Overview

- Likelihood principle (11 lectures)
- Bayesian inference (2 lectures)
- Decision theory (3 lectures)
- Multivariate analysis (2 lectures)
- Nonparametric inference & Monte Carlo techniques (6 lectures)

Books:

- Theory of point estimation - Lehmann & Casella
- “Asymptotic Statistics” - van der Vaart
- “Statistical Inference” - Casella & Berger
- “Intro to Multivariate Statistical Analysis” - Anderson

Introduction

Goal: Make inference about unknown probability distributions based on access to random samples.

Consider a real valued random variable X on a probability space Ω with distribution function

$$F(t) = \mathbb{P}(\omega \in \Omega : X(\omega) \leq t)\ \forall t \in \mathbb{R}$$

When X is discrete, $F(t) = \sum_{x \leq t} f(x)$, where f is the pmf of X.

When X is continuous, $F(t) = \int_{-\infty}^{t} f(s)\mathrm{d}s$, where f is the pdf of X.

For all the results in this course, we assume either pdf or pmf exists.

Often, the distribution of X is parameterised by an unknown value θ. The goal is to infer something about θ based on (iid) samples $X_1, \ldots, X_n$.

Definition. A *statistical model* for a sample from X is any family of probability distributions $\{P_\theta : \theta \in \Theta\}$ for the law of X. When P_θ has a pmf (pdf) $f(\cdot, \theta)$, this is also written as $\{f(\cdot, \theta) : \theta \in \Theta\}$. The index set Θ is the *parameter space*.

Example.

(i) $\mathcal{N}(\theta, 1)$; $\theta \in \Theta = \mathbb{R}$.

(ii) $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.

(iii) $\operatorname{Exp}(\theta)$; $\theta \in \Theta = (0, \infty)$.

(iv) $\mathcal{N}(\theta, 1)$; $\theta \in \Theta = [-1, 1]$.

Remark: for a variable X with distribution P, the model $\{P_\theta : \theta \in \Theta\}$ is *correctly specified* if there exists $\theta \in \Theta$ such that $P = P_\theta$. For instance, if $X \sim \mathcal{N}(2, 1)$, the model in (i) is correctly specified, but the model in (iv) is not.

In the case of a correctly specified model, we often use θ_0 to denote the "true value" of the parameter. We also say $\{X_1, \ldots, X_n\}$ are iid from a model $\{P_\theta : \theta \in \Theta\}$ in the case of a correctly specified model.

Statistical goals:

- Estimation: construct $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ such that $\hat{\theta}$ is close to θ_0 when $X_i \sim P_{\theta_0}$.
- Hypothesis testing: determine whether the null hypothesis $H_0 : \theta = \theta_0$ or the alternative hypothesis $H_1 : \theta \neq \theta_0$ is true, using a test $\psi_n = \psi(X_1, \ldots, X_n)$ such that $\psi_n = 0$ when H_0 is true and $\psi_n = 1$ when H_1 is true, with high probability.
- Inference: find confidence intervals (confidence sets) $\mathcal{C}_n = \mathcal{C}(X_1, \ldots, X_n)$ such that for some $0 < \alpha < 1$ we have $\mathbb{P}_\theta(\theta \in \mathcal{C}_n) \geq 1 - \alpha$, for all $\theta \in \Theta$, where α is the significance level.

1 The Likelihood Principle

Suppose $X_1, \ldots, X_n$ are iid from a Poisson model $\{\text{Poi}(\theta) : \theta \geq 0\}$ with numerical values $X_i = x_i$, for all $1 \leq i \leq n$. The joint distribution of the sample is

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

We can think of $L_n(\theta)$ as a random function from Θ to $\mathbb{R}$, where the randomness comes from $\{X_i\}_{i=1}^n$. This is the probability of occurence of the observed sample $(X_1 = x_1, \ldots, X_n = x_n)$, as a function of the unknown parameter θ.

The idea of the likelihood principle is to find θ which maximises $L_n(\theta)$, or equivalently $l_n(\theta) = \log L_n(\theta)$. In the example, we have

$$l_n(\theta) = -n\theta + \log(\theta) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!)$$

Setting $l_n'(\theta) = 0$ gives

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

and the solution is $\hat{\theta}_{\text{mle}} = \frac{1}{n}\sum_{i=1}^n x_i$, which is the sample mean. One can also check that $l_n''(\theta) < 0$ for all $\theta > 0$. When all X_i's are 0, one can check that maximising $l_n(\theta)$ is equivalent to maximising $-n\theta$, so $\hat{\theta}_{\text{mle}} = 0$ in this case.

