).$$

Now

$$\begin{aligned}\mathrm{Var}_\theta\, s(\underline{x},\theta) &= \mathrm{Var}\left[\sum_{i=1}^n \frac{\partial \log f(x_i,\theta)}{\partial\theta}\right] \\ &= n\,\mathrm{Var}\left[\frac{\partial \log f(x,\theta)}{\partial\theta}\right] \\ &= n\, E\left[\frac{\partial \log f(x,\theta)}{\partial\theta}\right]^2 = I_{\underline{x}}(\theta).\end{aligned}$$

So (4) gives:

$$\{1+b'(\theta)\}^2 \le \mathrm{Var}(S)\cdot I_{\underline{x}}(\theta) \;\Rightarrow\; \mathrm{Var}(S) \ge \frac{\{1+b'(\theta)\}^2}{I_{\underline{x}}(\theta)} = \frac{\{1+b'(\theta)\}^2}{n\,I_x(\theta)}. \quad \square$$

Corollary: If $S(\underline{x})$ is unbiased for θ, then

$$\mathrm{Var}_\theta(S(\underline{x})) \ge \frac{1}{n\,E\left[\dfrac{\partial \log f(x,\theta)}{\partial\theta}\right]^2} = \frac{1}{I_x(\theta)} \quad \text{(Fisher's information inequality).}$$

$\square$

Remark 8.1: Remarks on FRC Inequality

1. Equality in the FRC inequality is achieved iff $S(\underline{x})$ and $s(\underline{x},\theta)$ are linearly related with probability 1, i.e. there exist functions $d(\theta)$ and $\beta(\theta)$ such that
$$S(\underline{x}) + d(\theta)\,s(\underline{x},\theta) = \beta(\theta) \quad \text{with prob. 1.}$$
2. Under the regularity conditions:
$$E\left[\frac{\partial \log f(x,\theta)}{\partial\theta}\right]^2 = -E\left[\frac{\partial^2 \log f(x,\theta)}{\partial\theta^2}\right].$$

Proof of the identity:

$$\begin{aligned} E\frac{\partial^2 \log f}{\partial\theta^2} &= E\frac{\partial}{\partial\theta}\left[\frac{f'(x,\theta)}{f(x,\theta)}\right] \\ &= E\frac{f''(x,\theta)\,f(x,\theta) - [f'(x,\theta)]^2}{[f(x,\theta)]^2}. \\ &= \int f''(x,\theta)\,d\nu(x) - \int \left[\frac{f'(x,\theta)}{f(x,\theta)}\right]^2 f(x,\theta)\,d\nu(x) \\ &= 0 - E\left[\frac{\partial \log f'}{\partial\theta}\right]^2. \quad \square \end{aligned}$$

$\square$

Theorem 8.2: FRC Lower Bound for estimating $\phi = g(\theta)$

If $S(\underline{x})$ is unbiased for $g(\theta)$, then $f(x,\theta) = f^*(x,\phi)$ where f^* is parameterised by ϕ. The FRC lower bound is:

$$\mathrm{Var}(S) \ge \frac{\{g'(\theta)\}^2}{n\,E\left[\dfrac{\partial \log f(x,\theta)}{\partial\theta}\right]^2} = \frac{[g'(\theta)]^2}{I_x(\theta)}.$$

The condition for attaining the FRCLB remains the same: $S(\underline{x})$ must be linearly related to $s(x,\theta)$ with probability 1.

Theorem 8.3

If the FRC lower bound for the variance of an unbiased estimator of $g(\theta)$ is attained, then the class of parametric functions for whom the unbiased estimator attains FRC lower bound is the class of linear functions of $g(\theta)$.

Proof:

Let $T(\underline{x})$ be an unbiased estimator of $g(\theta)$ and $V(T(\underline{x}))$ equal the FRC LB. Then $T(\underline{x})$ and $S(\underline{x}, \theta)$ are linearly related with prob. 1, i.e. there exist functions $d(\theta)$ and $\beta(\theta)$ such that

$$T(\underline{x}) + d(\theta)\, s(\underline{x}, \theta) = \beta(\theta) \quad \text{wp } 1 \quad \forall\, \theta \in \Theta.$$

Taking expectations: $g(\theta) + d(\theta) \cdot 0 = \beta(\theta)$, so $g(\theta) = \beta(\theta)$.
Now let $h(\theta)$ be any other parametric function for which there is an unbiased estimator $U(\underline{x})$ whose variance attains the corresponding FRCLB. Then $U(\underline{x})$ and $S(\underline{x}, \theta)$ are also linearly related wp 1, i.e. there exist $d^*(\theta)$ and $\beta^*(\theta)$ such that

$$U(\underline{x}) + d^*(\theta)\, s(\underline{x}, \theta) = \beta^*(\theta) \quad \text{wp } 1 \quad \forall\, \theta \in \Theta.$$

Once again, taking expectations: $h(\theta) = \beta^*(\theta)$.
Fixing $\theta = \theta_0$:

$$T(\underline{x}) + d(\theta_0)\, s(\underline{x}, \theta_0) = g(\theta_0), \quad U(\underline{x}) + d^*(\theta_0)\, s(\underline{x}, \theta_0) = h(\theta_0).$$

Eliminating $s(\underline{x}, \theta_0)$:

$$d^*(\theta_0)\, T(\underline{x}) - d(\theta_0)\, U(\underline{x}) = d^*(\theta_0)\, g(\theta_0) - d(\theta_0)\, h(\theta_0) \quad \text{wp } 1,$$

or $a\, T(\underline{x}) + b\, U(\underline{x}) = c$. Taking expectations: $a\, g(\theta) + b\, h(\theta) = c$, so g and h are linearly related. □

Example 8.1

1. $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$, $\lambda > 0$. Let $g(\lambda) = \lambda^2$. The FRCLB for variance of an unbiased estimator of λ^2 is:

   $$[g'(\lambda)]^2 \cdot \{\text{FRCLB for } \lambda\} = 4\lambda^2 \cdot \frac{\lambda}{n} = \frac{4\lambda^3}{n}.$$

   Let $Y = \sum X_i \sim \text{Poi}(n\lambda)$. Define $U = \frac{1}{n^2} Y(Y-1)$. Then $E(U) = \frac{1}{n^2}(EY^2 - EY) = \frac{1}{n^2}(n\lambda + n^2\lambda^2 - n\lambda) = \lambda^2$.

   $$\text{Var}(U) = \frac{4\lambda^3}{n} + \frac{2\lambda^2}{n} > \frac{4\lambda^3}{n}.$$

   So the FRCLB is not attained.
2. $X_1, \ldots, X_n$ i.i.d. with mean μ and variance $\sigma^2 (< \infty)$. $T_1 = \bar{X}$, $T_2 = \frac{2}{n(n+1)} \sum_{i=1}^{n} i\, X_i$.
   $E(T_1) = \mu$, $\text{Var}(T_1) = \sigma^2/n$. So T_1 is unbiased and consistent for μ.
   $E(T_2) = \frac{2}{n(n+1)} \sum_{i=1}^{n} i \cdot \mu = \frac{2}{n(n+1)} \cdot \frac{n(n+1)}{2} \mu = \mu$.

$\text{Var}(T_2) = \frac{4}{n^2(n+1)^2}\sum i^2\,\sigma^2 = \frac{4\sigma^2}{n^2(n+1)^2}\cdot\frac{n(n+1)(2n+1)}{6} = \frac{2(2n+1)\sigma^2}{3n(n+1)} \to 0$ as $n\to\infty$.

So T_2 is also unbiased and consistent for σ^2.

Consider $\text{Var}(T_2)/\text{Var}(T_1) = \frac{2(2n+1)}{3(n+1)} > 1$ for $n > 1$.

So T_1 is more efficient than T_2.

