Statistical Theory

Krish Nigam [*]

March, 2025

Statistics is a set of principles and procedures for collecting and processing quantitative data in order to help make decisions, and in this document, we present a rigorous treatment of the mathematical theory behind statistical decision making and statistical inference, both from a frequentist viewpoint and Bayesian viewpoint. Inference here includes estimating parameters and testing hypotheses.

In this course, we assume familiarity with the basics of probability theory and statistics. This document constitutes my notes taken from lectures by *Dr Kolyan Ray* at *Imperial College London.*

Contents

[*] Email: krish.nigam@princeton.edu

§1 Principles of Point Estimation

This course will concern **parametric inference**. We assume that we have a random variable X drawn from a member of a known family of distributions, however the parameter of the distribution is unknown and we aim to estimate it from the data. For example, we may know $X \sim \text{Pois}(\lambda)$ for some unknown $\lambda > 0$ and we wish to estimate λ.

Typically, we repeat the experiment multiple times, and hence observe $X_1, \ldots, X_n$ independent and identically distributed (i.i.d) copies of X. If we know that the true distribution is P_θ for some $\theta \in \Theta$ (Θ, the parameter space), some of the main goals of statistical inference about the parameter θ are:

- **Estimation**, *i.e.* construct an estimate $\hat{\theta}\left(X_1, \ldots, X_n\right)$ of the true value of θ.
- **Hypothesis testing**, *i.e.* construct a test to determine between two (or more) hypotheses concerning θ, *e.g.* whether $\theta = 0$ or not.
- **Uncertainty quantification**, *i.e.* give a set of plausible values for θ, *e.g.* $\hat{C} = [\hat{\theta}_1(X_1, \ldots, X_n), \hat{\theta}_2(X_1, \ldots, X_n)]$.

§1.1 Statistical models and estimators

Consider a random variable X taking values in some sample space $\mathcal{X}$ (*e.g.* $\mathcal{X} = \mathbb{R}$ or $\mathbb{R}^k$), coming from some probability distribution P_θ, which is parameterised by an unknown parameter $\theta \in \mathbb{R}^p$.

Definition 1.1 (Statistical model) — A **statistical model** for X is any family $\{P_\theta : \theta \in \Theta\}$ of probability distributions P_θ for the distribution of X. The set Θ is called the parameter space.

Remark 1.2 If X is discrete/continuous with pmf/pdf f_θ, one can equivalently write the statistical model as $\{f_\theta : \theta \in \Theta\}$.

In this course, we will generally take $\Theta \subseteq \mathbb{R}^p$, so that θ can be a scalar or vector of parameters. When p is finite, this is known as a **parametric model**, but there also exist nonparametric models where Θ is infinite-dimensional, but these are beyond the scope for this course.

Example 1.3 If $X = (X_1, \ldots, X_n)$ with $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, then the joint pdf of X is

$$f_{\mu,\sigma^2}(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

and here $\theta = (\mu, \sigma^2)$ and the parameter space is $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 \geq 0\}$.

A necessary condition to estimate θ based on the data $X \sim P_\theta$ is that the model parameters θ can be identified from the probability distribution P_θ.

Definition 1.4 (Identifiable) — A statistical model $\{P_\theta : \theta \in \Theta\}$ is **identifiable** if $P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2$ for all $\theta_1, \theta_2 \in \Theta$.

Note that estimation in non-identifiable models is not well-defined since even if we exactly recover the true distribution P_θ, there are then multiple possible equivalent 'labels' θ that can be used (by 'label', I mean that we can think of $\theta \mapsto P_\theta$ as just a labelling, and thus estimating θ only makes sense when that parameterisation is injective).[1]

Definition 1.5 (Exponential family) — A family of distributions $\{P_\theta : \theta \in \Theta\}$ is a **k-parameter exponential family** if its pmf/pdf takes the form,

$$f_\theta(x) = \exp\left\{\sum_{i=1}^{k} c_i(\theta)T_i(x) - d(\theta) + S(x)\right\},$$

where $c_i(x)$, $T_i(x)$, $d(\theta)$ and $S(x)$ are known functions and the support of f_θ ($\{x : f_\theta(x) > 0\}$) does not depend on θ.

Example 1.6 One can directly show that the normal distributions with parameters $\theta = (\mu, \sigma^2)$ form a 2-parameter exponential family.

More generally, let $X = (X_1, \ldots, X_n)$ with $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Then the joint pdf of X is

$$\begin{aligned} f_\theta(x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \exp\left\{\underbrace{-\frac{1}{2\sigma^2}}_{c_1(\theta)} \underbrace{\sum_{i=1}^{n} x_i^2}_{T_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{c_2(\theta)} \underbrace{\sum_{i=1}^{n} x_i}_{T_2(x)} - \underbrace{\left(\frac{n\mu^2}{2\sigma^2} + \frac{n}{2}\log(2\pi\sigma^2)\right)}_{d(\theta)}\right\}. \end{aligned}$$

Example 1.7 (An important non-example of an exponential family) If $X = (X_1, \ldots, X_n)$ with $X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, \theta)$, then the joint density of X is

$$f_\theta(x) = \prod_{i=1}^{n} \frac{1}{\theta}\mathbb{1}_{(0,\theta)}(x_i).$$

Since

$$\mathbb{1}_{(0,\theta)}(x_1) \cdots \mathbb{1}_{(0,\theta)}(x_n) = \mathbb{1}_{(0,\theta)}(\min_i x_i)\mathbb{1}_{(0,\theta)}(\max_i x_i),$$

we may rewrite

$$f_\theta(x) = \frac{1}{\theta^n}\mathbb{1}_{(0,\theta)}(\max_i x_i)\mathbb{1}_{(0,\theta)}(\min_i x_i).$$

Thus the support of f_θ is $[0, \theta]^n$, which depends on θ, so not an exponential family.

[1] All statistical models we consider here are identifiable.

Exponential families include many other common parametric families, and we will return to them throughout the course.

§1.1.1 Estimators

Given an i.i.d. sample $X = (X_1, \ldots, X_n)$ drawn from P_θ for some unknown θ, the goal of estimation is to construct an estimator $\hat{\theta}(X)$ for θ. We often write $\hat{\theta}_n$ to emphasize dependence on the sample size n.

Definition 1.8 (Statistic) — A **statistic** is any function $T(X)$ of the observed data. The distribution of $T(X)$ is called its sampling distribution.

Example 1.9 $T(X) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $T(X) = (X_{(1)}, \ldots, X_{(n)})$ are both statistics.

A statistic may or may not contain information about θ, despite being a function of the data.

Example 1.10 Suppose $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$. Writing $X_i = \mu + Z_i$ for $Z_i \sim N(0,1)$,

$$T(X) = \sum_{i=1}^n (X_i - \overline{X}_n)^2 = \sum_{i=1}^n (Z_i - \overline{Z}_n)^2,$$

which does *not* depend on μ.

Definition 1.11 (Bias) — The **bias** of an estimator $\hat{\theta}$ is

$$b_\theta(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta.$$

If $b_\theta(\hat{\theta}) = 0$ for all $\theta \in \Theta$, then $\hat{\theta}$ is **unbiased**.

Remark 1.12 Note that this notation E_θ means that we are taking the expectation under the probability distribution $X \sim P_\theta$, and $\hat{\theta} = \hat{\theta}(X)$ is a function of X.

Here, for instance, to evaluate $E_\theta[\hat{\theta}]$, we can either find the distribution of $\hat{\theta}$ and find its expected value, or evaluate $\hat{\theta}$ as a function of X directly, and find its expected value.

Not all estimators are unbiased for finite n (and being unbiased does not necessarily imply an estimator is good), but we often desire them to be **asymptotically unbiased**, *i.e.*

$$\mathbb{E}_\theta[\hat{\theta}_n] \to \theta \quad \text{as } n \to \infty.$$

Example 1.13 If $X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, \theta)$, the estimator $X_{(n)} = \max_i X_i$ has

$$\mathbb{E}_\theta[X_{(n)}] = \frac{n}{n+1}\theta \neq \theta,$$

so $X_{(n)}$ is biased, but $\mathbb{E}_\theta[X_{(n)}] \to \theta$ as $n \to \infty$, so it is asymptotically unbiased.

Definition 1.14 (Mean squared error) — The **mean squared error** (MSE) of $\hat{\theta}$ is

$$\text{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \quad \left(= \text{Var}_\theta(\hat{\theta}) + b_\theta(\hat{\theta})^2\right) \; .$$

Technically, this last equality in brackets isn't part of the definition, but it is not difficult to show from the definitions of bias and variance, and it forms the basis of an important concept known as the *bias-variance tradeoff*.

Definition 1.15 (Standard error) — The **standard error** of an estimator $\hat{\theta} = \hat{\theta}(X)$ is the standard deviation of its sampling distribution:

$$\text{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

The standard error often depends on the unknown parameter θ being estimated, but it can usually be estimated from the data.

Example 1.16 If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, then the estimator $\hat{\theta} = \overline{X}_n$ has standard error

$$\text{se}(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}.$$

If σ^2 is unknown, we can estimate the standard error by

$$\widehat{\text{se}}(\overline{X}_n) = \frac{S}{\sqrt{n}}, \quad \text{where } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

§1.2 Sufficiency and the Rao-Blackwell theorem

We often collect data primarily to learn about the parameter θ. However, we might not be interested in the data points themselves and just want to understand the general population behaviour. This motivates the notion of a **sufficient statistic**, which captures all information in the sample regarding θ.

Definition 1.17 (Sufficiency) — A statistic $T(X)$ is **sufficient** for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

Equivalently, this says that for any measurable set A,

$$\mathbb{P}_\theta(X \in A \mid T = t) \quad \text{is free of } \theta.$$

Example 1.18 (Poisson) If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$, then $T(X) = \sum_{i=1}^n X_i$ is sufficient for θ. Indeed, for $x_1 + \cdots + x_n = t$,

$$\frac{\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)}{\mathbb{P}_\theta(T = t)} = \frac{\prod_{i=1}^n e^{-\theta}\theta^{x_i}/x_i!}{e^{-n\theta}(n\theta)^t/t!} = \frac{t!}{n^t \prod_{i=1}^n x_i!},$$

which does not depend on θ.

A sufficient statistic allows us to keep all the information about θ while reducing the *dimension* of our data. In the example above, the dimension of our full data is n, $(X_1, \ldots, X_n)$, while the dimension of the sufficient statistic $\sum_{i=1}^n X_i$ is 1.

A standard tool to find sufficient statistics is the **factorization criterion**.

Theorem 1.19 (Factorization criterion) Suppose X has pmf/pdf $f_\theta(x)$. Then $T(X)$ is sufficient for θ if and only if there exist functions g and h such that

$$f_\theta(x) = g(T(x), \theta)\, h(x).$$

Proof. We only prove the discrete case; the continuous case is similar conceptually, but requires use of measure theory hence we omit it here. We first prove the 'if' direction. Suppose that the pmf of X can be written as

$$f_\theta(x) = \mathbb{P}_\theta(X = x) = g(T(x), \theta)\, h(x),$$

for some functions g and h. If $T(x) = t$, then

$$\begin{aligned}
\mathbb{P}_\theta(X = x \mid T = t) &= \frac{\mathbb{P}_\theta(X = x,\, T(X) = t)}{\mathbb{P}_\theta(T = t)} \\
&= \frac{\mathbb{P}_\theta(X = x)}{\sum_{y:T(y)=t} \mathbb{P}_\theta(X = y)} \\
&= \frac{g(T(x), \theta)\, h(x)}{\sum_{y:T(y)=t} g(T(y), \theta)\, h(y)} \\
&= \frac{g(t, \theta)\, h(x)}{g(t, \theta) \sum_{y:T(y)=t} h(y)} \\
&= \frac{h(x)}{\sum_{y:T(y)=t} h(y)},
\end{aligned}$$

which does not depend on θ. If instead $T(x) \neq t$, then

$$\mathbb{P}_\theta(X = x \mid T = t) = 0,$$

which also does not depend on θ. Thus the conditional distribution of X given T is free of θ, and so T is sufficient for θ.