Maximum likelihood estimator

Suppose $\{f(\cdot,\theta) : \theta \in \Theta\}$ is a statistical model of pdfs/pmfs for the distribution of a random variable X, and $X_1, \ldots, X_n$ are iid copies of X.

Define the *likelihood function*

$$L_n(\theta) = \prod_{i=1}^n f(x_i, \theta)$$

the *log likelihood function*

$$l_n(\theta) = \log L_n(\theta) = \sum_{i=1}^n \log f(x_i, \theta)$$

and the *normalised log likelihood function*

$$\bar{l}_n(\theta) = \frac{1}{n} l_n(\theta) = \frac{1}{n}\sum_{i=1} \log f(x_i, \theta)$$

Definition. The *maximum likelihood estimator* is any element $\hat{\theta} = \hat{\theta}_{\text{mle}} = \hat{\theta}_{\text{mle}}(X_1, \ldots, X_n) \in \Theta$ for which $L_n(\hat{\theta}) = \max_{\theta\in\Theta} L_n(\theta)$.

Remark: the definition of MLE can be generalised to non-iid data, provided a joint pdf/pmf of $(X_1, \ldots, X_n)$ can be specified.

Example.

(i) For $X_i \sim \text{Poi}(\theta)$, $\theta \geq 0$, we calculated $\hat{\theta}_{\text{mle}} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$.

(ii) For $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, we have $\hat{\mu}_{\text{mle}} = \bar{X}_n$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ (see Example sheet).

(iii) In the Gaussian linear model $Y = X\theta + \varepsilon$, with a known $X \in \mathbb{R}^{n\times p}$, unknown $\theta \in \mathbb{R}^p$, and $\varepsilon \sim \mathcal{N}(0, I_n)$, the observations $(Y_1, \ldots, Y_n)$ are not iid, but a joint distribution $f(Y_1, \ldots, Y_n; \theta)$ can still be specified. The MLE is the least squares estimator (see Example sheet).

Definition. For $\Theta \subseteq \mathbb{R}^p$ and l_n differentiable at θ, the *score function* S_n is

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

Solving for a root of $S_n(\theta)$ is a common heuristic for maximising $l_n(\theta)$. In many cases, it is a necessary and sufficient condition.

Note: derivatives are taken with respect to θ, <u>not</u> the x_i's.

Information geometry

Recall that if X is a random variable with distribution P_θ on some space $\mathcal{X} \subseteq \mathbb{R}^d$, and $g : \mathcal{X} \to \mathbb{R}$ is a function, then

$$E_\theta[g(X)] = \int_{\mathcal{X}} g(x) \mathrm{d}P_g(x) = \int_{\mathcal{X}} g(x) f(x,\theta) \mathrm{d}x$$

if X has a pdf $f(x,\theta)$, and

$$\mathbb{E}_\theta[g(X)] = \sum_{x \in \mathcal{X}} g(x) f(x,\theta)$$

if X has a pmf $f(x,\theta)$

Theorem 1.1. *Consider a model $\{f(\cdot,\theta) : \theta \in \Theta\}$, where $f(\cdot,\theta)$ is a pdf/pmf and $f(x,\theta) > 0$ for all x, θ. Also suppose the model is correctly specified, with θ_0 equal to the true parameter, and $\mathbb{E}_{\theta_0}[|\log(f(X,\theta))|] < \infty$ for all $\theta \in \Theta$. Then the function defined by $l(\theta) = \mathbb{E}_{\theta_0}[\log(f(X,\theta))]$ is maximised at θ_0.*

Proof. Consider the case when X has a pdf (discrete case is analogous). For all $\theta \in \Theta$, we have

$$\begin{aligned} l(\theta) - l(\theta_0) &= \mathbb{E}_{\theta_0}[\log(f(X,\theta))] - \mathbb{E}_{\theta_0}[\log(f(x,\theta_0))] \\ &= \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X,\theta)}{f(X,\theta_0)}\right)\right] \end{aligned}$$

<u>Jensen's inequality</u>: $\mathbb{E}[\varphi(Z)] \leq \varphi(\mathbb{E}[Z])$ for any random variable Z and concave function φ.