3. $X_1,\ldots,X_n \sim \mathcal{N}(0,\sigma^2)$. Consider estimation of σ.

$$f(x,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-x^2/(2\sigma^2)}, \quad x\in\mathbb{R},\ \sigma>0.$$

$$I_x(\sigma) = \frac{2n}{\sigma^2}, \qquad \text{FRCLB for } \sigma = \frac{\sigma^2}{2n}.$$

Consider $V_d = d\sum_{i=1}^{n}|x_i|$.

$$E|x_i| = \sqrt{\frac{2}{\pi}}\,\sigma \;\Rightarrow\; E(V_d) = \frac{2n\sigma}{\sqrt{2\pi}}\,d = \sigma \;\Rightarrow\; d = \frac{1}{n}\sqrt{\frac{\pi}{2}}.$$

$$T_1 = \frac{1}{n}\sqrt{\frac{\pi}{2}}\sum_{i=1}^{n}|x_i| \text{ is an unbiased estimator of } \sigma.$$

$$\text{Var}T_1 = \frac{\pi}{2n^2}\,n\,\text{Var}(|x_i|) = \frac{\pi}{2n}\Big(Ex_i^2 - (E|x_i|)^2\Big) = \frac{\pi}{2n}\left(\sigma^2 - \frac{2\sigma^2}{\pi}\right) = \frac{(\pi-2)\sigma^2}{2n\pi}.$$

Since $(\pi-2)/\pi > 1/2$ (equivalently $\pi > 2$, which is always true), $\text{Var}(T_1) > \sigma^2/(2n) = \text{FRCLB}$. So T_1 does not attain FRCLB.

Define further $w_\beta = \beta(\sum x_i^2)^{1/2}$, $U = \sum x_i^2/\sigma^2 \sim \chi^2_n$.

$$Ew_\beta = \beta\cdot\sigma\cdot EU^{1/2} = \beta\cdot\sigma\cdot\frac{\sqrt{2}\,\Gamma((n+1)/2)}{\Gamma(n/2)}.$$

Setting $Ew_\beta = \sigma$: $\beta = \frac{\Gamma(n/2)}{\sqrt{2}\,\Gamma((n+1)/2)}$.

So $T_2 = \frac{\Gamma(n/2)}{\sqrt{2}\,\Gamma((n+1)/2)}(\sum x_i^2)^{1/2}$ is also an unbiased estimator of σ.

$$\text{Var}(T_2) = \frac{1}{2}\left[\frac{\Gamma(n/2)}{\Gamma((n+1)/2)}\right]^2 \text{Var}\Big(\sum x_i^2\Big)^{1/2}.$$

It can be shown that $V(T_2) > \sigma^2/(2n)$, and $\text{Var}(T_2) < \text{Var}(T_1)$. So T_2 is more efficient than T_1.

Efficiency of Estimators

Concept 8.2

Let T_1 and T_2 be two unbiased estimators of a parameter $g(\theta)$. Let $ET_1^2 < \infty$ and $ET_2^2 < \infty$. Define the **efficiency** of T_2 relative to T_1 by

$$e_{ff_\theta}(T_2|T_1) = \frac{\text{Var}_\theta(T_2)}{\text{Var}_\theta(T_1)}.$$

We say T_2 is more efficient than T_1 if $e_{ff}(T_2|T_1) < 1$.
We can also define the efficiency of an unbiased estimator w.r.t. FRCLB:

$$Ef(T) = \frac{\text{Var}(T)}{\text{FRCLB}}.$$

If $\lim_{n\to\infty} Ef(T) = 1$, we say T is **asymptotically efficient**.

Example 8.2

1. $X \sim \text{Poi}(\lambda)$; parameter of interest is $P(X = 0) = e^{-\lambda}$. FRCLB for $e^{-\lambda}$:

$$[g'(\lambda)]^2 \cdot \{\text{FRCLB for } \lambda\} = \lambda e^{-2\lambda} \cdot \frac{\lambda}{n} = \frac{\lambda^2 e^{-2\lambda}}{n}.$$

Wait–actually FRC LB $= [g'(\lambda)]^2/I_x(\lambda) = \lambda^2 e^{-2\lambda}/(n/\lambda) = \lambda e^{-2\lambda}/n \cdot \lambda = \lambda e^{-2\lambda}/n$.
More carefully: $g(\lambda) = e^{-\lambda}$, $g'(\lambda) = -e^{-\lambda}$, $I_x(\lambda) = n/\lambda$.

$$\text{FRCLB} = \frac{[g'(\lambda)]^2}{I_x(\lambda)} = \frac{e^{-2\lambda}}{n/\lambda} = \frac{\lambda e^{-2\lambda}}{n}.$$

Consider an estimator $\beta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x = 1, 2, \ldots \end{cases}$. Then $E\beta(x) = e^{-\lambda}$, so $\beta(x)$ is unbiased for $e^{-\lambda}$.

$E\beta^2(x) = e^{-\lambda} \Rightarrow \text{Var}(\beta(x)) = e^{-\lambda} - e^{-2\lambda} > \lambda e^{-2\lambda}/n$ (FRCLB not attained).

This is equivalent to $e^\lambda > 1 + \lambda$, which is always true. Hence FRCLB is not attained. We can actually show $\beta(x)$ is the only unbiased estimator: let $d(x)$ be any unbiased estimator of $e^{-\lambda}$:

$$\sum_{x=0}^{\infty} d(x) \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \Rightarrow d(0) + d(1)\lambda + d(2)\frac{\lambda^2}{2!} + \cdots = 1.$$

Comparing coefficients: $d(0) = 1$, $d(1) = d(2) = \cdots = 0$. So $d(x) = \beta(x)$ for all x, and $\beta(x)$ is UMVUE.

9 Exponential Family

Concept 9.1: One-Parameter Exponential Family

A distribution belongs to the **one-parameter exponential family** if its pdf/pmf can be written as:

$$f(x,\theta) = c(\theta)\, h(x)\, e^{Q(\theta)\, T(x)}.$$

Example 9.1

1. $X \sim \text{Bin}(n,p)$, n known:

$$f(x,p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (1-p)^n \left(\frac{p}{1-p}\right)^x = (1-p)^n \binom{n}{x} e^{x \log(p/(1-p))}.$$

So the binomial distribution (with n known) is in the exponential family.

2. $X \sim \text{Poi}(\lambda)$:

$$f(x,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \cdot \frac{1}{x!} \cdot e^{x \log \lambda}.$$

Concept 9.2: Multiparameter Exponential Family

$$f(x,\theta) = c(\theta)\, h(x) \exp\left\{\sum_{i=1}^{K} Q_i(\theta)\, T_i(x)\right\}.$$

Example 9.2

1. $X \sim \mathcal{N}(\mu, \sigma^2)$, both unknown:

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)} \\ &= \frac{e^{-\mu^2/(2\sigma^2)}}{\sigma} \cdot \frac{1}{\sqrt{2\pi}}\, e^{-x^2/(2\sigma^2)+\mu x/\sigma^2}. \end{aligned}$$

This is a two-parameter exponential distribution family.

Remark 9.1: One-parameter exponential family and FRCLB

For the one-parameter exponential family $f(x,\theta) = c(\theta)\, h(x)\, e^{Q(\theta)T(x)}$:

$$\log f(x,\theta) = \ln c(\theta) + \ln h(x) + Q(\theta)T(x).$$

$$\frac{\partial \log f(x,\theta)}{\partial\theta} = \frac{c'(\theta)}{c(\theta)} + Q'(\theta)T(x).$$

$$s(x,\theta) = \sum_{i=1}^{n} \frac{\partial \log f(x_i,\theta)}{\partial\theta} = n\frac{c'(\theta)}{c(\theta)} + Q'(\theta)\sum_{i=1}^{n} T(x_i).$$

Thus $w = \frac{1}{n}\sum_{i=1}^{n} T(x_i)$ is linearly related to $s(\underline{x},\theta)$ wp 1. Hence any linear function of w will attain the FRCLB for the variance of an unbiased estimator of $E(w)$.