We now prove the 'only if' direction. Suppose $T = T(X)$ is sufficient for θ, so that the conditional distribution of X given $T = t$ does not depend on θ. Then for any x,

$$\mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x,\, T = T(x)) = \mathbb{P}_\theta(X = x \mid T = T(x)) \cdot \mathbb{P}_\theta(T = T(x)).$$

By sufficiency, the conditional probability $\mathbb{P}_\theta(X = x \mid T = T(x))$ does not depend on θ; let us denote

$$h(x) := \mathbb{P}_\theta(X = x \mid T = T(x)),$$

which is a function only of x. Let

$$g(t,\theta) := \mathbb{P}_\theta(T=t).$$

Then

$$\mathbb{P}_\theta(X=x) = g(T(x),\theta)\,h(x),$$

which is precisely the required factorization. □

Example 1.20 (Uniform) If $X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,\theta]$, then

$$f_\theta(x) = \prod_{i=1}^n \frac{1}{\theta}\,\mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\,\mathbb{1}_{\{\max_i x_i \le \theta\}}\,\mathbb{1}_{\{\min_i x_i \ge 0\}}.$$

Taking $T(x) = \max_i x_i$, $g(t,\theta) = \theta^{-n}\mathbb{1}_{\{t\le\theta\}}$, and $h(x) = \mathbb{1}_{\{\min_i x_i \ge 0\}}$, theorem 1.19 implies $T = \max_i X_i$ is sufficient for θ.

Example 1.21 (Exponential families) If X has pmf/pdf in a k-parameter exponential family,

$$f_\theta(x) = \exp\left\{\sum_{i=1}^k c_i(\theta)\,T_i(x) - d(\theta) + S(x)\right\},$$

then by thoerem 1.19, $T(X) = (T_1(X),\dots,T_k(X))$ is sufficient for θ.

Sufficient statistics are not unique (indeed, by the factorization criterion, any bijective function of a sufficient statistic is also sufficient) and they may have different dimensions. We prefer those that *reduce* the data as much as possible while retaining all information about θ. For example, the full data X is always sufficient for θ, though this is not of much use, so how can we decide if a sufficient statistic is 'good'?

Definition 1.22 (Minimal sufficiency) — A sufficient statistic $T(X)$ is **minimal** if it is a function of every other sufficient statistic. Equivalently, if $T'(X)$ is sufficient, then $T'(X) = T'(Y) \implies T(X) = T(Y)$.

Thus for any other sufficient statistic T', there exists a function h such that $T(x) = h(T'(x))$A minimal sufficient statistic represents the maximal reduction of the data that contains as much information about the unknown parameter as the full data itself.

Remark 1.23 Minimal sufficient statistics are not unique: any bijective transform of a minimal sufficient statistic is also minimal (they have the same information and dimension).

Example 1.24 Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and consider the following statistics:

$$T_1(X) = (X_1, \dots, X_n), \qquad T_2(X) = (X_1^2, \dots, X_n^2),$$

$$T_3(X) = \left(\sum_{i=1}^{m} X_i^2, \sum_{i=m+1}^{n} X_i^2 \right), \qquad T_4(X) = \sum_{i=1}^{n} X_i^2.$$

All of these statistics are sufficient for σ^2. Moreover, each T_i provides a greater reduction of the data than the previous one. We will later show that $T_4(X)$ is a **minimal sufficient** statistic for σ^2.

A useful characterisation for minimal sufficiency is the following.

Theorem 1.25 Suppose X has pmf/pdf $f_\theta(x)$ and $T = T(X)$ satisfies

$$\frac{f_\theta(x)}{f_\theta(x')} \text{ is free of } \theta \quad \Longleftrightarrow \quad T(x) = T(x').$$

Then T is minimal sufficient for θ.

Proof. We first show that T is sufficient. By hypothesis, $T = T(X)$ satisfies

$$\frac{f_\theta(x)}{f_\theta(x')} \text{ is free of } \theta \iff T(x) = T(x').$$

For each possible value t in the range of T, choose some x_t such that $T(x_t) = t$. Now take any x, and write $t = T(x)$, so that $T(x) = T(x_t)$. By the hypothesis of the theorem,

$$\frac{f_\theta(x)}{f_\theta(x_t)}$$

does not depend on θ; denote this ratio by $h(x)$. Define

$$g(t, \theta) := f_\theta(x_t).$$

Then

$$f_\theta(x) = f_\theta(x_t) \cdot \frac{f_\theta(x)}{f_\theta(x_t)} = g(t, \theta)\, h(x),$$

which is exactly the factorization required by the factorization criterion. Therefore, T is sufficient for θ.

We now show that T is minimal sufficient. Let $S = S(X)$ be any other sufficient statistic. By the factorization criterion, there exist functions g_S and h_S such that

$$f_\theta(x) = g_S(S(x), \theta)\, h_S(x).$$

Let x and x' be any two sample points with $S(x) = S(x')$. Then

$$\frac{f_\theta(x)}{f_\theta(x')} = \frac{g_S(S(x), \theta)\, h_S(x)}{g_S(S(x'), \theta)\, h_S(x')} = \frac{h_S(x)}{h_S(x')},$$

which does not depend on θ. By the defining condition of the theorem, this implies that

$$T(x) = T(x').$$

Thus, whenever $S(x) = S(x')$, we have $T(x) = T(x')$. This means that T is a function of S. Since this holds for every sufficient statistic S, T is minimal sufficient. □

Example 1.26 (Normal with unknown mean and variance) Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. For $x, x^* \in \mathbb{R}^n$,

$$\frac{f_\theta(x)}{f_\theta(x^*)} = \exp\left\{ -\frac{1}{2\sigma^2}\Big(\sum_i x_i^2 - \sum_i (x_i^*)^2\Big) + \frac{\mu}{\sigma^2}\Big(\sum_i x_i - \sum_i x_i^*\Big) \right\}.$$

This ratio is constant in (μ, σ^2) if and only if $\sum_i x_i^2 = \sum_i (x_i^*)^2$ and $\sum_i x_i = \sum_i x_i^*$. Hence

$$T(X) = \left(\sum_{i=1}^n X_i^2, \ \sum_{i=1}^n X_i \right)$$

is minimal sufficient for (μ, σ^2) by theorem 1.25. Because $(\overline{X}, S^2)$ is a bijection of T, it is also minimal sufficient for (μ, σ^2).

Example 1.27 (Uniform with two parameters) If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}(\theta_1, \theta_2)$, then

$$f_\theta(x) = \frac{1}{(\theta_2 - \theta_1)^n} \mathbb{1}_{\{\max_i x_i < \theta_2\}} \mathbb{1}_{\{\min_i x_i > \theta_1\}}.$$

For x, x^*,

$$\frac{f_\theta(x)}{f_\theta(x^*)} = \frac{\mathbb{1}_{\{\max_i x_i < \theta_2\}} \mathbb{1}_{\{\min_i x_i > \theta_1\}}}{\mathbb{1}_{\{\max_i x_i^* < \theta_2\}} \mathbb{1}_{\{\min_i x_i^* > \theta_1\}}}.$$

This ratio is constant in (θ_1, θ_2) if and only if $\min_i x_i = \min_i x_i^*$ and $\max_i x_i = \max_i x_i^*$. Therefore

$$T(X) = \big(\min_i X_i, \ \max_i X_i\big)$$

is minimal sufficient for $\theta = (\theta_1, \theta_2)$.

§1.2.1 Rao-Blackwell theorem

With the Rao-Blackwell theorem, it turns out that we use sufficient statistics to *improve* any estimator.

Theorem 1.28 (Rao-Blackwell Theorem) Let $T = T(X)$ be a sufficient statistic for θ and let $\tilde{\theta}(X)$ be an estimator of θ with $\mathbb{E}_\theta[\tilde{\theta}^2] < \infty$ for all $\theta \in \Theta$. Define

$$\hat{\theta}(X) = \mathbb{E}_\theta[\tilde{\theta}(X) \mid T(X)].$$

Then for all $\theta \in \Theta$,

$$b_\theta(\hat{\theta}) = b_\theta(\tilde{\theta}), \qquad \text{and} \qquad \text{Var}_\theta(\hat{\theta}) \le \text{Var}_\theta(\tilde{\theta}),$$

with strict inequality unless $\tilde{\theta}$ is already a function of T. In particular,

$$\text{MSE}_\theta(\hat{\theta}) \le \text{MSE}_\theta(\tilde{\theta}).$$

Note that $\hat{\theta}$ is defined as the conditional expectation of $\tilde{\theta}(X)$ given T. At first glance, this may appear to depend on the unknown parameter θ. However, because T is sufficient, the conditional distribution of X given T does not depend on θ. Therefore $\hat{\theta}$ is a genuine estimator.

The Rao-Blackwell Theorem states: if an estimator is not already a function of a sufficient statistic, then replacing it by its conditional expectation given that statistic *reduces variance without changing bias.* If $\tilde{\theta}$ is unbiased, then $\hat{\theta}$ is also unbiased.

Proof. For the bias,

$$\mathbb{E}_\theta[\hat{\theta}] = \mathbb{E}_\theta\left[\mathbb{E}_\theta(\tilde{\theta} \mid T)\right] = \mathbb{E}_\theta[\tilde{\theta}],$$

so $b_\theta(\hat{\theta}) = b_\theta(\tilde{\theta})$.
For the variance, the conditional variance identity gives:

$$\mathrm{Var}_\theta(\tilde{\theta}) = \mathbb{E}_\theta\Big[\mathrm{Var}_\theta(\tilde{\theta} \mid T)\Big] + \mathrm{Var}_\theta\Big(\mathbb{E}_\theta[\tilde{\theta} \mid T]\Big) = \mathbb{E}_\theta[\mathrm{Var}_\theta(\tilde{\theta} \mid T)] + \mathrm{Var}_\theta(\hat{\theta}).$$

Since $\mathrm{Var}_\theta(\tilde{\theta} \mid T) \geq 0$, we have $\mathrm{Var}_\theta(\hat{\theta}) \leq \mathrm{Var}_\theta(\tilde{\theta})$.
Using the bias–variance decomposition yields the MSE statement. Equality holds only if $\mathrm{Var}_\theta(\tilde{\theta} \mid T) = 0$, i.e. $\tilde{\theta}$ is a function of T. □

Example 1.29 Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ and define $\theta = e^{-\lambda}$, the probability that $X_1 = 0$. Since $\lambda = -\log\theta$,

$$f_\theta(x) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = \theta^n(-\log\theta)^{\sum x_i}\Big/\prod x_i!.$$

By the factorization criterion, $T = \sum_{i=1}^n X_i$ is sufficient for θ. Note that $T \sim \text{Poisson}(n\lambda)$.
Consider the estimator $\tilde{\theta} = \mathbb{1}(X_1 = 0)$, which is unbiased since $P_\theta(X_1 = 0) = \theta$.
Then,

$$\hat{\theta} = \mathbb{E}_\theta[\tilde{\theta} \mid T = t] = \frac{P_\theta(X_1 = 0, \sum_{i=1}^n X_i = t)}{P_\theta(\sum_{i=1}^n X_i = t)} = \left(\frac{n-1}{n}\right)^t.$$

Thus the Rao-Blackwell estimator is

$$\hat{\theta} = \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^n X_i}.$$

Using $\sum X_i \approx n\overline{X}$ for large n,

$$\hat{\theta} \approx e^{-\overline{X}} = e^{-\hat{\lambda}},$$

which is intuitively much more reasonable than the crude estimator $\tilde{\theta}$.

A natural question is - which sufficient statistic should we use in the conditional expectation?

Lemma 1.30 Let T_1 and T_2 be sufficient statistics for θ, and suppose $T_2 = h(T_1)$. For any estimator $\tilde{\theta}$ with $\mathbb{E}_\theta[\tilde{\theta}^2] < \infty$, define

$$\hat{\theta}_1 = \mathbb{E}[\tilde{\theta} \mid T_1], \qquad \hat{\theta}_2 = \mathbb{E}[\tilde{\theta} \mid T_2].$$

Then for all $\theta \in \Theta$,

$$\mathrm{Var}_\theta(\hat{\theta}_2) \leq \mathrm{Var}_\theta(\hat{\theta}_1).$$

Proof. Since $T_2 = h(T_1)$, by the tower property,

$$\mathbb{E}[\hat{\theta}_1 \mid T_2] = \mathbb{E}[\mathbb{E}[\tilde{\theta} \mid T_1] \mid T_2] = \mathbb{E}[\tilde{\theta} \mid T_2] = \hat{\theta}_2.$$

Applying the Rao-Blackwell theorem to $\hat{\theta}_1$ gives $\mathrm{Var}_\theta(\hat{\theta}_2) \leq \mathrm{Var}_\theta(\hat{\theta}_1)$. □

This shows that the *best* variance reduction is achieved by conditioning on a **minimal sufficient statistic**. For a given starting estimator $\tilde{\theta}$ and minimal sufficient statistic T, the best Rao-Blackwell estimator is $\mathbb{E}[\tilde{\theta} \mid T]$.