We also determine $E(w)$:

$$\int f(x,\theta)\,d\nu(x) = 1 \;\Rightarrow\; \int c(\theta)\,h(x)\,e^{Q(\theta)T(x)}\,d\nu(x) = 1.$$

Differentiating under the integral sign:

$$c'(\theta)\int h(x)\,e^{Q(\theta)T(x)}\,d\nu(x) + c(\theta)\int h(x)\,Q'(\theta)T(x)\,e^{Q(\theta)T(x)}\,d\nu(x) = 0.$$

$$\frac{c'(\theta)}{c(\theta)} + Q'(\theta)\,E_\theta\,T(x) = 0 \;\Rightarrow\; E_\theta\,T(x) = -\frac{c'(\theta)}{Q'(\theta)\,c(\theta)}.$$

Example 9.3: Geometric distribution

$f(x,\theta) = \theta(1-\theta)^x$, $x = 0, 1, \ldots$, $0 < \theta < 1$.

$$= \theta\,e^{x\log(1-\theta)}.$$

$c(\theta) = \theta$, $h(x) = 1$, $T(x) = x$, $Q(\theta) = \log(1-\theta)$.

$$E(w) = E(\bar{X}) = -\frac{c'(\theta)}{Q'(\theta)\,c(\theta)} = \frac{1-\theta}{\theta} - 1 = \frac{1}{\theta} - 1.$$

So $E(\bar{X}) = 1/\theta - 1$, and $\mathrm{Var}(\bar{X})$ will be the same as FRCLB for the estimator of $1/\theta - 1$. Then $\bar{X} - 1$ is UMVUE for $1/\theta$.

10 Bhattacharyya Bound

Theorem 10.1: Bhattacharyya Bound

Let $X_1, \ldots, X_n$ be a random sample from a population with pdf (pmf) $f(x,\theta)$, $\theta \in \Theta$.

Let $S_i = \left(\frac{\partial^i}{\partial\theta^i}\prod_{j=1}^n f(x_j;\theta)\right) \Big/ \prod_{j=1}^n f(x_j;\theta)$, $i = 1, \ldots, K$.

Let the following conditions hold:

1. $\frac{\partial^i f(x,\theta)}{\partial\theta^i}$ exists for all $\theta \in \Theta$ and almost all x.
2. $\int\cdots\int \prod_{j=1}^n f(x_j;\theta)\,d\nu(\underline{x})$ can be differentiated under the integral sign i times.
3. $S_1, \ldots, S_K$ are linearly dependent.
4. $\int\cdots\int S(\underline{x})\prod_{j=1}^n f(x_j;\theta)\,d\nu(\underline{x})$ can be differentiated under the integral sign i times $(i = 1, \ldots, K)$ for any integrable function S.

Let $\lambda_{ij} = \mathrm{Cov}(S_i, S_j)$, $i \neq j = 1, \ldots, K$; $\lambda_{ii} = \mathrm{Var}(S_i)$.

Let $\Lambda = [\lambda_{ij}]_{i,j=1,\ldots,K}$, λ^{rs} denote the (r,s)-th term of Λ^{-1}, and $\eta_i = \mathrm{Cov}(T, S_i) = d^i g/d\theta^i$ where $E_\theta T(x) = g(\theta)$. Let $\eta' = (\eta_1, \ldots, \eta_K)$. Then the **Bhattacharyya bound** is:

$$\mathrm{Var}_\theta(T) \geq \eta'\,\Lambda^{-1}\eta = \sum_r\sum_s \lambda^{rs}\,\frac{d^r g}{d\theta^r}\,\frac{d^s g}{d\theta^s}.$$

For $K = 1$, this reduces to FRCLB.

Proof:

$E_\theta T(\underline{x}) = g(\theta)$:

$$\int \cdots \int T(\underline{x}) \prod_{j=1}^{n} f(x_j;\theta)\, d\nu(\underline{x}) = g(\theta).$$

Differentiating i times w.r.t. θ:

$$\int \cdots \int T(\underline{x})\, S_i \prod_{j=1}^{n} f(x_j;\theta)\, d\nu(\underline{x}) = \frac{d^i g}{d\theta^i},$$

$$\Rightarrow \quad E(T(\underline{x})\, S_i) = \frac{d^i g}{d\theta^i}.$$

Also $E(S_i) = 0$, so $\mathrm{Cov}(T, S_i) = \dfrac{d^i g}{d\theta^i}$.
Consider the multiple correlation coefficient between T and $(S_1, \ldots, S_K)$:

$$R^2 = \frac{\eta' \Lambda^{-1} \eta}{\mathrm{Var}(T)} \leq 1.$$

So $\mathrm{Var}(T) \geq \eta' \Lambda^{-1} \eta$. □

Remark 10.1: Remarks on Bhattacharyya Bound

1. Equality in Bhattacharyya bound is attained iff T is linearly related to $S_1, S_2, \ldots, S_K$ wp 1.
2. Bhattacharyya bound (BhLB) is sharper than FRCLB since the multiple correlation coefficient between T and $(S_1, \ldots, S_K)$ is larger than the correlation between T and S_1.
3. BhLB gets sharper as K increases, because the multiple correlation coefficient between T and $(S_1, \ldots, S_{K+1})$ is larger than that between T and $(S_1, \ldots, S_K)$.
4. **Note:** Although BhLB is sharper, its use is limited due to complications in evaluating higher order derivatives.

Example 10.1

$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)}.$$

$$\frac{\partial f}{\partial \sigma^2} = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)} \left[\frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right].$$

$$S_1 = \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} = \frac{1}{2\sigma^2}\left[\frac{(x-\mu)^2}{\sigma^2} - 1\right].$$

Let $w = (x-\mu)^2/\sigma^2 \sim \chi^2_1$.

$$ES_1^2 = \mathrm{Var}S_1 = \frac{1}{4\sigma^4}\, E(w-1)^2 = \frac{2}{4\sigma^4} = \frac{1}{2\sigma^4}.$$

FRCLB for variance of unbiased estimator of σ^2 is $2\sigma^4/n$.

$$\frac{\partial^2 f}{\partial(\sigma^2)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \left[\frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right]^2 + \cdots$$

$$S_2 = \frac{1}{4\sigma^4}(w^2 - 6w + 3).$$

$$ES_2^2 = \mathrm{Var}S_2 = \frac{1}{16\sigma^8} E(w^2 - 6w + 3)^2 = \frac{33}{32\sigma^8}.$$

$$E(S_1S_2) = \mathrm{Cov}(S_1, S_2) = -\frac{2}{\sigma^6(1-\sigma^6)} = -\frac{2}{\sigma^6(1-\sigma)} \approx -\frac{2}{\sigma^3(1-\sigma)}.$$

(Detailed calculation shows $E(S_1S_2) = -2/(\sigma^3(1-\sigma))$.)
The variance-covariance matrix of (S_1, S_2):

$$\Lambda = \begin{pmatrix} \frac{1}{2\sigma^4} & -\frac{2}{\sigma^3(1-\sigma)} \\ -\frac{2}{\sigma^3(1-\sigma)} & \frac{33}{32\sigma^8} \end{pmatrix}, \quad |\Lambda| = \frac{4}{\sigma^6(1-\sigma)^3}.$$

$$\Lambda^{-1} = \begin{pmatrix} (2-\sigma)\sigma^2(1-\sigma) & \sigma^3(1-\sigma)^2/2 \\ \sigma^3(1-\sigma)^2/2 & \sigma^4(1-\sigma)^3/4 \end{pmatrix},$$

$$\eta_1 = \frac{dg}{d\theta} = 1, \qquad \eta_2 = \frac{d^2g}{d\theta^2} = 0.$$

(Here $g(\sigma^2) = \sigma^2$, so $\eta' = (1, 0)$.)
Bhattacharyya lower bound for estimating σ^2 unbiasedly:

$$\mathrm{BhLB} = \eta'\Lambda^{-1}\eta = \sigma^2(2-\sigma)(1-\sigma).$$

$$\mathrm{FRCLB} = \sigma^2(1-\sigma) = E(1/S_1^2).$$

$$\mathrm{Var}(T) = \sigma(1-\sigma) > \mathrm{BhLB} > \mathrm{FRCLB}.$$

11 Chapman–Robbins–Kiefer (CRK) Inequality

Theorem 11.1: CRK Inequality

Let x have pdf (pmf) $f(x, \theta)$, $\theta \in \Theta$. Let T be an unbiased estimator of $g(\theta)$. Define

$$A(\phi, \theta) = \mathrm{Var}_\theta\left[\frac{f(x,\phi)}{f(x,\theta)}\right], \quad \phi \neq \theta \;\&\; \{x : f(x,\phi) > 0\} \subset \{x : f(x,\theta) > 0\}.$$