However, the result still depends on the choice of the initial estimator $\tilde{\theta}$. There may exist another unbiased estimator $\tilde{\theta}^*$ such that $\mathbb{E}[\tilde{\theta}^* \mid T]$ has strictly smaller variance. However, the Rao-Blackwell procedure alone does not guarantee the minimum-variance unbiased estimator (UMVUE). To guarantee uniqueness, we will need the notion of a **complete statistic** (see section section 5.3).

§2 Likelihood-Based Estimation

§2.1 The likelihood function

Definition 2.1 (Likelihood and log-likelihood) — Let $X = (X_1, \dots, X_n)$ have joint pmf/pdf $f_\theta(x) = f_{n,\theta}(x)$ for $\theta \in \Theta$, and suppose we observe a realization x. The **likelihood function** is $L : \Theta \to \mathbb{R}$,

$$L(\theta) = L_n(\theta) = f_{n,\theta}(x),$$

regarded as a function of θ with x fixed. The **log-likelihood function** is $l : \Theta \to \mathbb{R}$,

$$l(\theta) = l_n(\theta) = \log L_n(\theta) = \log f_{n,\theta}(x).$$

In the i.i.d. case where $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_{1,\theta}$,

$$L_n(\theta) = \prod_{i=1}^n f_{1,\theta}(x_i), \qquad l_n(\theta) = \sum_{i=1}^n \log f_{1,\theta}(x_i).$$

Definition 2.2 (Maximum Likelihood Estimator) — A **maximum likelihood estimator** (MLE) is any $\hat{\theta} \in \Theta$ satisfying

$$L_n(\hat{\theta}) = \max_{\theta \in \Theta} L_n(\theta).$$

Remark 2.3 Since $\log(\cdot)$ is strictly increasing, maximizing $L_n(\theta)$ is equivalent to maximizing $l_n(\theta)$. The MLE may not be unique.

Moreover, if $l_n(\theta)$ is differentiable in $\theta = (\theta_1, \dots, \theta_p)$, then an interior maximizer must satisfy the **likelihood equations**

$$\frac{\partial}{\partial \theta_k} l_n(\hat{\theta}) = 0, \qquad k = 1, \dots, p.$$

Solutions must still be checked to ensure they correspond to maxima.

Example 2.4 (Exponential Distribution) Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, $\lambda > 0$. Then

$$L_n(\lambda) = \lambda^n \exp\Big(-\lambda \sum_{i=1}^n x_i \Big), \qquad l_n(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n x_i.$$

Differentiating,

$$\frac{dl_n}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{1}{\overline{X}_n}.$$

Since $\frac{d^2 l_n}{d\lambda^2} = -n/\lambda^2 < 0$, this is the global maximum.

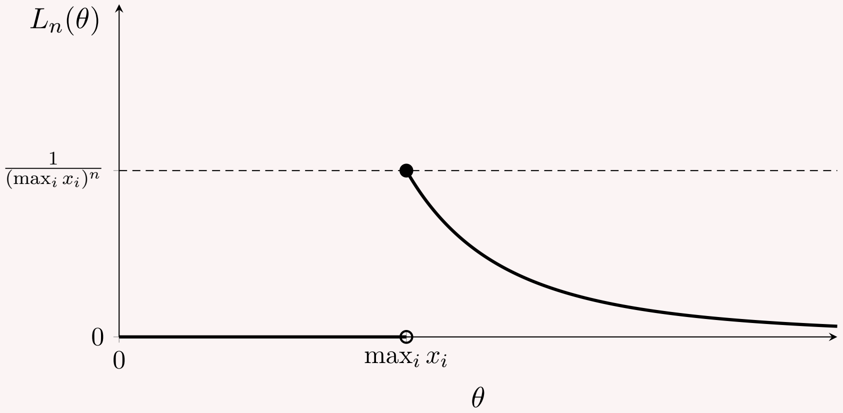
Example 2.5 (Uniform Distribution) Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} U(0, \theta)$, $\theta > 0$. The likelihood is

$$L_n(\theta) = f_{n,\theta}(x) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x_i) = \theta^{-n} \mathbb{1}\{\max X_i \leq \theta\}.$$

Since θ^{-n} decreases in θ, the likelihood is maximized by choosing the smallest θ that is still valid:

$$\hat{\theta}_{\text{ML}} = \max_{1 \leq i \leq n} X_i.$$

We can plot the likelihood function as below:



Remark 2.6 (How does the MLE relate to sufficient statistics?) If $T(X)$ is sufficient for θ, the factorization criterion gives for the likelihood,

$$L(\theta) = g(T(x), \theta) h(x).$$

Since $h(x)$ does not depend on θ, maximizing $L(\theta)$ is equivalent to maximizing $g(T(x), \theta)$. Thus, the MLE is always a *function* of a *sufficient statistic* (when one exists).

§2.1.1 Invariance of the MLE

A key property of maximum likelihood estimators is **invariance**: if $\hat{\theta}_{\text{ML}}$ is an MLE for θ and g is any function, then $g(\hat{\theta}_{\text{ML}})$ is an MLE for $g(\theta)$.
Let $\eta = g(\theta)$ and define the induced likelihood for η by

$$L^*(\eta) = \sup_{\theta : g(\theta) = \eta} L(\theta).$$

The value $\hat{\eta}$ which maximises L^* is the MLE of η. Since maximising L over θ or maximising L^* over η yields the same likelihood value, the maxima coincide.

Theorem 2.7 (Invariance of the MLE) If $\hat{\theta}_{\text{ML}}$ is an MLE for θ and $g(\theta)$ is any function, then $g(\hat{\theta}_{\text{ML}})$ is an MLE for $g(\theta)$.

Proof. Let $\hat{\eta}$ be the maximiser of L^*. Then

$$L^*(\hat{\eta}) = \sup_{\eta} \sup_{\theta : g(\theta)=\eta} L(\theta) = \sup_{\theta} L(\theta) = L(\hat{\theta}_{\mathrm{ML}}).$$

Moreover,

$$L(\hat{\theta}_{\mathrm{ML}}) = \sup_{\theta : g(\theta)=g(\hat{\theta}_{\mathrm{ML}})} L(\theta) = L^*(g(\hat{\theta}_{\mathrm{ML}})).$$

Thus $L^*(\hat{\eta}) = L^*(g(\hat{\theta}_{\mathrm{ML}}))$, showing that $g(\hat{\theta}_{\mathrm{ML}})$ is the maximiser of L^*, i.e. the MLE of $g(\theta)$. □

Example 2.8 If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda)$, then $\hat{\lambda}_{\mathrm{ML}} = 1/\bar{X}_n$. The variance is $g(\lambda) = 1/\lambda^2$, hence the MLE is

$$\widehat{\mathrm{Var}}(X_1) = \left(\hat{\lambda}_{\mathrm{ML}}\right)^{-2} = \bar{X}_n^2.$$

In general, if $\hat{\sigma}$ is the MLE of the standard deviation, then $\hat{\sigma}^2$ is the MLE of the variance.

§2.2 Geometry of the likelihood: score and Fisher information

Recall that $l_n(\theta) = \log L_n(\theta)$ is a random function of θ. To understand the behaviour of the MLE, we compare $l_n(\theta)$ with its expectation.

Lemma 2.9 Assume $E_\theta|\log f_\theta(X)| < \infty$ for all θ. If $X \sim f_{\theta_0}$, then

$$E_{\theta_0}[l(\theta)] \leq E_{\theta_0}[l(\theta_0)], \qquad \text{for all } \theta \in \Theta.$$

Thus, the expected log-likelihood is maximised at the true parameter θ_0.

This lemma suggests that if we knew the function $\theta \mapsto E_{\theta_0}[l(\theta)]$, we could recover the true θ_0 exactly by maximisation. Since the (normalized) log-likelihood, $\frac{1}{n} l_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f_\theta(X_i)$, is an empirical average, the law of large numbers gives $\frac{1}{n} l_n(\theta) \to E_{\theta_0}[l(\theta)]$ for every $\theta \in \Theta$.

This provides a heuristic justification for maximum likelihood estimation.

§2.2.1 The Score Function

Definition 2.10 (Score function) — For $\Theta \subseteq \mathbb{R}^p$ and differentiable $l_n(\theta)$, the **score function** is

$$S_n(\theta) = \nabla_\theta l_n(\theta) = \begin{pmatrix} \partial l_n / \partial \theta_1 \\ \vdots \\ \partial l_n / \partial \theta_p \end{pmatrix}.$$

The likelihood equations are $S_n(\hat{\theta}) = 0$.

Lemma 2.11 Consider a model $f_\theta : \theta \in \Theta$ that is regular enough that differentiation (in θ) and integration (in x) can be exchanged.[a] Then for all $\theta \in \operatorname{int}(\Theta)$,

$$E_\theta[\nabla_\theta \log f_\theta(X)] = 0.$$

[a] This rule of exchanging of integration and differentiation is formalized properly in remark 3.4.

In particular, this implies $E_{\theta_0}[\nabla_\theta \log f_{\theta_0}(X)] = 0$ at the true parameter θ_0.

Remark 2.12 Conditions for interchanging the order of differentiation and integration can be found in a measure theory course. But, note that when the support of f_θ depends on θ, this is generally not true.

For example, we saw if $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} U[0, \theta]$, one cannot obtain the MLE for θ by differentiating the likelihood.

§2.2.2 Fisher Information

Definition 2.13 (Fisher information (matrix)) — The **Fisher information matrix** is

$$I(\theta) = E_\theta\left[\nabla_\theta \log f_\theta(X)\ \nabla_\theta \log f_\theta(X)^\top\right],$$

or coordinate-wise,

$$I_{ij}(\theta) = E_\theta\left[\frac{\partial}{\partial\theta_i} \log f_\theta(X) \frac{\partial}{\partial\theta_j} \log f_\theta(X)\right] \quad , \quad 1 \le i, j \le p.$$

In one dimension,

$$I(\theta) = \operatorname{Var}_\theta\left(\frac{d}{d\theta} \log f_\theta(X)\right).$$

Lemma 2.14 Under the same regularity assumptions as in lemma 2.11,

$$I(\theta) = -E_\theta\left[\nabla_\theta^2 \log f_\theta(X)\right],$$

or coordinate-wise,

$$I_{ij}(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta_i\, \partial\theta_j} \log f_\theta(X)\right] \quad , \quad 1 \le i, j \le p.$$

In dimension $p = 1$, this lemma becomes the more ‘friendly’, well-known result,

$$I(\theta) = E_\theta\left[(l'(\theta))^2\right] = -E_\theta\left[l''(\theta)\right].$$

Proposition 2.15 If $X_1, \ldots, X_n$ are i.i.d. with Fisher information $I(\theta)$ for one observation of random variable X_i, then the Fisher information for $X = (X_1, \ldots, X_n)$ is

$$I_n(\theta) = nI(\theta).$$

§2.3 The Cramer-Rao lower bound

The Fisher information controls the best possible precision of unbiased estimation. In one dimension, it yields a lower bound on the variance of any *unbiased* estimator.

Theorem 2.16 (Cramer-Rao Lower Bound) Consider a model $\{f_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ for which differentiation in θ and integration in x can be interchanged. Let $\hat{\theta} = \hat{\theta}(X)$ be an unbiased estimator of θ based on an observation $X \sim f_\theta$. Then, for all $\theta \in \operatorname{int}(\Theta)$,

$$\operatorname{Var}_\theta(\hat{\theta}) \;=\; \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \;\geq\; \frac{1}{I(\theta)},$$

where $I(\theta)$ is the Fisher information of X.

Proof. By Cauchy-Schwarz, for any random variables Y, Z,

$$\operatorname{Cov}_\theta(Y, Z)^2 \;\leq\; \operatorname{Var}_\theta(Y)\operatorname{Var}_\theta(Z).$$

Apply this with $Y = \hat{\theta}$ and $Z = \frac{d}{d\theta}\log f_\theta(X) = l'(\theta)$. Then

$$\operatorname{Var}_\theta(\hat{\theta}) \;\geq\; \frac{\operatorname{Cov}_\theta(\hat{\theta}, Z)^2}{\operatorname{Var}_\theta(Z)}. \tag{1}$$

From lemma 2.11, $\mathbb{E}_\theta[Z] = 0$ and $\operatorname{Var}_\theta(Z) = I(\theta)$. In the continuous case,

$$\begin{aligned}\operatorname{Cov}_\theta(\hat{\theta}, Z) = \mathbb{E}_\theta[\hat{\theta}Z] - \mathbb{E}_\theta[\hat{\theta}]\mathbb{E}_\theta[Z] &= \int \hat{\theta}(x)\,\frac{1}{f_\theta(x)}\frac{d}{d\theta}f_\theta(x)\,f_\theta(x)\,dx \\ &= \frac{d}{d\theta}\int \hat{\theta}(x) f_\theta(x)\,dx \\ &= \frac{d}{d\theta}\,\mathbb{E}_\theta[\hat{\theta}] = \frac{d}{d\theta}\,\theta = 1,\end{aligned}$$

where we used interchange of differentiation and integration and the unbiasedness of $\hat{\theta}$. Substituting into (1) gives the result. The discrete case is identical with integrals replaced by sums. □

Remark 2.17 (i.i.d. sample) If $X = (X_1, \ldots, X_n)$ with X_i i.i.d., then $I_n(\theta) = n\,I_{X_1}(\theta)$ and

$$\operatorname{Var}_\theta(\hat{\theta}) \;\geq\; \frac{1}{n\,I_{X_1}(\theta)}.$$

Remark 2.18 (Biased estimators for $g(\theta)$) For scalar θ and differentiable $g : \Theta \to \mathbb{R}$, if $\hat{g}$ estimates $g(\theta)$ with bias $b(\theta) = \mathbb{E}_\theta[\hat{g}] - g(\theta)$, then

$$\operatorname{Var}_\theta(\hat{g}) \;\geq\; \frac{\left(g'(\theta) + b'(\theta)\right)^2}{I(\theta)}. \tag{2}$$

This follows by repeating the last proof with $\mathbb{E}_\theta[\hat{g}] = g(\theta) + b(\theta)$.

Proposition 2.19 (Multivariate extension) For $\theta \in \mathbb{R}^p$ and differentiable $g : \Theta \to \mathbb{R}$, any unbiased estimator $\hat{g}$ satisfies

$$\mathrm{Var}_\theta(\hat{g}) \;\geq\; \nabla_\theta g(\theta)^\top I(\theta)^{-1}\, \nabla_\theta g(\theta).$$

For $g(\theta) = u^\top \theta$, this reduces to $\mathrm{Var}_\theta(\hat{g}) \geq u^\top I(\theta)^{-1} u$.

Example 2.20 (Normal mean, variance known) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma_0^2)$ with σ_0^2 known. Then $I_{X_1}(\mu) = 1/\sigma_0^2$, so

$$\mathrm{Var}_\mu(\hat{\mu}) \;\geq\; \frac{1}{I(\mu)} = \frac{1}{nI_{X_1}(\mu)} = \frac{1}{n/\sigma_0^2} = \frac{\sigma_0^2}{n}.$$

The unbiased estimator $\bar{X}_n$ attains the bound: $\mathrm{Var}(\bar{X}_n) = \sigma_0^2/n$.

A natural follow-up question is then, whe in general is this bound attainable?

Proposition 2.21 Assume regularity conditions and $p = 1$. An unbiased statistic $\hat{\theta}(X)$ attains the Cramer-Rao lower bound if and only if X belongs to the exponential family

$$f_\theta(x) = \exp\bigl(A(\theta)\,\hat{\theta}(x) + B(\theta) + S(x)\bigr)$$

for some functions A, B, S.

Proof. From the proof of theorem 2.16, with $Z = l'(\theta)$ we have $\mathrm{Var}_\theta(Z) = I(\theta)$ and $\mathrm{Cov}_\theta(\hat{\theta}, Z) = 1$. Equality in (1) (hence attainment of the bound) holds iff Z is an affine function of $\hat{\theta}$, *i.e.*

$$l'(\theta) = A^*(\theta)\,\hat{\theta}(x) + B^*(\theta) \quad \text{for all } x.$$

Integrating in θ and exponentiating yields

$$f_\theta(x) = \exp\bigl(A(\theta)\,\hat{\theta}(x) + B(\theta) + S(x)\bigr),$$

with $S(x)$ absorbing the constant of integration. Conversely, such a representation implies the linear relation and hence equality. □

Note that in general the Cramer-Rao lower bound cannot always be attained, but if an unbiased estimator attains the lower bound, then it has the best possible variance among all unbiased estimators.

We will see an example where it cannot be attained in section 5.3.2.

§3 Asymptotic Theory for MLEs

§3.1 Consistency

A desirable property of an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is that it be *asymptotically unbiased*, meaning

$$\mathbb{E}_\theta[\hat{\theta}_n] \ \to \ \theta, \quad \text{as } n \to \infty.$$

This guarantees that the estimator is centered correctly in the limit, but does not rule out the estimator continuing to fluctuate significantly.

We now study a stronger property: the estimator itself should converge to the true parameter with high probability as the sample size grows.

Definition 3.1 (Consistency) — Let $(X_1, \ldots, X_n)$ be i.i.d. from a statistical model $\{P_\theta : \theta \in \Theta\}$. A sequence of estimators $\hat{\theta}_n$ is said to be **consistent** for θ_0 if

$$\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0,$$

i.e. for every $\varepsilon > 0$,

$$P_{\theta_0}\Big(|\hat{\theta}_n - \theta_0| > \varepsilon\Big) \ \to \ 0 \quad \text{as } n \to \infty.$$

Example 3.2 If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, then by the *weak law of large numbers*,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mu,$$

so $\bar{X}_n$ is consistent for μ.

In this same model,

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \ \to \ 0,$$

so $\bar{X}_n$ is consistent for μ. Similarly, the sample variance

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

is consistent for σ^2, since $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ and hence $S^2 \to_P \sigma^2$.

Remark 3.3 By Markov's inequality,

$$P_{\theta_0}(|\hat{\theta}_n - \theta_0| > \varepsilon) = P_{\theta_0}((\hat{\theta}_n - \theta_0)^2 > \varepsilon^2) \le \frac{\mathbb{E}_{\theta_0}[(\hat{\theta}_n - \theta_0)^2]}{\varepsilon^2} = \frac{\mathrm{MSE}_{\theta_0}(\hat{\theta}_n)}{\varepsilon^2}.$$

Thus, if $\mathrm{MSE}_{\theta_0}(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n$ is consistent. In particular, consistency follows if both bias and variance vanish asymptotically.

The maximum likelihood estimator often performs well in large samples. To illustrate, consider $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. The MLE (and sample mean) $\hat{\lambda}_n = \bar{X}_n$ is unbiased and attains the Cramer-Rao lower bound, $\mathrm{Var}_\lambda(\hat{\lambda}_n) = \frac{\lambda}{n} = \frac{1}{nI_{X_1}(\lambda)}$, so its variance is optimal among unbiased estimators. Moreover, by the *central limit theorem*,

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{d} N(0, \lambda) \implies \hat{\lambda}_n \approx N\left(\lambda,\, \frac{1}{nI_{X_1}(\lambda)}\right).$$

More generally, we will later show that the MLE is *asymptotically efficient*, which is exactly what is meant by the MLE for λ in the above having that distribution.

We now establish the consistency of the MLE, under some regularity conditions on the model. These regularity conditions are more of just a formality for rigor and are not super important to know in detail in the bigger picture.

Remark 3.4 (Model regularity) For model regularity, suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta$, where $\theta \in \Theta \subset \mathbb{R}$ and the log-likelihood of one observation is $l_{X_1}(\theta) = \log f_\theta(X_1)$. Assume:

1. Θ is an open subset of $\mathbb{R}$.
2. For each x, $l_{X_1}(\theta)$ is twice continuously differentiable in θ.
3. $\mathbb{E}_\theta[\, l''_{X_1}(\theta)\,] < \infty$ for all θ.
4. Differentiation in θ may be interchanged with integration in x up to second order.

Theorem 3.5 (Consistency of the MLE) Let $\{f_\theta : \theta \in \Theta\}$ satisfy assumptions 3.4, and suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_{\theta_0}$ for some $\theta_0 \in \Theta$. Then

$$\hat{\theta}_{MLE} \xrightarrow{P_{\theta_0}} \theta_0.$$

§3.2 Asymptotic normality of the MLE

The MLE $\hat{\theta}$ is consistent (theorem 3.5), *i.e.* $\hat{\theta} \xrightarrow{P} \theta_0$. We now quantify its stochastic fluctuations for large n.

Theorem 3.6 (Asymptotic normality of the MLE) Let $\{f_\theta : \theta \in \Theta\}$ satisfy assumptions 3.4 and suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_{\theta_0}$ for some true $\theta_0 \in \Theta$. Then,

$$\sqrt{n}\,(\hat{\theta} - \theta_0) \;\xrightarrow{d}\; N\left(0,\, \frac{1}{I_{X_1}(\theta_0)}\right),$$

where $I_{X_1}(\theta_0)$ is the Fisher information for one observation.

Proof. Write $l_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)$, so $l'_n(\hat{\theta}) = 0$. By the mean value theorem, there exists $\tilde{\theta}$ between θ_0 and $\hat{\theta}$ such that

$$l''_n(\tilde{\theta}) \;=\; \frac{l'_n(\hat{\theta}) - l'_n(\theta_0)}{\hat{\theta} - \theta_0} \implies 0 = l'_n(\hat{\theta}) = l'_n(\theta_0) + l''_n(\tilde{\theta})(\hat{\theta} - \theta_0).$$

Hence

$$\hat{\theta} - \theta_0 = -\frac{l_n'(\theta_0)}{l_n''(\tilde{\theta})} \quad \text{and} \quad \sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\frac{1}{\sqrt{n}} l_n'(\theta_0)}{\frac{1}{n} l_n''(\tilde{\theta})}. \tag{3}$$

For the numerator, by lemma 2.9 (score has mean 0) and the CLT,

$$\frac{1}{\sqrt{n}} l_n'(\theta_0) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_\theta(X_i)\Big|_{\theta=\theta_0} - \mathbb{E}_{\theta_0}\left[l_1'(\theta_0)\right]\right) \xrightarrow{d} N(0,\ I_{X_1}(\theta_0)).$$

For the denominator, write $l_1(\theta; X_i) = \log f_\theta(X_i)$. Then

$$\frac{1}{n} l_n''(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n} l_1''(\tilde{\theta}; X_i) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} (l_1''(\tilde{\theta}; X_i) - \mathbb{E}_{\theta_0}[l_1''(\tilde{\theta}; X_1)])}_{\to_P 0} + \mathbb{E}_{\theta_0}[l_1''(\tilde{\theta}; X_1)].$$

Consistency gives $\tilde{\theta} \to_P \theta_0$. By continuity (assumptions 3.4) and lemma 2.14,

$$\mathbb{E}_{\theta_0}[l_1''(\tilde{\theta}; X_1)] \ \to_P \ \mathbb{E}_{\theta_0}[l_1''(\theta_0; X_1)] = \ -I_{X_1}(\theta_0).$$

The uniform law of large numbers yields the first term $\to_P 0$. Hence,

$$\frac{1}{n} l_n''(\tilde{\theta}) \ \xrightarrow{P} \ -I_{X_1}(\theta_0).$$

Finally, combining numerator and denominator limits in (3) via *Slutsky's theorem*, we obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) \ \xrightarrow{d} \ \frac{1}{I_{X_1}(\theta_0)} N(0, I_{X_1}(\theta_0)) \ = \ N\Big(0,\ I_{X_1}(\theta_0)^{-1}\Big).$$

□

Remark 3.7 (Multivariate version) For $\theta \in \mathbb{R}^p$ $(p \geq 1)$,

$$\sqrt{n}\,(\hat{\theta} - \theta_0) \ \xrightarrow{d} \ N_p(0,\ I_{X_1}(\theta_0)^{-1}),$$

where the inverse Fisher information matrix (for one observation) $I_{X_1}^{-1}(\theta_0)$ is now the limiting $p \times p$ covariance matrix.

This results is useful in constructing (*e.g. Wald-type*) confidence intervals and tests for MLEs; see section 6.2.

§3.3 Asymptotic efficiency and the delta method

It is often difficult to compute the variance of an estimator exactly for finite n. Instead, we can approximate this through its *asymptotic variance.*

Definition 3.8 (Asymptotic efficiency) — In a parametric model $\{f_\theta : \theta \in \Theta\}$, a consistent estimator $\hat{\theta}_n$ is **asymptotically efficient** if

$$n \operatorname{Var}_{\theta_0}(\hat{\theta}_n) \ \to \ I(\theta_0)^{-1}$$

for all $\theta_0 \in \operatorname{int}(\Theta)$ (or, equivalently, $n\mathrm{Cov}_{\theta_0}(\hat{\theta}_n) \to I(\theta_0)^{-1}$ if $\theta \in \mathbb{R}^p$).