The CRK inequality states:

$$\mathrm{Var}_\theta T \geq \sup_{\phi\in\Theta} \frac{\{g(\phi) - g(\theta)\}^2}{A(\phi,\theta)},$$

where the sup is taken over all ϕ for which the above condition satisfies.

Proof:

$$\begin{aligned} g(\phi) - g(\theta) &= E_\phi\, T(\underline{x}) - E_\theta\, T(\underline{x}) \\ &= \int T(\underline{x})(f(\underline{x},\phi) - f(\underline{x},\theta))\; d\nu(\underline{x}) \\ &= \int T(\underline{x})\, f(\underline{x},\theta)\left[\frac{f(\underline{x},\phi)}{f(\underline{x},\theta)} - 1\right] d\nu(\underline{x}). \end{aligned}$$

Now $E_\theta\left[\dfrac{f(\underline{x},\phi)}{f(\underline{x},\theta)}\right] = \displaystyle\int \frac{f(\underline{x},\phi)}{f(\underline{x},\theta)}\, f(\underline{x},\theta)\, d\nu(\underline{x}) = 1$, so $E_\theta\left[\dfrac{f(\underline{x},\phi)}{f(\underline{x},\theta)} - 1\right] = 0.$

Therefore $g(\phi) - g(\theta) = \mathrm{Cov}_\theta\left(T(\underline{x}),\ \dfrac{f(\underline{x},\phi)}{f(\underline{x},\theta)}\right).$

$$\begin{aligned} \{g(\phi) - g(\theta)\}^2 &= \mathrm{Cov}^2_\theta\left(T,\ \frac{f(\underline{x},\phi)}{f(\underline{x},\theta)}\right) \le \mathrm{Var}_\theta\, T \cdot \mathrm{Var}_\theta\left[\frac{f(\underline{x},\phi)}{f(\underline{x},\theta)}\right] \\ &= \mathrm{Var}_\theta\, T \cdot A(\phi,\theta). \end{aligned}$$

$$\Rightarrow \quad \mathrm{Var}_\theta\, T \ge \frac{\{g(\phi) - g(\theta)\}^2}{A(\phi,\theta)}.$$

Taking the sup on RHS over all ϕ subject to the condition on ϕ:

$$\mathrm{Var}_\theta\, T \ge \sup_\phi \frac{\{g(\phi) - g(\theta)\}^2}{A(\phi,\theta)}. \quad \square$$

$\square$

Example 11.1

1. $X \sim \mathrm{Unif}(0,\theta)$, $\theta > 0$. Consider unbiased estimator of θ.

$$f(x,\theta) = \begin{cases} 1/\theta & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}, \qquad f(x,\phi) = \begin{cases} 1/\phi & 0 < x < \phi \\ 0 & \text{otherwise.} \end{cases}$$

$$\frac{f(x,\phi)}{f(x,\theta)} = \begin{cases} \theta/\phi & x < \phi < \theta \\ 0 & \phi < x < \theta \end{cases} \quad (\text{for } \phi < \theta).$$

$$E_\theta\left[\frac{f(x,\phi)}{f(x,\theta)}\right] = \int_0^\phi \frac{\theta}{\phi} \cdot \frac{1}{\theta}\, dx = 1.$$

$$E_\theta\left[\frac{f(x,\phi)}{f(x,\theta)}\right]^2 = \int_0^\phi \frac{\theta^2}{\phi^2} \cdot \frac{1}{\theta}\, dx = \frac{\theta}{\phi}.$$

$$A(\phi,\theta) = \mathrm{Var}_\theta\left[\frac{f(x,\phi)}{f(x,\theta)}\right] = \frac{\theta}{\phi} - 1.$$

The CRKLB for the variance of an unbiased estimator of θ is:

$$\sup_{\phi<\theta} \frac{(\phi-\theta)^2}{\theta/\phi - 1} = \sup_{\phi<\theta} \frac{\phi(\theta-\phi)^2}{\theta-\phi} = \sup_{\phi<\theta} \phi(\theta-\phi).$$

We find $\sup_{\phi<\theta} \phi(\theta-\phi) = \theta^2/4$ attained at $\phi = \theta/2$ (since $\frac{d}{d\phi}[\phi(\theta-\phi)] = \theta - 2\phi = 0 \Rightarrow \phi = \theta/2$).
Consider $T = 2x$: $E\,T = \theta$, $\mathrm{Var}T = 4\,\mathrm{Var}x^0 = 4 \cdot \theta^2/12 = \theta^2/3 > \theta^2/4$.

2. $X \sim f(x,\theta) = e^{\theta - x}$, $x > \theta$ (one-parameter exponential shifted). We want CRKLB for unbiased estimator of θ.

$$f(x,\phi) = e^{\phi - x}, \quad x > \phi, \quad \text{for } \phi > \theta.$$

$$\frac{f(x,\phi)}{f(x,\theta)} = \begin{cases} e^{\phi-\theta} & x > \phi > \theta \\ 0 & \phi > x > \theta \end{cases}.$$

$$E_\theta\left[\frac{f(x,\phi)}{f(x,\theta)}\right] = \int_\phi^\infty e^{\phi-\theta} \cdot e^{\theta - x}\, dx = 1.$$

$$E_\theta\left[\frac{f(x,\phi)}{f(x,\theta)}\right]^2 = \int_\phi^\infty e^{2(\phi-\theta)} \cdot e^{\theta - x}\, dx = e^{\phi-\theta}.$$

$$A(\phi,\theta) = e^{\phi-\theta} - 1.$$

CRKLB:

$$\sup_{\phi>\theta} \frac{(\phi-\theta)^2}{e^{\phi-\theta}-1} = \sup_{t>0} \frac{t^2}{e^t - 1} \quad (\text{let } t = \phi - \theta).$$

$$h(t) = \frac{t^2}{e^t - 1}, \quad \lim_{t\to 0} \frac{2t}{e^t} = 0, \quad \lim_{t\to\infty} \frac{2t}{e^t} = 0.$$

$$h'(t) = \frac{t\{(2-t)e^t - 2\}}{(e^t-1)^2} < 0 \ \text{ for } t \geq 2, \quad > 0 \ \text{ for } t \leq 1.$$

We can solve numerically to get $t = 1.59362$, where $h(t) = 0.6476$. So CRKLB is 0.6476.
In this case, an unbiased estimator for θ is $T = x - 1$, $\mathrm{Var}(T) = \mathrm{Var}x = 1 > 0.6476$.