This states that the asymptotic variance attains the Cramer-Rao lower bound.

Remark 3.9 Under the usual regularity conditions, theorem 3.6 gives that the MLE is asymptotically normal with variance $n^{-1}I_{X_1}(\theta_0)^{-1}$ and is therefore asymptotically efficient.

If the regularity assumptions fail however, the MLE may not be efficient; *e.g.* when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$ and the likelihood is discontinuous.

Example 3.10 If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, then $\hat{\theta}_{\text{ML}} = \bar{X}_n$ and $I_{X_1}(\theta) = 1$. Thus,

$$\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_0) \xrightarrow{d} N(0, 1),$$

so $\hat{\theta}_{\text{ML}}$ is asymptotically efficient.

Example 3.11 If the parameter of interest is $g(\theta) = e^{t\theta}$ (with t known), then by invariance of the MLE, the MLE of $g(\theta)$ is $g(\hat{\theta}_{\text{ML}}) = e^{t\bar{X}_n}$, which is also asymptotically efficient.

The invariance of the MLE in the previous example allows us to transfer statements about the asymptotic variance to the MLE of $g(\theta)$.

Theorem 3.12 (Delta method) Let $g : \Theta \to \mathbb{R}$ be continuously differentiable at θ_0 with $\nabla_\theta g(\theta_0) \neq 0$. If (Y_n) and Z are random variables such that $\sqrt{n}(Y_n - \theta_0) \xrightarrow{d} Z$ $(\in \mathbb{R}^p)$, then

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \xrightarrow{d} \nabla_\theta g(\theta_0)^\top Z.$$

Proof. Define $h(t) = g(tY_n + (1-t)\theta_0)$ for $t \in [0, 1]$. By the mean value theorem, for some $\tilde{\theta}_n$ between Y_n and θ_0,

$$g(Y_n) - g(\theta_0) = \nabla_\theta g(\tilde{\theta}_n)^\top (Y_n - \theta_0).$$

Consistency implies $Y_n \to_P \theta_0$, hence $\tilde{\theta}_n \to_P \theta_0$. By continuity, $\nabla_\theta g(\tilde{\theta}_n) \to_P \nabla_\theta g(\theta_0)$. Slutsky's theorem now gives the result,

$$\sqrt{n}\,(g(Y_n) - g(\theta_0)) = \nabla_\theta g(\tilde{\theta}_n)^\mathsf{T} \sqrt{n}\,(Y_n - \theta_0) \xrightarrow{d} \nabla_\theta g(\theta_0)^\mathsf{T} Z.$$

□

Corollary 3.13 If $\sqrt{n}(Y_n - \theta_0) \xrightarrow{d} N(0, \Sigma)$, then

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \xrightarrow{d} N(0,\ \nabla_\theta g(\theta_0)^\top \Sigma\, \nabla_\theta g(\theta_0)).$$

Proof. The proof remains the same, just with the specific normal limiting distribution. If A is an $m \times p$ matrix and $Z \sim N_p(0, \Sigma)$, then $AZ \sim N_m(0, A\Sigma A^T)$ as required. □

Remark 3.14 In dimension $p = 1$, the previous corollary yields the well-known result: if $\sqrt{n}(Y_n - \theta_0) \to_d N(0, \sigma^2)$, then

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \to_d N(0,\ g'(\theta_0)^2\sigma^2).$$

Remark 3.15 Applying the delta method to the MLE, theorem 3.6 yields

$$\sqrt{n}(g(\hat{\theta}_{\mathrm{ML}}) - g(\theta_0)) \xrightarrow{d} N(0,\ \nabla_\theta g(\theta_0)^\top I(\theta_0)^{-1} \nabla_\theta g(\theta_0)),$$

which is exactly the Cramer-Rao bound, so $g(\hat{\theta}_{\mathrm{ML}})$ is asymptotically efficient.

Example 3.16 If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda)$, then $\hat{\lambda}_{\mathrm{ML}} = 1/\bar{X}_n$. Since $\sqrt{n}(\bar{X}_n - 1/\lambda) \xrightarrow{d} N(0, 1/\lambda^2)$ by the central limit theorem, we set $\theta = 1/\lambda$ and $g(\theta) = 1/\theta$ to obtain by the delta method,

$$\sqrt{n}(\hat{\lambda}_{\mathrm{ML}} - \lambda) \xrightarrow{d} N(0, \lambda^2).$$

§3.3.1 Estimating the standard error

The delta method is useful when one is interested in obtaining the asymptotic distribution of a function of the estimator, such as the standard (*i.e.* the standard deviation of the estimation). For example, in constructing confidence intervals, an estimate of the standard error would be required.

In the previous exponential example, $\mathrm{se}(\hat{\lambda}_{\mathrm{ML}}) \approx \lambda/\sqrt{n}$ but λ is unknown. We therefore estimate it by plugging in the MLE.

$$\widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{ML}}) = \frac{\hat{\lambda}_{\mathrm{ML}}}{\sqrt{n}} = \frac{1}{\sqrt{n}\,\bar{X}_n} \approx N\left(\frac{\lambda}{\sqrt{n}}, \frac{\lambda^2}{n^2}\right)$$

This limiting distribution approximation above is simply obtained from the delta method, $n\left(\frac{1}{\sqrt{n}\bar{X}_n} - \frac{\lambda}{\sqrt{n}}\right) \to^d N\left(0, \lambda^2\right)$. More generally, from theorem 3.6 in dimension $p = 1$, we have,

$$\mathrm{se}_{\theta_0}(\hat{\theta}_{\mathrm{ML}}) = \mathrm{Var}_{\theta_0}(\hat{\theta}_{ML})^{\frac{1}{2}} \approx \frac{1}{\sqrt{n\,I(\theta_0)}},$$

we estimate the Fisher information by plugging in $\hat{\theta}_{\mathrm{ML}}$:

$$\widehat{\mathrm{se}}(\hat{\theta}_{\mathrm{ML}}) = \frac{1}{\sqrt{n\,I(\hat{\theta}_{\mathrm{ML}})}}.$$

Alternatively, since $I(\theta_0) = -E_{\theta_0}[l''(\theta_0)]$, we can replace the expectation by the sample mean of the observed second derivative,

$$\hat{I}(\hat{\theta}_{\mathrm{ML}}) = -\frac{1}{n}\sum_{i=1}^{n} \ell''(\hat{\theta}_{\mathrm{ML}}; X_i).$$

This is called the *observed* Fisher information.

§4 Bayesian Inference

§4.1 Priors and posteriors

In the frequentist approach, the parameter θ is a fixed unknown. In the Bayesian approach, we treat θ as a *random variable* with its own distribution π on Θ. This can encode real uncertainty in the data-generating process, or it can represent subjective beliefs or outside information about the true value θ_0.

Suppose $X \sim f_\theta$ from a statistical model $\{f_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^p$. The **prior** $\pi(\theta)$ describes beliefs about the probability distribution of θ *before* observing data. After observing $X = x$, we update the distribution of θ using **Bayes' theorem** to obtain the **posterior** distribution,

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\,\pi(\theta)}{f_X(x)} = \frac{f_\theta(x)\,\pi(\theta)}{\int_\Theta f_{\theta'}(x)\,\pi(\theta')\,d\theta'} \propto f_\theta(x)\,\pi(\theta).$$

The denominator $f_X(x)$ is the **marginal likelihood** (or 'evidence'). Since it does not depend on θ, we can write this, conveniently, as

$$\underbrace{\pi(\theta \mid x)}_{\text{posterior}} \propto \underbrace{L(\theta)}_{\text{likelihood}} \times \underbrace{\pi(\theta)}_{\text{prior}},$$

and the constant of proportionality makes the posterior integrate to 1. Because the data enters through the likelihood $L(\theta)$, inference is automatically based on any sufficient statistic, by the factorization criterion.

The posterior gives us a way of blending what we believe before seeing x (the prior) and a summary of what the data says (the likelihood).

In Bayesian analysis, all inference about θ is based on the posterior distribution. Commonly used point estimators are the posterior mean and mode, but to determine the 'best' Bayesian estimator, one typically needs to consider a **loss function** (see section 5.1.1).

Example 4.1 (Normal-Normal) Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$ and place the prior $\theta \sim N(0, \tau_0^2)$ (known τ_0^2). Then,

$$\begin{aligned}\pi(\theta \mid x) &\propto \exp\left(-\frac{\theta^2}{2\tau_0^2}\right)\prod_{i=1}^n \exp\left(-\frac{(x_i-\theta)^2}{2}\right)\\ &\propto \exp\left(-\frac{1+n\tau_0^2}{2\tau_0^2}\left(\theta - \frac{\tau_0^2}{1+n\tau_0^2}n\bar{x}\right)^2\right),\end{aligned}$$

so,

$$\pi(\theta \mid x) = N\left(\frac{\tau_0^2}{1+n\tau_0^2}n\bar{x},\ \frac{\tau_0^2}{1+n\tau_0^2}\right).$$

The interpretation of this is that the posterior mean is a shrinkage estimator of $\bar{x}$ toward 0, with shrinkage factor $\frac{\tau_0^2}{1+n\tau_0^2}$. As $n \to \infty$, the mean approaches $\bar{x}$ and the variance decays like $1/n$.

Remark 4.2 (Conjugacy) In the previous example 4.1, the posterior belongs to the *same* family as the prior; such priors are known as **conjugate priors**. Conjugate priors are typically used for computational convenience, since the posterior can be computed analytically, specifically because they allow you to compute the integral in the denominator of Bayes' theorem, $f_X(x) = \int_\Theta f_{\theta'}(x)\pi(\theta')d\theta'$. For non-conjugate priors, this can be difficult computationally.

Even if $\int_\Theta \pi(\theta)\, d\theta = \infty$, the posterior is still well-defined provided that,

$$\int_\Theta f_\theta(x)\, \pi(\theta)\, d\theta \;<\; \infty.$$

Definition 4.3 (Improper prior) — A nonnegative prior function π with $\int_\Theta \pi(\theta)\, d\theta = \infty$ is called an **improper prior**.

Example 4.4 (Improper prior for a normal mean) With $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$ and $\pi(\theta) \propto 1$ (improper on $\mathbb{R}$),

$$\pi(\theta \mid x) \;\propto\; \prod_{i=1}^{n} \exp\Big(-\tfrac{1}{2}(x_i - \theta)^2\Big) \;\propto\; \exp\Big(-\tfrac{n}{2}(\theta - \bar{x})^2\Big),$$

so $\pi(\theta \mid x) = N(\bar{x}, 1/n)$, *i.e.* this improper prior can be viewed as the limit of $N(0, \tau_0^2)$, from example 4.1, as $\tau_0^2 \to \infty$ (indeed, in the limit $\tau_0 \to \infty$, the posterior distribution satisfies $N(\frac{\tau_0^2}{1+n\tau_0^2}\, n\bar{x},\ \frac{\tau_0^2}{1+n\tau_0^2}) \to N(\bar{x}, 1/n)$).

§4.2 Jeffreys priors

When there is no compelling prior information available, one often seeks a 'default', 'objective', or 'non-informative' prior. While there is no accepted definition or choice for this, there are several possibilities.

A naive idea is to pick a **uniform prior** (also known as a 'flat prior') that assigns equal weight to all possible values of θ, and so is as uninformative as possible (e.g. $\pi(\theta) \propto 1$ on an unbounded Θ). However, uniformity depends on the *parameterization.*

Example 4.5 (Parametrization dependence) If $X \sim \text{Binomial}(n, p)$ and $p \sim \text{Unif}[0, 1]$, then $q = \sqrt{p} \in [0, 1]$ has

$$\Pi(q \le t) = \Pi(p \le t^2) = t^2, \qquad \pi(q) = 2q,$$

which heavily weights values near $q = 1$ and de-weights near 0. A "uniform" prior in p is not uniform in q (nor in the odds $p/(1-p)$).

This motivates a prior that is *invariant* under reparameterization. Jeffreys proposed such a prior based on the Fisher information.

Definition 4.6 (Jeffreys prior) — For a model with Fisher information matrix $I(\theta)$, the **Jeffreys prior** is

$$\pi(\theta) \propto \sqrt{\det I(\theta)}.$$

For $p = 1$, this reduces to $\pi(\theta) \propto \sqrt{I(\theta)}$.

By the Cramer-Rao bound, $I(\theta)^{-1}$ is a lower bound for the variance of an unbiased estimator of θ. Thus, large $I(\theta)$ indicates the parameter θ is easier to estimate from data.

The Jeffreys prior weights the parameter space by local 'statistical resolution': when the data is informative (large I), the prior exerts less influence on the posterior; when the data is less informative (small I), the prior mass is lighter, avoiding overstatement. Since information adds over i.i.d. samples (proposition 2.15), the Jeffreys prior for n observations equals that for one observation.

Lemma 4.7 (Reparametrization invariance) If θ has Jeffreys prior and $\phi = h(\theta)$ is a smooth reparametrization, then ϕ also has Jeffreys prior.

Proof. For $p = 1$ (monotone differentiable h). Write $\Pi(\phi \le t) = \Pi(\theta \le h^{-1}(t))$. Differentiating:

$$\pi(\phi) = \pi(h^{-1}(\phi))\,\frac{d\,h^{-1}(\phi)}{d\phi} = \pi(\theta)\,\frac{d\theta}{d\phi} \propto \sqrt{I(\theta)}\,\frac{d\theta}{d\phi}.$$

Since

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\phi}\log f_\theta(X)\cdot\frac{d\phi}{d\theta}\right)^2\right] = \left(\frac{d\phi}{d\theta}\right)^2 I(\phi),$$

we obtain $\pi(\phi) \propto \sqrt{I(\phi)}$. For $p > 1$, the same conclusion follows from the change-of-variables formula $\pi(\phi) = \pi(\theta)\,|\det(\partial\theta/\partial\phi)|$ and the transformation rule $I(\phi) = J^\top I(\theta) J$ with $J = \partial\theta/\partial\phi$. □

Example 4.8 (Poisson mean) For $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$, one-observation information is $I_{X_1}(\lambda) = 1/\lambda$, so $I_n(\lambda) = n/\lambda$. Jeffreys prior:

$$\pi(\lambda) \propto \sqrt{\frac{n}{\lambda}} \propto \lambda^{-1/2}, \qquad \lambda > 0,$$

which is improper since $\int_0^\infty \lambda^{-1/2} d\lambda = \infty$. Note that this does not depend on n (so we obtain the same Jeffreys prior $\forall n \ge 1$).

§4.3 Frequentist analysis of Bayesian methods

We can evaluate Bayesian procedures under the frequentist model $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_{\theta_0}$. We first compute the Bayesian posterior as usual by conditioning on the data, and then treat the posterior as a random quantity due to the randomness in $X_1, \dots, X_n$.

Example 4.9 (Posterior mean under a normal prior) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$ and prior $\theta \sim N(0,1)$. The posterior is

$$\pi(\theta \mid x_1, \ldots, x_n) = N\left(\frac{n}{n+1}\,\bar{x},\ \frac{1}{n+1}\right).$$

The posterior mean $\bar{\theta}_n = \frac{n}{n+1}\bar{X}_n$ differs slightly from the MLE $\hat{\theta}_{ML} = \bar{X}_n$. Under the frequentist model with true θ_0:

$$\bar{\theta}_n = \frac{n}{n+1}\bar{X}_n \xrightarrow{P} \theta_0 \quad \text{(consistency)},$$

and

$$\begin{aligned}\sqrt{n}(\bar{\theta}_n - \theta_0) &= \sqrt{n}(\bar{\theta}_n - \hat{\theta}_{ML}) + \sqrt{n}(\hat{\theta}_{ML} - \theta_0) \\ &= -\frac{\sqrt{n}}{n+1}(\bar{X}_n - \theta_0 + \theta_0) + \sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0,1),\end{aligned}$$

since the first term converges in probability to 0 and the second is $N(0,1)$ by the CLT (or theorem 3.6). Thus the posterior mean is asymptotically normal with the same limiting variance as the MLE.

Bayesian statistics is especially popular for uncertainty quantification; a $100(1-\alpha)\%$ **posterior credible set** is any set $C \subseteq \Theta$ with

$$\Pi(C \mid X) = 1 - \alpha.$$

Credible sets are often simpler to compute than frequentist confidence sets (see section 6.3). In general, understanding the performance of such credible sets requires studying the posterior $\Pi(\cdot \mid X_1, \ldots, X_n)$ as a random probability distribution (this is beyond the scope of these notes however).

The posterior also satisfies a much stronger form of asymptotic normality, given by the **Bernstein-von Mises theorem**. This theorem states that for large n, the posterior behaves like a normal distribution centered at an efficient estimator (such as the MLE) and variance equal to the Cramer-Rao lower bound.

Explicitly, under regularity conditions for a parametric model $\{f_\theta : \theta \in \Theta \subset \mathbb{R}\}$, and a prior density π continuous and positive at the true θ_0, if $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_{\theta_0}$, then

$$\sup_A \left| \Pi(A \mid X_1, \ldots, X_n) \ - \ \Pr\Big(N\Big(\hat{\theta}_{ML},\, I(\theta_0)^{-1}/n\Big) \in A\Big) \right| \xrightarrow{P} 0,$$

as $n \to \infty$, where the supremum is over Borel sets.

As a consequence, any $100(1-\alpha)\%$ posterior credible set is also an *asymptotic* $100(1-\alpha)\%$ frequentist confidence set (see see section 6.3), providing a frequentist justification for Bayesian uncertainty quantification.

§5 Optimality in Estimation

§5.1 Decision theory, Bayes risk and minimax risk

Up to now, we have discussed desirable properties of estimators, such as unbiasedness, small mean squared error, consistency and asymptotic normality. These properties capture only certain aspects of performance. To provide a general framework for evaluating and comparing statistical procedures, we now adopt the perspective of *decision theory*.

Given a statistical model $\{f_\theta : \theta \in \Theta\}$ and data $X \in \mathcal{X}$, many statistical problems can be viewed as **decision problems** with an **action space** $\mathcal{A}$ and **decision rules** $\delta : \mathcal{X} \to \mathcal{A}$.

Example 5.1 Common statistical tasks naturally fit into this framework:

- **Estimation:** $\mathcal{A} = \Theta$ and $\delta(X) = \hat{\theta}(X)$ is an estimator.
- **Hypothesis testing:** $\mathcal{A} = \{0, 1\}$ and $\delta(X)$ is a test that selects a hypothesis.
- **Uncertainty quantification:** $\mathcal{A}$ = subsets of Θ, and $\delta(X) = C(X)$ is a confidence set.

We need a measure to assess the performance of a decision rule $\delta(X)$, and thus give a notion of optimality. To compare different decision rules, we quantify how ‘costly’ an action $a \in \mathcal{A}$ is when the true parameter is θ.

Definition 5.2 (Loss function) — A **loss function** is a non-negative function

$$L : \mathcal{A} \times \Theta \to [0, \infty)$$

that measures the cost of taking action $a \in \mathcal{A}$ when the true parameter is θ.

Example 5.3 Common loss functions include:

- Squared error loss (estimation): $L(a, \theta) = (a - \theta)^2$.
- Absolute error loss: $L(a, \theta) = |a - \theta|$.
- $0 - 1$ loss (hypothesis testing): if $a, \theta \in \{0, 1\}$, $L(a, \theta) = 1\{a \neq \theta\}$.

Since the decision depends on the random data X, the loss is also random, so it is natural to consider the *expected* loss function under the distribution of X.

Definition 5.4 (Risk function) — For loss function L, a decision rule δ and an observation $X \sim f_\theta$, the **risk function** is

$$R(\delta, \theta) = \mathbb{E}_\theta \left[L(\delta(X), \theta) \right] = \int_{\mathcal{X}} L(\delta(x), \theta) f_\theta(x) \, dx.$$

This measures the average performance of δ when the true parameter is θ.

Minimizing this expected loss gives one notion of optimality.
Since the risk function is a function of the parameter θ, different decision rules may each perform better in different regions of the parameter space, so a single estimator rarely achieves the smallest risk for all θ.

Example 5.5 If $X \sim \text{Binomial}(n, \theta)$ with $\theta \in [0,1]$, consider:

$$\hat{\theta}_1(X) = X/n, \qquad \hat{\theta}_2(X) = 1/2.$$

Then

$$R(\hat{\theta}_1, \theta) = E_\theta\left[(\hat{\theta}_1(X) - \theta)^2\right] = \frac{1}{n}\text{Var}_\theta(X) = \frac{\theta(1-\theta)}{n},$$
$$R(\hat{\theta}_2, \theta) = E_\theta\left[(\hat{\theta}_2(X) - \theta)^2\right] = \left(\theta - \frac{1}{2}\right)^2.$$

Comparing these risk functions for $\theta \in [0,1]$, we observe $R(\hat{\theta_2}, \theta) \leq R(\hat{\theta_1}, \theta)$ if and only if $\theta \in [\frac{1}{2} - \frac{1}{\sqrt{n+1}}, \frac{1}{2} + \frac{1}{\sqrt{n+1}}]$. This means always guessing $\frac{1}{2}$ is objectively better if the true parameter is close to $\frac{1}{2}$, though if not it does very badly otherwise.

This shows that we cannot expect one estimator to always dominate uniformly over the parameter space Θ, however there are some estimators that are genuinely worse than others.

Definition 5.6 (Admissibility) — A decision rule δ is **inadmissible** if there exists another rule $\delta^*(X)$ such that

$$R(\delta^*, \theta) \leq R(\delta, \theta) \quad \forall \theta \in \Theta,$$

and strict inequality holds for some θ. If no such δ^* exists, then δ is **admissible**.

Clearly then, inadmissible decision rules are not good, since one can then find a decision rule that performs at least as well for every parameter value, and strictly better for some parameter value.

Example 5.7 Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$. The MLE is $\hat{\lambda}_{ML} = 1/\bar{X}_n$, which is biased since

$$\mathbb{E}_\lambda[\hat{\lambda}_{ML}] = \frac{n}{n-1}\lambda.$$

An unbiased estimator is

$$\tilde{\lambda} = \frac{n-1}{n}\hat{\lambda}_{ML}.$$

This estimator has strictly smaller MSE than $\hat{\lambda}_{ML}$ for all $\lambda > 0$, so $\hat{\lambda}_{ML}$ is inadmissible under the squared error loss.

Remark 5.8 Showing that an estimator is admissible or not is generally difficult. However, some estimators (including certain Bayes estimators and shrinkage estimators) are known to be admissible. We will explore this shortly.

§5.1.1 Bayes risk

To compare risks across the entire parameter space, we can average risk with respect to a prior $\pi(\theta)$ on Θ.

Definition 5.9 (Bayes risk and Bayes rule) — Given a prior $\pi(\theta)$ and loss L, the **π-Bayes risk** of a decision rule δ is

$$R_\pi(\delta) = \mathbb{E}_{\theta\sim\pi}[R(\delta,\theta)] = \int_\Theta R(\delta,\theta)\,\pi(\theta)\,d\theta = \int_\Theta \int_{\mathcal{X}} L(\delta(x),\theta)\,f_\theta(x)\,dx\,\pi(\theta)\,d\theta.$$

A **π-Bayes decision rule** δ_π is any rule minimizing $R_\pi(\delta)$.

Note that the expectation above is over θ, not over X. In estimation problems where $\delta(X) = \hat{\theta}(X)$ is an estimator, the π-Bayes rule is called the **Bayes estimator** of θ.

Example 5.10 (Uniform-prior Bayes risk for Binomial mean) Let $X \sim \text{Binomial}(n,\theta)$ with prior $\theta \sim \text{Unif}[0,1]$. Under squared error loss and estimator $\hat{\theta}_1(X) = X/n$,

$$R(\hat{\theta}_1,\theta) = \frac{\theta(1-\theta)}{n} \quad\Rightarrow\quad R_\pi(\hat{\theta}_1) = \frac{1}{n}\int_0^1 \theta(1-\theta)\,d\theta = \frac{1}{6n}.$$

Since Bayesians update their prior π to the posterior $\pi(\theta \mid x)$ after observing x, it is therefore natural to update the risk function based on the posterior also.

Definition 5.11 (Posterior risk) — Given an observation $x \in \mathcal{X}$, the **posterior risk** is defined as the average loss under the posterior distribution of δ,*i.e.*

$$R_\pi(\delta(x)) = \mathbb{E}_\pi[L(\delta(x),\theta) \mid x] = \int_\Theta L(\delta(x),\theta)\,\pi(\theta \mid x)\,d\theta.$$

Note that this expectation above is now taken over θ instead, and $R_\pi(\delta(x)$ is a function of the observation $x \in \mathcal{X}$.

The example below shows how one can sometimes minimize the posterior risk explicitly.