3. Both FRC and CRK lower bounds can be found for $X \sim \mathcal{N}(\theta, 1)$. Here FRCLB for estimating θ is $\sigma^2/n = 1$ (in this case).
For CRK:

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}}\, e^{-(x-\theta)^2/2}, \quad f(x,\phi) = \frac{1}{\sqrt{2\pi}}\, e^{-(x-\phi)^2/2}.$$

$$\begin{aligned} \frac{f(x,\phi)}{f(x,\theta)} &= \exp\Big\{-\frac{1}{2}\Big[(x-\phi)^2 - (x-\theta)^2\Big]\Big\} \\ &= e^{(\theta^2-\phi^2)/2}\, e^{(\phi-\theta)x}. \end{aligned}$$

$$\begin{aligned} E_\theta\left[\frac{f(x,\phi)}{f(x,\theta)}\right]^2 &= e^{\theta^2-\phi^2}\, E_\theta e^{2(\phi-\theta)x} \\ &= e^{\theta^2-\phi^2}\, M_x(2(\phi-\theta)) \\ &= e^{\theta^2-\phi^2}\, e^{2\theta(\phi-\theta)+2(\phi-\theta)^2} \\ &= e^{\theta^2-\phi^2} \cdot e^{2(\phi-\theta)\theta+2(\phi-\theta)^2} \\ &= e^{\theta^2-\phi^2} \cdot e^{(2\phi-\theta)\phi} = e^{(\phi-\theta)^2}. \end{aligned}$$

$$A(\phi,\theta) = \mathrm{Var}_\theta\left[\frac{f(x,\phi)}{f(x,\theta)}\right] = e^{\theta^2-\phi^2}\, e^{(\phi-\theta)^2} - 1 = e^{(\phi-\theta)^2} - 1.$$

$$\mathrm{CRKLB} = \sup_{\phi\in\mathbb{R}} \frac{(\phi-\theta)^2}{e^{(\phi-\theta)^2}-1} = \sup_{t\in\mathbb{R}} \frac{t^2}{e^{t^2}-1} \xrightarrow{t\to 0} 1.$$

CRKLB $= 1 =$ FRCLB (attained as $t \to 0$).

12 FRC Lower Bound in Higher Dimensions

Theorem 12.1: Matrix Form of FRCLB

Let $X_1, \ldots, X_n$ be a random sample from a population with density/mass function $f(x,\theta)$, $\theta = (\theta_1, \ldots, \theta_K) \in \Theta \subset \mathbb{R}^K$.
We consider estimation of parametric functions $g_1(\theta), \ldots, g_r(\theta)$. Let $E_\theta T_i(x) = g_i(\theta)$ for all $\theta \in \Theta$, i.e. $T_1, T_2, \ldots, T_r$ are unbiased for $g_1, g_2, \ldots, g_r$.
Let $T' = (T_1, \ldots, T_r)$. Define

$$\mathrm{Var}(T_i) = V_{ii}, \quad 1 \le i \le r, \qquad \mathrm{Cov}(T_i, T_j) = V_{ij}, \quad 1 \le i, j \le r,\ i \ne j.$$

V is the dispersion matrix of T.
Further define

$$\frac{\partial g_i}{\partial \theta_j} = \Delta_{ij}, \quad 1 \le i \le r,\ 1 \le j \le K.$$

$\Delta = (\Delta_{ij})_{r\times K}$.

$$\mathcal{I}_{ij} = E\left[-\frac{\partial^2 \log f(x,\theta)}{\partial\theta_i\, \partial\theta_j}\right], \quad 1 \le i, j \le K.$$

$\mathcal{I} = (\mathcal{I}_{ij})_{K\times K}$ is the **Fisher's Information matrix**.
Regularity Conditions

1. $\dfrac{\partial^2 f(x,\theta)}{\partial\theta_i\, \partial\theta_j}$ exists for all $1 \le i, j \le K$ and for all $\theta \in \Theta$ a.e.
2. $\displaystyle\int \cdots \int S(x) \prod_{j=1}^{n} f(x_j;\theta)\, d\nu(x)$ can be differentiated under the integral sign for any integrable function $S(x)$.
3. $\displaystyle -E\left[\frac{\partial^2 \log f(x,\theta)}{\partial\theta_i^2}\right] > 0$ for all $\theta \in \Theta$.

Under the above regularity conditions, $V - \Delta\mathcal{I}^{-1}\Delta'$ is a non-negative definite matrix.
In particular:

$$V_{ii} \ge \sum_s \sum_t \mathcal{I}^{st} \frac{\partial g_i}{\partial\theta_s} \frac{\partial g_i}{\partial\theta_t},$$

where $\mathcal{I}^{st}$ are terms in $\mathcal{I}^{-1}$.

Proof:

Consider $E_\theta T_i = g_i(\theta)$ for all $\theta \in \Theta \Leftrightarrow \displaystyle\int \cdots \int T_i(\underline{x}) \prod_{j=1}^{n} f(x_j;\theta)\, d\nu(\underline{x}) = g_i(\theta)$.

Differentiating w.r.t. θ_j:

$$\int \cdots \int T_i(\underline{x}) \left(\frac{\partial}{\partial \theta_j}\right) \prod_{j'=1}^{n} f(x_{j'};\theta)\, d\nu(\underline{x}) = \frac{\partial g_i}{\partial \theta_j} = \Delta_{ij}.$$

Also note $E\dfrac{\partial}{\partial \theta_j} \log f = ET_i \cdot \dfrac{1}{f}\dfrac{\partial f}{\partial \theta_j}$ etc., and $\displaystyle\int f(x,\theta)\, d\nu(x) = 1 \Rightarrow E\frac{1}{f}\frac{\partial f}{\partial \theta_j} = 0.$

Consider the dispersion matrix of $T_1, \ldots, T_r,\ \dfrac{1}{f}\dfrac{\partial f}{\partial \theta_1}, \ldots, \dfrac{1}{f}\dfrac{\partial f}{\partial \theta_K}$:

$$\begin{pmatrix} V & \Delta \\ \Delta' & \mathcal{I} \end{pmatrix}.$$

The determinants $\left|\begin{smallmatrix} \mathcal{I} & -\Delta\mathcal{I}^{-1} \\ 0 & \mathcal{I}^{-1} \end{smallmatrix}\right|$ and $\left|\begin{smallmatrix} V & \Delta \\ \Delta' & \mathcal{I} \end{smallmatrix}\right|$ are non-negative and their product is also non-negative:

$$|V - \Delta\,\mathcal{I}^{-1}\Delta'| \geq 0.$$

This statement remains true for a subset of $T_1, \ldots, T_r$, which means $V - \Delta\mathcal{I}^{-1}\Delta'$ is non-negative definite. □

Example 12.1

$X_i \sim \mathcal{N}(\mu, \sigma^2)$, μ, σ^2 unknown. $\theta = (\mu, \sigma^2)$, $g_1(\theta) = \mu$, $g_2(\theta) = \sigma^2$.

$$\log f(x;\mu,\sigma^2) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi) - \frac{(x-\mu)^2}{2\sigma^2}.$$

$$\frac{\partial \log f}{\partial \mu} = \frac{x-\mu}{\sigma^2}, \qquad \frac{\partial \log f}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4},$$

$$\frac{\partial^2 \log f}{\partial \mu^2} = -\frac{1}{\sigma^2}, \qquad \frac{\partial^2 \log f}{\partial (\sigma^2)^2} = -\frac{(x-\mu)^2}{\sigma^6} + \frac{1}{2\sigma^4}, \qquad \frac{\partial^2 \log f}{\partial \mu\, \partial \sigma^2} = -\frac{(x-\mu)}{\sigma^4}.$$

$$\mathcal{I}_{11} = \frac{1}{\sigma^2}, \quad \mathcal{I}_{22} = \frac{1}{2\sigma^4}, \quad \mathcal{I}_{12} = 0.$$

$$\mathcal{I} = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix}, \qquad \mathcal{I}^{-1} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}.$$

$$\mathrm{Var}(T_1) \geq \sigma^2/n \ \text{ if } E(T_1) = \mu, \qquad \mathrm{Var}(T_2) \geq 2\sigma^4/n \ \text{ if } E(T_2) = \sigma^2.$$

13 Sufficiency

Concept 13.1: Sufficient Statistic

Let $X_1, \ldots, X_n$ be a random sample from a population $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. A statistic $T = T(X_1, \ldots, X_n) = T(X)$ is said to be **sufficient** for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ is independent of θ (except perhaps a null set A, i.e. $P_\theta(T \in A) = 0$ for all $\theta \in \Theta$).

Concept 13.2: Data Reduction

This signifies that if one does not have the original observations $X_1, \ldots, X_n$ but has T and a knowledge of the known conditional distribution of $X_1, \ldots, X_n$ given $T = t$, then one can generate $x_1', \ldots, x_n'$ which will have the same distribution as $X_1, \ldots, X_n$. This is **data reduction**.

Example 13.1

Let $X_1, \ldots, X_n \sim \text{Ber}(p)$, $0 < p < 1$. Consider $T = \sum_{i=1}^n X_i$. The conditional distribution of $X_1, \ldots, X_n$ given $T = t$:

$$P(X_1 = x_1, \ldots, X_n = x_n \mid T = t) = \begin{cases} \dfrac{1}{\binom{n}{t}} & \text{if } t = \sum x_i \\ 0 & \text{if } t \neq \sum x_i \end{cases}$$

(derived by direct calculation using the Binomial pmf). This is independent of p. So $T = \sum X_i$ is sufficient for $\text{Ber}(p)$.

Example 13.2

Let $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$, $\lambda > 0$. Let $T = \sum X_i$.

$$P(X_1 = x_1, \ldots, X_n = x_n \mid T = t) = \begin{cases} \dfrac{t!}{x_1! \cdots x_{n-1}! (t - \sum_{i=1}^{n-1} x_i)!} & t = \sum x_i \\ 0 & t \neq \sum x_i \end{cases}$$

This is independent of λ. So $T = \sum X_i$ is sufficient for $\text{Poi}(\lambda)$.

Remark 13.1

1. Let T be sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and let T be a function of U. Then U is also sufficient for $\mathcal{P}$.
2. $P_\theta(X_1 = x_1, \ldots, X_n = x_n \mid X_1 = t_1, \ldots, X_n = t_n) = \begin{cases} 1 & \text{if } t = x \\ 0 & \text{if } t \neq x \end{cases}$ which is always free from the parameter. So $X = (X_1, \ldots, X_n)$ is always sufficient. This is called a **trivial sufficient statistic**.

Theorem 13.1

Let X have distribution P_θ, $\theta \in \Theta$ and let $T(X)$ be sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then given T, we can generate r.v. Y which has the same distribution P_θ.

Proof:

The conditional distribution of $X \mid T = t$ is independent of θ. After observing $T = t$, we conduct a random experiment with outcome y following the known conditional distribution of $X \mid T = t$. Then Y and X have the same distribution. □

13.1 Rao–Blackwell Theorem

Theorem 13.2: Rao–Blackwell

Let $X_1, \ldots, X_n$ be a random sample from a population with distribution P_θ, $\theta \in \Theta$. Let $S(X)$ be an unbiased estimator of $g(\theta)$ and $T(X)$ be sufficient. Then there exists an unbiased estimator based on T alone whose variance is not more than that of $S(X)$.

Proof:

Let $h(t) = E[S(X) \mid T(X) = t]$. Since T is sufficient, $h(t)$ is independent of θ.

$$E_\theta\, h(T) = E_\theta\, E[S(X) \mid T(X) = t] = E_\theta\, S(X) = g(\theta) \quad \forall\, \theta \in \Theta.$$

So $h(T)$ is unbiased for $g(\theta)$.

$$\begin{aligned} \mathrm{Var}_\theta(S(X)) &= E_\theta\{\mathrm{Var}(S(X) \mid T)\} + \mathrm{Var}_\theta\{E(S(X) \mid T)\} \\ &= \underbrace{E_\theta\{\mathrm{Var}(S(X) \mid T)\}}_{\geq 0} + \mathrm{Var}_\theta(h(T)). \end{aligned}$$

So $\mathrm{Var}_\theta(h(T)) \leq \mathrm{Var}_\theta(S(X))$. □

13.2 Neyman–Fisher Factorisation Theorem

Theorem 13.3: Neyman–Fisher Factorisation

For a discrete r.v. X with pmf $p(x, \theta)$, $\theta \in \Theta$, $T(X)$ is sufficient if and only if

$$p(x, \theta) = g(T(x), \theta)\, h(x) \qquad \forall\, \theta \in \Theta.$$

Proof:

($\Rightarrow$) Let $p(x, \theta) = g(T(x), \theta)\, h(x)$.

$$P_\theta(T(X) = t) = \sum_{x:\, T(x)=t} p(x, \theta) = \sum_{x:\, T(x)=t} g(T(x), \theta)\, h(x) = g(t, \theta) \sum_{x:\, T(x)=t} h(x).$$

For $T(x') = t$:

$$\begin{aligned} P_\theta(X = x' \mid T(X) = t) &= \frac{P_\theta(X = x',\, T(X) = T(x'))}{P_\theta(T(X) = t(x'))} \\ &= \frac{g(t, \theta)\, h(x')}{g(t, \theta) \sum_{x:\, T(x)=t} h(x)} = \frac{h(x')}{\sum_{x:\, T(x)=t} h(x)}, \end{aligned}$$

which is independent of θ. So T is sufficient.

($\Leftarrow$) Conversely, let T be sufficient. Then

$$P_\theta(X = x' \mid T(X) = t) = c(x', t) \quad \text{(independent of } \theta\text{)}.$$

$$P_\theta(X = x',\, T(X) = T(x')) = c(x', t)\, P(T(X) = T(x')).$$

$$\Rightarrow \quad p_\theta(x') = c(x', t)\, g(t, \theta) \cdot h(x)\, g(t, \theta).$$

(More precisely: $P_\theta(x') = c(x', T(x')) \cdot P_\theta(T(X) = T(x')) = c(x', t)\, g(t, \theta)$, taking $h(x) = c(x', t)$ and $g(t, \theta) = P_\theta(T(X) = t)$.) □

Remark 13.2

The theorem holds if θ and T are vectors with same or different dimensions.
If T is sufficient and T is a function of U of U, then U is also sufficient. However, if V is a function of T, then V need not be sufficient. V is sufficient if V is a one-to-one function of T.

Example 13.3

1. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$.
   Case I: σ^2 known (WLOG $\sigma^2 = 1$).

$$\begin{aligned} f(\underline{x}, \mu) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum(x_i - \mu)^2\right\} \\ &= (2\pi)^{-n/2} e^{-\sum x_i^2/2} \cdot e^{-n\mu^2/2 + n\mu\bar{x}} \\ &= h(\underline{x}) \cdot g(\bar{x}, \mu). \end{aligned}$$

   So $\bar{X}$ is sufficient for $\{N(\mu, 1) : \mu \in \mathbb{R}\}$.
   Case II: μ known, $\mu = \mu_0$.

$$f(\underline{x}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - \mu_0)^2\right\} = h(\underline{x}) \cdot g\left(\sum(x_i - \mu_0)^2, \sigma^2\right).$$

   So $T(X) = \sum(X_i - \mu_0)^2$ is sufficient for $\{N(\mu_0, \sigma^2) : \sigma^2 > 0\}$.
   Case III: Both μ and σ^2 unknown.

$$\begin{aligned} f(\underline{x}, \mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n\mu^2}{2\sigma^2} - \frac{\sum x_i^2}{2\sigma^2} + \frac{\mu\sum x_i}{\sigma^2}\right\} \\ &= h(\underline{x}) \cdot g\left(\sum x_i, \sum x_i^2, \mu, \sigma^2\right). \end{aligned}$$

   So $(\sum X_i, \sum X_i^2)$ is sufficient, and so is $(\bar{X}, S^2)$.
   Case IV: $\sigma = \mu$.

$$\begin{aligned} f(\underline{x}, \mu) &= (2\pi)^{-n/2}\mu^{-n} \exp\left\{-\frac{n}{2} - \frac{\sum x_i^2}{2\mu^2} - \frac{\sum x_i}{\mu}\right\} \\ &= h(\underline{x}) \cdot g\left(\sum x_i, \sum x_i^2, \mu\right). \end{aligned}$$

   So $(\sum X_i, \sum X_i^2)$ is sufficient for $\{N(\mu, \mu^2) : \mu > 0\}$.
2. $X_i \sim \lambda e^{-\lambda x_i}$, $x_i, \lambda > 0$.

$$f(\underline{x}, \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}.$$

   By FT, $\sum X_i$ is sufficient.
3. $X_i \sim e^{\theta - x_i}$, $x_i > \theta$.

$$f(\underline{x}, \theta) = e^{-\sum x_i} e^{n\theta}\, \mathbf{1}(x_{(1)}, \theta) \prod_{i=2}^n \mathbf{1}(x_{(i)}, x_{(1)}) = g(x_{(1)}, \theta)\, h(\underline{x}).$$

   So $x_{(1)}$ is sufficient.