Example 5.12 For $L(a,\theta) = (a-\theta)^2$ and fixed x,

$$R_\pi(\delta(x)) = E\left[(\delta(x)-\theta)^2|x\right] = \int_\Theta (\delta(x)^2 - 2\delta(x)\theta + \theta^2)\,\pi(\theta \mid x)\,d\theta.$$

This is a quadratic in $\delta(x)$ minimized when

$$\frac{d}{d\delta(x)}R_\pi(\phi(x)) = \frac{d}{d\delta(x)}\left[\delta(x)^2 - 2\delta(x)\int_\Theta \theta\,\pi(\theta \mid x)\,d\theta\right] = 2\delta(x) - 2\int_\Theta \theta\,\pi(\theta \mid x)\,d\theta = 0.$$

This is minimized at the posterior mean, $\delta(x) = \int_\Theta \theta\,\pi(\theta \mid x)\,d\theta = \mathbb{E}_\pi[\theta \mid x]$.

Proposition 5.13 If δ minimizes the π-posterior risk pointwise in x, then δ minimizes the π-Bayes risk.

Proof. Write the Bayes risk as

$$\begin{aligned} R_\pi(\delta) &= \int_\Theta\int_{\mathcal{X}} L(\delta(x),\theta)\, f_\theta(x)\, dx\, \pi(\theta)\, d\theta \\ &= \int_{\mathcal{X}} \underbrace{\left(\int_\Theta L(\delta(x),\theta)\, \pi(\theta \mid x)\, d\theta\right)}_{R_\pi(\delta(x))} \underbrace{\left(\int_\Theta f_\theta(x)\, \pi(\theta)\, d\theta\right)}_{\varphi(x)} dx, \end{aligned}$$

where $\varphi(x) = f_\pi(x)$ is the prior predictive density and $\varphi(x) \geq 0$. If δ_π minimizes $R_\pi(\delta(x))$ for every x, then multiplying by $\varphi(x)$ and integrating in x yields $R_\pi(\delta_\pi) \leq R_\pi(\delta)$. □

It is typically easier to minimize the posterior risk instead of the Bayes risk, and this is convenient since the previous proposition 5.13 shows that this immediately gives a minimizer of the Bayes risk too.

Proposition 5.14 Suppose δ_π minimizes $R_\pi(\delta)$ and $R_\pi(\delta_\pi) < \infty$. Then $\delta_\pi(x)$ minimizes the posterior risk $R_\pi(\delta(x))$ (for f_π-almost every x).

Example 5.15 (Beta–Bernoulli Bayes estimator) If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ with prior $\theta \sim \text{Beta}(\alpha,\beta)$, then $\theta \mid x \sim \text{Beta}(\sum_i x_i + \alpha,\ n - \sum_i x_i + \beta)$ and the Bayes estimator (squared error) is

$$\hat\theta_{\text{Bayes}} = \mathbb{E}[\theta \mid x] = \frac{\sum_i x_i + \alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta}\bar X_n + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta}.$$

It is close to the MLE $\bar X_n$ for large n.

Proposition 5.16 If a Bayes estimator $\hat\theta_{\text{Bayes}}$ is unique, then it is admissible.

§5.1.2 Minimax risk

Another global criterion considers the *worst-case* risk over the entire parameter space Θ.

Definition 5.17 (Minimax risk and minimax rule) — The **minimax risk** is

$$\inf_\delta \sup_{\theta\in\Theta} R(\delta,\theta).$$

Any decision rule achieving this value is called **minimax**.

A unique minimax rule is automatically admissible. Generally, finding minimax rules is hard. The following basic inequality relates Bayes risk and the worst-case risk.

Lemma 5.18 For any decision rule δ and prior π,

$$R_\pi(\delta) \;\leq\; \sup_{\theta\in\Theta} R(\delta,\theta).$$

Proof. By definition, $R_\pi(\delta) = \mathbb{E}_\pi[R(\delta, \theta)] \leq \sup_\theta R(\delta, \theta)$. □

Proposition 5.19 Let π be a prior on Θ such that for its Bayes rule δ_π,

$$R_\pi(\delta_\pi) \;=\; \sup_{\theta \in \Theta} R(\delta_\pi, \theta).$$

Then δ_π is minimax. If, in addition, δ_π is the unique π-Bayes rule, then δ_π is the unique minimax.

Proof. For any rule δ^*,

$$\sup_\theta R(\delta^*, \theta) \;\geq\; R_\pi(\delta^*) \;\geq\; R_\pi(\delta_\pi) \;=\; \sup_\theta R(\delta_\pi, \theta),$$

where the first inequality is lemma 5.18 and the second uses Bayes optimality of δ_π. Taking $\inf_{\delta^*}$ proves minimaxity. Uniqueness follows since $R_\pi(\delta^*) > R_\pi(\delta_\pi)$ for $\delta^* \neq \delta_\pi$ if the Bayes rule is unique. □

Therefore, if the maximal risk of a Bayes rule equals the Bayes risk, the corresponding Bayes rule is minimax.

Corollary 5.20 If a (unique) Bayes rule δ_π has constant risk in θ, then it is (unique) minimax.

Proof. If $R(\delta_\pi, \theta)$ is constant in θ, then $R_\pi(\delta_\pi) = \mathbb{E}_\pi[R(\delta_\pi, \theta)] = \sup_\theta R(\delta_\pi, \theta)$, so it is minimax by proposition 5.19. □

Example 5.21 (Minimax estimator for a Bernoulli mean) For $\theta \in [0, 1]$, consider squared-error estimation with prior $\theta \sim \mathrm{Beta}(\alpha, \beta)$. The Bayes estimator is

$$\bar{\theta}_{\alpha,\beta}(x) = \frac{\sum_i x_i + \alpha}{n + \alpha + \beta}.$$

If one can choose α, β so that $R(\bar{\theta}_{\alpha,\beta}, \theta)$ is constant in θ, then by corollary 5.20 $\bar{\theta}_{\alpha,\beta}$ is (unique) minimax. Note that this minimax rule differs from the MLE $\bar{X}_n$.

Lemma 5.22 If δ is admissible and has constant risk, then δ is minimax.

§5.2 Minimum variance unbiased estimators

We have now seen several notions of optimality via decision theory, where we find an optimal estimator by minimizing some concept of risk: a *Bayes estimator* minimizes the average risk with respect to a prior, while a *minimax estimator* minimizes the worst-case risk. In this section, we focus on optimality within the class of unbiased estimators under squared error loss

$$L(a, \theta) = (a - \theta)^2.$$

For unbiased estimators, the risk reduces to

$$R(\hat{\theta}, \theta) = E_\theta[(\hat{\theta}(X) - \theta)^2] = \mathrm{Var}_\theta(\hat{\theta}),$$

so minimizing risk corresponds to minimizing the estimator's variance.

Definition 5.23 (Uniformly minimum variance unbiased estimator (UMVUE)) — Let $X \sim P_\theta$ for a family $\{P_\theta : \theta \in \Theta\}$, and let $g(\theta)$ be a parameter of interest. An unbiased estimator $\hat{g}(X)$ of $g(\theta)$ is called a **uniformly minimum variance unbiased estimator (UMVUE)** if

$$\mathrm{Var}_\theta(\hat{g}) \leq \mathrm{Var}_\theta(\tilde{g}) \quad \forall\, \theta \in \Theta,$$

for any other unbiased estimator $\tilde{g}(X)$ of $g(\theta)$.

Since θ is unknown, the inequality must hold for all $\theta \in \Theta$.

Remark 5.24 Recall that the Cramer-Rao lower bound provides a theoretical lower bound on the variance of any unbiased estimator, hence if an estimator achieves this bound for every $\theta \in \Theta$, then it is a UMVUE.

Note the converse however; not every UMVUE comes from the Cramer-Rao lower bound, since it is possible for the Cramer-Rao lower bound to be unattainable (of course, if the model is an exponential family however, then this bound is in fact attainable by proposition 2.21).

§5.2.1 Unbiased estimation

Unbiasedness is often desirable in an estimator, although unbiased estimators can have potential drawbacks.

Example 5.25 (Unbiased estimators may not exist) Let $X \sim \mathrm{Binom}(n, \theta)$ for $0 < \theta < 1$, and consider $g(\theta) = 1/\theta$. Suppose an unbiased estimator $T(X)$ existed such that

$$\mathbb{E}_\theta[T(X)] = \sum_{x=0}^{n} T(x) \binom{n}{x} \theta^x (1-\theta)^{n-x} = \frac{1}{\theta}, \quad \forall\, 0 < \theta < 1.$$

As $\theta \to 0$, the left-hand side tends to $T(0)$ (finite), while the right-hand side diverges to ∞. Hence no such unbiased estimator $T(X)$ exists.

Example 5.26 (Unbiased estimators can be nonsensical) Let $X \sim \text{Poisson}(\lambda)$ and consider $g(\lambda) = e^{-2\lambda}$. We seek $T(X)$ satisfying

$$\mathbb{E}_\lambda[T(X)] = \sum_{x=0}^{\infty} T(x)\,\frac{e^{-\lambda}\lambda^x}{x!} = e^{-2\lambda}, \quad \forall\, \lambda > 0.$$

Multiplying both sides by e^{λ} gives

$$\sum_{x=0}^{\infty} T(x)\frac{\lambda^x}{x!} = e^{-\lambda} = \sum_{x=0}^{\infty} \frac{(-1)^x \lambda^x}{x!},$$

so $T(x) = (-1)^x$. Thus, $T(X)$ equals 1 if X is even and -1 if X is odd. Since $0 < e^{-2\lambda} < 1$, such estimates are meaningless and not even within the parameter range.

In the previous example, if instead we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poiss}(\lambda)$, then using the sufficient statistic $\sum_i X_i = n\bar{X}_n$, the unbiased estimator

$$T(X) = \left(1 - \frac{2}{n}\right)^{n\bar{X}_n}$$

satisfies $\mathbb{E}_\lambda[T(X)] = e^{-2\lambda}$. For large n, $T(X) \approx e^{-2\bar{X}_n} \approx e^{-2\lambda}$, making it far more reasonable.

Example 5.27 (Unbiased estimators are not unique) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$, $\theta > 0$. Then both $2\bar{X}_n$ and $\frac{n+1}{n}\max_i\{X_i\}$ are unbiased estimators of θ.

It can further be shown that $\frac{n+1}{n}X_{(n)}$ has smaller variance and hence is better under squared error loss.

The above example raises a key question: does $\frac{n+1}{n}X_{(n)}$ have *minimum* variance among *all* unbiased estimators of θ?

The **Rao–Blackwell theorem** provides a constructive way to obtain the best unbiased estimator given a sufficient statistic; namely, conditioning the current estimator on the sufficient statistic and taking its expectation reduces the variance of the original estimator.

Moreover, the best possible Rao–Blackwell estimator arises when conditioning on a **minimal sufficient statistic**. Thus, any candidate for the UMVUE must be a function of a minimal sufficient statistic $T(X)$, since otherwise one could further reduce its variance by conditioning on T.

§5.3 Complete statistics

We have already seen that sufficient statistics allow us to compress the data X without losing information about the parameter θ. In this section, we add one more crucial property: **completeness**. Completeness, together with sufficiency, will give us a powerful optimality result: the **Lehmann-Scheffe theorem**.

Definition 5.28 (Complete statistic) — Let X have distribution P_θ from a parametric family $\{P_\theta : \theta \in \Theta\}$, and let $T = T(X)$ be a statistic. We say T is **complete** if for every (measurable) function g,

$$\mathbb{E}_\theta[g(T)] = 0 \quad \forall \theta \in \Theta \qquad \Longrightarrow \qquad P_\theta(g(T) = 0) = 1 \quad \forall \theta \in \Theta,$$

i.e. the only unbiased estimator of 0 that is a function of T is the zero function (almost surely).

To motivate this definition, think of each distribution of T under θ as a "direction" in a vector space. A family of vectors $v_1, \ldots, v_n$ in $\mathbb{R}^n$ is sometimes called **complete** (or *basis*) if it spans $\mathbb{R}^n$, meaning that the only vector orthogonal to all of them is the zero vector. Analogously, in the discrete case, completeness of T means,

$$\sum_t g(t)\, P_\theta(T = t) = 0 \quad \forall \theta \in \Theta \qquad \Longrightarrow \qquad g(t) = 0,$$

where this sum can be viewed as an inner product. As such, completeness means the family $\{P_\theta : \theta \in \Theta\}$ provides a sufficiently 'rich' set of 'vectors' $\{P_\theta(T = t)\}_{t \in \mathcal{T}}$ to give an orthogonality condition.

This is important:, since completeness is a property of the entire model $\{P_\theta : \theta \in \Theta\}$, not just of T.

Example 5.29 (Uniform case: maximum is complete) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$ with $\theta > 0$, and define $T = \max_i X_i$. Then T has pdf

$$f_T(t) = \frac{n}{\theta^n} t^{n-1} \mathbf{1}_{(0,\theta)}(t).$$

Suppose g satisfies

$$\mathbb{E}_\theta[g(T)] \;=\; \int_0^\theta g(t)\, \frac{n}{\theta^n} t^{n-1}\, dt \;=\; 0 \quad \forall\, \theta > 0.$$

Equivalently,

$$\int_0^\theta g(t)\, t^{n-1}\, dt = 0 \quad \forall\, \theta > 0.$$

Differentiate both sides with respect to θ:

$$g(\theta)\, \theta^{n-1} = 0 \quad \forall\, \theta > 0 \qquad \Longrightarrow \qquad g(\theta) = 0 \quad \forall\, \theta > 0.$$

Thus $P_\theta(g(T) = 0) = 1$, and so $T = \max_i X_i$ is complete for θ.

A broad source of complete statistics comes from exponential families. Statistics of the form $T = (T_1(X), \ldots, T_k(X))$ (where $T_i(X)$ are the corresponding T_i's from the exponential family definition 1.5) are complete for *most* k-parameter exponential family joint distributions. (Strictly speaking, this does not hold when the exponential family does not have full rank, but this not apply for most of the common distributions.)

Theorem 5.30 If T is sufficient and complete for θ, then T is minimal sufficient.

Proof (sketch). We want to show that if S is any other sufficient statistic, then T is a function of S (*a.s.*). For simplicity assume T is one-dimensional. Define

$$W(s) = \mathbb{E}_\theta[T \mid S = s], \qquad Y(t) = \mathbb{E}_\theta[W(S) \mid T = t].$$

By sufficiency, these conditional expectations do not depend on θ.
Step 1: show $P_\theta(T = Y(T)) = 1$ for all θ. Indeed,

$$\mathbb{E}_\theta[Y(T)] = \mathbb{E}_\theta\big[\mathbb{E}_\theta[W(S) \mid T]\big] = \mathbb{E}_\theta[W(S)] = \mathbb{E}_\theta\big[\mathbb{E}_\theta[T \mid S]\big] = \mathbb{E}_\theta[T].$$

So $\mathbb{E}_\theta[Y(T) - T] = 0$. Since T is complete, the only way $\mathbb{E}_\theta[g(T)] = 0$ for all θ is if $g(T) = 0$ *a.s.*, so $Y(T) = T$ *a.s.* .
Step 2: show $P_\theta(W(S) = Y(T)) = 1$ for all θ. By the tower property, $\mathbb{E}_\theta[Y(T) \mid S] = W(S)$. But from step 1, $Y(T) = T$ *a.s.*, so the conditional variance $\mathrm{Var}_\theta(Y(T) \mid S)$ must be 0 *a.s.* Hence $W(S) = Y(T) = T$ *a.s.* . This shows $T = W(S)$ *a.s.*, *i.e.* T is a function of S with probability 1, so T is minimal sufficient. □

Remark 5.31 The converse of this last theorem is not true, since one can have a minimal sufficient statistic that is not complete (see example 5.36)

While sufficient statistics contain all information about θ present in a sample, a statistic that carries no information about θ is called **ancillary** (this term of 'ancillary' is also commonly used in industry in this same context).

Definition 5.32 (Ancillary statistic) — A statistic is an **ancillary statistic** if its distribution does not depend on θ.

Example 5.33 If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, then $\bar{X}_n \sim N(\theta, 1/n)$ and so $\bar{X}_n$ is *not* ancillary.

However, the sample range $\max_i X_i - \min_i X_i$ has a distribution that depends only on the noise distribution, not on θ, so it *is* ancillary.

If T is complete, it turns out from the following theorem that T contains no ancillary information about θ, and thus a complete statistic T is independent of *any* ancillary statistic.

Theorem 5.34 (Basu's Theorem) If T is sufficient and complete for θ, and V is ancillary, then T and V are independent.

Proof (sketch). We want to show $P_\theta(V \in A \mid T) = P_\theta(V \in A) \quad$ *a.s.*, for all measurable sets A. Let $g(T) = P_\theta(V \in A \mid T) - P_\theta(V \in A)$. Because V is ancillary, $P_\theta(V \in A)$ does not depend on θ. Because T is sufficient, $P_\theta(V \in A \mid T)$ also does not depend on θ. Thus, $g(T)$ does not depend on θ. Also,

$$\mathbb{E}_\theta[g(T)] = \mathbb{E}_\theta[P_\theta(V \in A \mid T) - P_\theta(V \in A)] = P_\theta(V \in A) - P_\theta(V \in A) = 0.$$

By completeness of T, $g(T) = 0$ *a.s.*, *i.e.* conditional and unconditional probabilities match. Therefore, T and V are independent. □

Remark 5.35 A complete sufficient statistic T is a most efficient representation of the information about θ in the sample, since it is independent of any random variable whose distribution does not depend on θ (*i.e.* θ captures *all* the θ-information in the data).

Example 5.36 (Shifted uniform - not complete) Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(\theta - 1, \theta + 1)$ with $\theta \in \mathbb{R}$. Then $T = (\min_i X_i, \max_i X_i)$ is (minimal) sufficient for θ, but not complete. Indeed one can check

$$\mathbb{E}_\theta[\max_i X_i] = \theta + \frac{n-1}{n+1}, \qquad \mathbb{E}_\theta[\min_i X_i] = \theta - \frac{n-1}{n+1},$$

so

$$\mathbb{E}_\theta\left[\max_i X_i - \min_i X_i - \frac{2(n-1)}{n+1}\right] = 0 \quad \forall \theta,$$

but that random variable is *not* (almost surely) 0. Thus, T is not complete.

In fact $\max_i X_i - \min_i X_i - \frac{2(n-1)}{n+1}$ is ancillary (its distribution does not depend on θ), which aligns with Basu's theorem failing here because T is not complete.

§5.3.1 The Lehmann-Scheffe theorem

We now combine sufficiency, completeness, and Rao–Blackwell to get an extremely strong result: if a complete sufficient statistic exists, then it automatically gives us the unique, best, unbiased estimator.

Theorem 5.37 (Lehmann-Scheffe Theorem) Let $T = T(X)$ be a sufficient and complete statistic for θ. Let $\tilde{g}(X)$ be any unbiased estimator of $g(\theta)$ with $\text{Var}_\theta(\tilde{g}) < \infty$ for all θ. Define

$$\hat{g}(T(X)) \;=\; \mathbb{E}[\tilde{g}(X) \mid T(X)],$$

then $\hat{g}$ is the **unique** uniformly minimum variance unbiased estimator (UMVUE) of $g(\theta)$.

Proof. Let V be another unbiased estimator of $g(\theta)$. By the Rao-Blackwell theorem, $V^*(T) = E[V \mid T]$ is unbiased and satisfies $\text{Var}_\theta(V^*) \le \text{Var}_\theta(V),\ \forall\, \theta \in \Theta$.

If we can show that $P_\theta(\hat{g} = V^*) = 1\ \forall\, \theta \in \Theta$, this establishes that there is a unique $\hat{g}$ satisfying $\text{Var}_\theta(\hat{g}) \le \text{Var}_\theta(V),\ \forall\, \theta \in \Theta$, as required.

Since both $\hat{g}$ and V^* are unbiased estimators of $g(\theta)$, $E_\theta[\hat{g} - V^*] = 0,\ \forall\, \theta \in \Theta$.

Since they are both functions of the complete statistic T, this implies $P_\theta(\hat{g} - V^* = 0) = 1,\ \forall\, \theta \in \Theta$, as desired. □

Remark 5.38 The theorem tells us: if a complete sufficient statistic T exists, then *any* unbiased estimator based on T is the unique UMVUE.

Moreover, the UMVUE need not attain the Cramer-Rao lower bound.

Lehmann-Scheffe gives two equivalent constructive approaches of finding the UMVUE in practice, when a complete and sufficient statistic T exists.

1. Start with an unbiased estimator $\tilde{g}(X)$ of $g(\theta)$, and set $\hat{g}(T) = \mathbb{E}[\tilde{g}(X) \mid T]$. This conditional expectation is the unique UMVUE of $g(\theta)$.

2. Alternatively, find a function $h(T)$ that satisfies $\mathbb{E}_\theta[h(T)] = g(\theta)$ for all θ. Then $h(T)$ is the unique the UMVUE of $g(\theta)$.

Example 5.39 Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with both μ and σ^2 unknown. We know that $T = (\bar{X}_n, S^2)$ is sufficient and complete for (μ, σ^2), where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Also, $\mathbb{E}[\bar{X}_n] = \mu$, $\mathbb{E}[S^2] = \sigma^2$.

Since $\bar{X}_n$ and S^2 are already functions of the complete sufficient statistic T, Lehmann-Scheffe implies that $\bar{X}_n$ is the UMVUE of μ, and S^2 is the UMVUE of σ^2.

Example 5.40 Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$ with $\theta > 0$. Then $T = X_{(n)} = \max_i X_i$ is sufficient and complete for θ. We know

$$\mathbb{E}_\theta[X_{(n)}] = \frac{n}{n+1}\theta,$$

so

$$\frac{n+1}{n} X_{(n)}$$

is unbiased for θ. Since it's a function of the complete sufficient statistic T, Lehmann-Scheffe tells us it is the UMVUE of θ.
Similarly, for $\mathbb{E}_\theta[X] = \theta/2$, the UMVUE is

$$\frac{n+1}{2n} X_{(n)}.$$

More generally, suppose we want an unbiased estimator of θ^r for some fixed $r \leq n$. We seek $h(T)$ with $\mathbb{E}_\theta[h(T)] = \theta^r$ for all $\theta > 0$. A calculation using the pdf of $T = X_{(n)}$ shows

$$h(t) = \frac{n+r}{n} t^r$$

satisfies $\mathbb{E}_\theta[h(T)] = \theta^r$. Therefore, $\frac{n+r}{n} X_{(n)}^r$ is the UMVUE of θ^r.

Remark 5.41 UMVUEs do not always exist. For instance, if no unbiased estimator exists for the target $g(\theta)$, then clearly no UMVUE can exist. Furthermore, if no complete statistic can be found, then finding the UMVUE is hard.

§5.3.2 Revisiting the Cramer-Rao lower bound

Below, we provide an example where the UMVUE does not attain the Cramer-Rao lower bound, and thus no unbiased estimator attains it.

Example 5.42 Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0,1)$. Then

$$\begin{aligned} f_\theta(x) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} &= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} \\ &= \exp\left\{ \underbrace{\log \frac{\theta}{1-\theta}}_{=:c(\theta)} \sum_{i=1}^n x_i \; + \; n\log(1-\theta) \right\}. \end{aligned}$$

This is an exponential family with natural statistic $T(X) = \sum_{i=1}^n X_i$, so an estimator attains the Cramer-Rao lower bound if and only if it is of the form $aT(X) + b$ (proposition 2.21).

It is not hard to show that the unbiased estimator $\bar{X}_n$ attains the lower bound and is the UMVUE, since $T(X)$ is a sufficient and complete statistic.

Suppose we wish to estimate $g(\theta) = \theta^2$. For an unbiased estimator $\tilde{g}$ of θ^2, the Cramer-Rao lower bound (theorem 2.16) gives

$$\text{Var}_\theta(\tilde{g}) \;\geq\; \frac{(g'(\theta))^2}{I_n(\theta)} \;=\; \frac{(2\theta)^2}{n\, I_{X_1}(\theta)} \;=\; \frac{4\theta^2}{n\,(1/[\theta(1-\theta)])} \;=\; \frac{4\,\theta^3(1-\theta)}{n}.$$

From completeness of $T = \sum_i X_i$, the UMVUE of θ^2 is

$$\hat{g}(T) \;=\; \frac{T(T-1)}{n(n-1)} \;=\; \frac{1}{n(n-1)}\Big(\sum_{i=1}^n X_i\Big)\Big(\sum_{i=1}^n X_i - 1\Big),$$

which is unbiased for θ^2. However, $\hat{g}(T)$ is *quadratic* in T, not affine of the form $a\,T + b$, and therefore it *cannot* achieve the Cramer-Rao lower bound.

§6 Hypothesis Testing and Confidence Intervals

§6.1 Hypothesis testing

§6.2 Uniformly most powerful tests and likelihood ratio tests

§6.2.1 Likelihood ratio tests

§6.3 Confidence intervals