4. $X_i \sim \frac{1}{\sigma} e^{-(x_i-\mu)/\sigma}$, $x_i > \mu$, $\mu \in \mathbb{R}$, $\sigma > 0$.

$$f(\underline{x}, \mu, \sigma) = \frac{1}{\sigma^n} \exp\left\{\frac{n\mu}{\sigma} - \frac{\sum x_i}{\sigma}\right\} \mathbf{1}(x_{(1)}, \mu) \prod_{i=2}^{n} \mathbf{1}(x_{(i)}, x_{(1)}).$$

So $(x_{(1)}, \sum x_i)$ is sufficient, and so is $(x_{(1)}, \bar{X})$.

5. $X_i \sim \frac{1}{2} e^{-|x_i-\theta|}$, $x_i \in \mathbb{R}$, $\theta \in \mathbb{R}$ (Laplace).

$$\begin{aligned} f(\underline{x}, \theta) &= \frac{1}{2^n} \exp\Big\{-\sum |x_i - \theta|\Big\} = \frac{1}{2^n} \exp\Big\{-\sum |x_{(i)} - \theta|\Big\} \\ &= g(x_{(1)}, \ldots, x_{(n)}, \theta)\, h(\underline{x}). \end{aligned}$$

So $(x_{(1)}, \ldots, x_{(n)})$ is sufficient. $\hat{\theta}_{\text{MLE}}$ = median = function of $x_{(i)}$'s.

6. $X_i \sim \text{Unif}(0, \theta)$, $\theta > 0$:

$$\begin{aligned} f(\underline{x}, \theta) &= \frac{1}{\theta^n} \mathbf{1}(x_{(n)}, \theta) \prod_{i=1}^{n-1} \mathbf{1}(x_{(i)}, x_{(n)}) \\ &= g(x_{(n)}, \theta)\, h(\underline{x}). \end{aligned}$$

So $x_{(n)}$ is sufficient.

7. $X_i \sim \text{Unif}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$:

$$\begin{aligned} f(\underline{x}, \theta) &= \mathbf{1}(x_{(1)}, \theta - \tfrac{1}{2})\, \mathbf{1}(x_{(n)}, \theta + \tfrac{1}{2}) \prod_{i=2}^{n-1} \mathbf{1}(x_{(i)}, x_{(n)}) \\ &= g(x_{(1)}, x_{(n)}, \theta)\, h(\underline{x}). \end{aligned}$$

So $(x_{(1)}, x_{(n)})$ is sufficient.

Example 13.4: Exponential Family and Sufficiency

Let $f(x, \theta) = c(\theta)\, h(x) \exp\left\{\sum_{j=1}^{K} Q_j(\theta)\, T_j(x)\right\}$.

Based on a random sample $X_1, \ldots, X_n$:

$$\begin{aligned} f(\underline{x}, \theta) &= c^n(\theta) \prod_{j=1}^{n} h(x_j) \exp\left\{\sum_{j=1}^{K} Q_j(\theta) \sum_{i=1}^{n} T_j(x_i)\right\} \\ &= g\Big(\sum T_1(x_i), \ldots, \sum T_K(x_i)\Big)\, h(\underline{x}). \end{aligned}$$

So $\left(\sum_{j=1}^{n} T_1(x_j), \ldots, \sum_{j=1}^{n} T_K(x_j)\right)$ is sufficient.

13.3 Relation between Sufficiency and Information

Theorem 13.4: Information Mesaure and Sufficiency

Let $T(X)$ have pdf/pmf $\phi(t,\theta)$. Suppose $\dfrac{\partial}{\partial\theta}\phi(t,\theta)$ exists.

$$\frac{d}{d\theta}\int_B \phi(t,\theta)\,d\mu(t) = \int_B \frac{\partial\phi(t,\theta)}{\partial\theta}\,d\mu(t) \quad \text{for any measurable set } B.$$

Then:

1. $E\left[\dfrac{\partial \log f(x,\theta)}{\partial\theta}\,\middle|\, T=t\right] = \dfrac{\partial\log\phi(t,\theta)}{\partial\theta}$ almost everywhere.
2. $I_X(\theta) \geq I_T(\theta)$ with equality holding iff T is sufficient.

Proof of (i):

Let $\mathcal{X}$ be the space of x-values and $\mathcal{Y}$ the space of T-values. $T:\mathcal{X}\to\mathcal{Y}$.
For any set $B\in\mathcal{B}$:

$$\begin{aligned} E\left[\frac{\partial\log f(x,\theta)}{\partial\theta}\,I_B(x)\right] &= \int_B \frac{\partial f(x,\theta)}{\partial\theta}\cdot\frac{1}{f(x,\theta)}\cdot f(x,\theta)\,d\mu(x) \\ &= \frac{d}{d\theta}\int_B f(x,\theta)\,d\mu(x) = \frac{d}{d\theta}P(x\in B). \end{aligned}$$

Similarly:

$$\begin{aligned} &= \frac{d}{d\theta}P(T\in C) = \frac{d}{d\theta}\int_C \phi(t,\theta)\,d\mu(t) = \int_C \frac{\partial\phi(t,\theta)}{\partial\theta}\,d\mu(t) \\ &= E\left[\frac{\partial\log\phi(T,\theta)}{\partial\theta}\,I_C(T)\right] \\ &= E\left[\frac{\partial\log\phi(T,\theta)}{\partial\theta}\,I_B(x)\right]. \end{aligned}$$

By the definition of conditional expectation:

$$E\left[\frac{\partial\log f(x,\theta)}{\partial\theta}\,\middle|\, T=t\right] = \frac{\partial\log\phi(t,\theta)}{\partial\theta} \quad \text{a.e.} \quad \square$$

Remark. A function $g(t)$ is said to be $E(Y\mid T=t)$ if $E[Y\,I_B(T)] = E[g(T)\,I_B(T)]$ for all $B\in\mathcal{B}$. $\square$

Proof of (ii):

Consider

$$E\left[\frac{\partial\log\phi(T,\theta)}{\partial\theta}\cdot\frac{1}{\phi(T,\theta)} - \frac{\partial\log f(x,\theta)}{\partial\theta}\cdot\frac{1}{f(x,\theta)}\right]^2 \geq 0.$$

$$\Rightarrow \quad I_T(\theta) + I_X(\theta) - 2E\left[\frac{\partial\log\phi(T,\theta)}{\partial\theta}\cdot\frac{\partial\log f(x,\theta)}{\partial\theta}\right] \geq 0.$$

By (i):

$$E\left[\frac{\partial \log \phi(T,\theta)}{\partial\theta}\cdot\frac{\partial \log f(x,\theta)}{\partial\theta}\right] = E\,E\left[\frac{\partial \log \phi(T,\theta)}{\partial\theta}\cdot\frac{\partial \log f(x,\theta)}{\partial\theta}\,\middle|\,T\right]$$
$$= E\left[\frac{\partial \log \phi(T,\theta)}{\partial\theta}\right]^2 = I_T(\theta).$$

$$\Rightarrow \quad I_T(\theta) + I_X(\theta) - 2I_T(\theta) \geq 0 \;\Rightarrow\; I_X(\theta) \geq I_T(\theta). \quad \square$$

Necessary and sufficient condition for equality: $I_X(\theta) = I_T(\theta)$:

$$\frac{\partial \log \phi(t,\theta)}{\partial\theta} = \frac{\partial \log f(x,\theta)}{\partial\theta} \;\Leftrightarrow\; \log\phi(t,\theta) = \log f(x,\theta) + k(x)$$
$$\Leftrightarrow\; f(x,\theta) = \phi(t,\theta)\,h(x)$$
$$\Leftrightarrow\; T \text{ is sufficient.} \quad \square$$

$\square$

Example 13.5

1. $X_1,\ldots,X_n \sim \text{Poi}(\lambda)$. Let $T_1 = X_1$, $T_2 = X_1 + X_2$, …, $T_n = X_1 + \cdots + X_n$. We know $T_k \sim \text{Poi}(k\lambda)$ for $1 \leq k \leq n$.

   $$\frac{\partial \log f(x,\lambda)}{\partial\lambda} = \frac{x-\lambda}{\lambda}, \quad I_{T_1}(\lambda) = \frac{1}{\lambda}, \quad I_X(\lambda) = \frac{n}{\lambda}, \quad I_{T_2}(\lambda) = \frac{2}{\lambda},\ldots, I_{T_n}(\lambda) = \frac{n}{\lambda}.$$

   Observe that $I_X(\lambda) = I_{T_n}(\lambda)$ as T_n is sufficient.
   Information is additive: Let X and Y be independent r.v.'s with distributions $f_1(x,\theta)$ and $f_2(y,\theta)$. Then $I_{X+Y}(\theta) = I_X(\theta) + I_Y(\theta)$.
2. $X_1,\ldots,X_n \sim \mathcal{N}(\mu,1)$. Consider $T_1 = X_1 - X_2$, $T_2 = \sum X_i$. $T_2 \sim \mathcal{N}(n\mu, n)$.

   $$\frac{\partial \log f(t_2)}{\partial\mu} = t_2 - n\mu, \quad I_{T_2}(\mu) = E(t_2 - n\mu)^2 = n.$$

   $I_X(\mu) = n = I_{T_2}(\mu)$: T_2 is sufficient.
   $T_1 \sim \mathcal{N}(0,2)$: $\dfrac{\partial \log f(t_1)}{\partial\mu} = 0 \Rightarrow I_{T_1}(\mu) = 0$. T_1 is an **ancillary statistic**.

14 Minimal Sufficiency

Further Results on Sufficiency

For $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$: $(\bar{X}, S^2)$ is sufficient, and also $(X_1+X_2, X_3+\cdots+X_n, (X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2,\ldots)$ is sufficient.. So we need a better method of data reduction

Concept 14.1: Partition and Induced Partition

A **partition** of a space $\mathcal{X}$ is a collection $\{E_i\}$ of subsets of $\mathcal{X}$ such that:
1. $\bigcup_i E_i = \mathcal{X}$,

2. $E_i \cap E_j = \emptyset$ for $i \neq j$.

The sets E_i are called partition sets.
Let $T : \mathcal{X} \to \mathcal{Y}$. The partition of $\mathcal{X}$ induced by the function T is the collection of sets $T_y = \{x : T(x) = y\}$.

Example 14.1: Partition Reduction

Let $\mathcal{X} = \{\text{red, white, green, blue, yellow, violet}\}$.
Define a function T_1 such that:

$$T_1(\text{red}) = T_1(\text{white}) = 1$$
$$T_1(\text{green}) = T_1(\text{blue}) = 2$$
$$T_1(\text{yellow}) = T_1(\text{violet}) = 3$$

The partition induced by T_1 is:

$$P_1 = \{\{\text{red, white}\}, \{\text{green, blue}\}, \{\text{yellow, violet}\}\}$$

Now, let $T_2(X) = (T_1(X) - 2)^2$. The partition induced by T_2 is:

$$P_2 = \{\{\text{green, blue}\}, \{\text{red, white, yellow, violet}\}\}$$

We say that P_2 is a reduction of P_1 if each partition set of P_2 is a union of some members of P_1. In this case, P_2 is indeed a reduction of P_1. Furthermore, notice that T_2 is a function of T_1.

Example 14.2

1. $\mathcal{X} = \mathbb{R}$, $T(x) = x^2$, $\mathcal{Y} = [0, \infty)$. $T_y = \{-\sqrt{y}, \sqrt{y}\}$ for $y > 0$, $T_0 = \{0\}$. Then $\{T_0, T_y : y > 0\}$ is a partition of $\mathbb{R}$ induced by $T(x) = x^2$.
2. $\mathcal{X} = \{a, b, \ldots, y, z\}$. $T(a) = T(e) = T(i) = T(o) = T(u) = 1$, $T(b) = T(c) = \cdots = T(z) = 2$. Then $\{\{a, e, i, o, u\}, \{b, c, \ldots, z\}\}$ is the partition induced by T.

Remark 14.1

A function induces a partition. However, given a partition we may not define a function uniquely. However, if two functions T_1 and T_2 induce the same partition, then they must be one-to-one functions of each other.

Concept 14.2: Partition Reduction

We say that P_2 is a **reduction** of P_1 if each partition set of P_2 is a union of some member of P_1.
If T_i induces partition P_i, $i = 1, 2$ and P_2 is a reduction of P_1, then T_2 is a function of T_1.
Let T be sufficient. Then the conditional distribution of x given $\{x : T(x) = t\}$ is independent of the parameter.

Example 14.3

$\Theta = \{\theta_1, \theta_2, \theta_3\}$, $\mathcal{X} = \{x_1, x_2, x_3\}$.

	x_1	x_2	x_3
θ_1	0.1	0.2	0.7
θ_2	0.2	0.4	0.4
θ_3	0.3	0.6	0.1

Let $P = \{\{x_1, x_2\}, \{x_3\}\}$.
Conditional distribution of X given partition sets $A = \{x_1, x_2\}$ and $B = \{x_3\}$:

$$P_{\theta_1}(x = x_1 \mid x \in A) = \frac{P_{\theta_1}(x = x_1)}{P_{\theta_1}(x \in A)} = \frac{0.1}{0.1 + 0.2} = \frac{1}{3}.$$

$$P_{\theta_2}(x = x_1 \mid x \in A) = \frac{0.2}{0.2 + 0.4} = \frac{1}{3}, \quad P_{\theta_3}(x = x_1 \mid x \in A) = \frac{1}{3}.$$

Similarly $P_{\theta_i}(x = x_2 \mid x \in A) = 2/3$ for all i. And $P_{\theta_i}(x = x_3 \mid x \in B) = 1$ for all i.
So P is a sufficient partition.
Let $T(x_1) = T(x_2) = a$ and $T(x_3) = b$ with $a \neq b$. Then T is sufficient.
Now consider $P^* = \{\{x_1\}, \{x_2, x_3\}\}$:

$$P_{\theta_1}(x = x_2 \mid x \in \{x_2, x_3\}) = \frac{0.2}{0.2 + 0.7} = \frac{2}{9}, \quad P_{\theta_2}(x = x_2 \mid x \in \{x_2, x_3\}) = \frac{0.4}{0.4 + 0.4} = \frac{1}{2}.$$

So P^* is *not* a sufficient partition.

Remark 14.2

1. A sufficient statistic induces a sufficient partition and conversely given a sufficient partition we can define a sufficient statistic.

Concept 14.3: Minimal Sufficient Statistic

A sufficient statistic T is called **minimal sufficient** if the partition induced by T is a reduction of the partition induced by any other sufficient statistic.
Equivalently, T is minimal sufficient if T is a function of every other sufficient statistic.