Since log is concave,

$$\begin{aligned} l(\theta) - l(\theta_0) &\leq \log\left(\mathbb{E}_{\theta_0}\left[\frac{f(X,\theta)}{f(X,\theta_0)}\right]\right) \\ &= \log\left(\int_{\mathcal{X}} \frac{f(x,\theta)}{f(x,\theta_0)} f(x,\theta_0) \mathrm{d}x\right) = \log 1 = 0 \qquad (*) \end{aligned}$$

□

Remark: under the assumption of "strict identifiability of the model parameterisation", i.e,

$$f(\cdot,\theta) = f(\cdot,\theta') \iff \theta = \theta'$$

the inequality $(*)$ is strict, since equality occurs in Jensen only when φ is linear or Z is constant.

Remark: the quantity $l(\theta_0) - l(\theta)$ computed above can be written as

$$\mathrm{KL}(P_{\theta_0}, P_\theta) = \int_{\mathcal{X}} f(x,\theta_0) \log\left(\frac{f(x,\theta_0)}{f(x,\theta)}\right) \mathrm{d}x$$

and is the Kullback-Leibler divergence in information theory. It is a "distance" between distributions. Maximising $l(\theta)$ is equivalent to minimising KL.

Fisher information

We consider the gradient and Hessian of the likelihood function.

Theorem 1.2. *For a parametric model $\{f(\cdot,\theta) : \theta \in \Theta\}$, "regular enough" so integration and differentiation can be interchanged, we have $\mathbb{E}_\theta[\nabla_\theta \log(f(X,\theta))] = 0$ for all $\theta \in \mathrm{int}(\Theta)$.*

Proof. We write the expectation

$$\begin{aligned}
\mathbb{E}_\theta[\nabla_\theta \log(f(X,\theta))] &= \int_{\mathcal{X}} (\nabla_\theta \log f(x,\theta)) f(x,\theta) \mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\nabla_\theta f(x,\theta)}{f(x,\theta)} f(x,\theta) \mathrm{d}x \\
&= \nabla_\theta \left(\int_X f(x,\theta) \mathrm{d}x \right) = \nabla_\theta(1) = 0
\end{aligned}$$

□

Remark: in particular, when $\theta_0 \in \mathrm{int}(\Theta)$, then $\mathbb{E}_{\theta_0}[\nabla_\theta \log(f(X,\theta))] = 0$.

Definition. For a parameter space $\Theta \subseteq \mathbb{R}^p$, the *Fisher information* matrix is defined by

$$I(\theta) = \mathbb{E}_\theta\left[(\nabla_\theta \log f(X,\theta)) (\nabla_\theta \log f(X,\theta))^T \right], \ \forall \theta \in \mathrm{int}(\Theta)$$

in other words,

$$I_{ij}(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i} \log f(X,\theta) \frac{\partial}{\partial\theta_j} \log f(X,\theta) \right]$$

Remark: in 1 dimension, we have

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

Thus I_{θ_0} describes random variations of $S_n(\theta_0)$ about its mean. This in turn will help quantify the precision of $\hat{\theta}$, a zero of $S_n(\hat{\theta}) = 0$, about θ_0.

Theorem 1.3. *Under the same regularity assumptions as the previous theorem*

$$I(\theta) = -\mathbb{E}_\theta\left[\nabla_\theta^2 \log(f(X,\theta))\right],\ \forall\theta \in \mathrm{int}(\Theta)$$

i.e,

$$I_{ij}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X,\theta)\right]$$

Proof. We write

$$\nabla_\theta^2 \log f(X,\theta) = \nabla_\theta\left(\frac{\nabla_\theta f(X,\theta)}{f(X,\theta)}\right) = \frac{\nabla_\theta^2 f(X,\theta)}{f(X,\theta)} - \frac{\nabla_\theta f(X,\theta)\nabla_\theta f(X,\theta)^T}{f(X,\theta)^2}$$

note that

$$\mathbb{E}\left[\frac{\nabla_\theta^2 f(X,\theta)}{f(X,\theta)}\right] = \int_{\mathcal{X}} \nabla_\theta^2 f(X,\theta)\mathrm{d}x = \nabla_\theta^2 \int_{\mathcal{X}} f(X,\theta)\mathrm{d}x = 0$$

Therefore

$$\begin{aligned}
-\mathbb{E}_\theta\left[\nabla_\theta^2 \log f(X,\theta)\right] &= \mathbb{E}_\theta\left[\frac{\nabla_\theta f(X,\theta)\nabla_\theta f(X,\theta)^T}{f^2(X,\theta)}\right] \\
&= \mathbb{E}\left[\frac{\nabla_\theta f(X,\theta)}{f(X,\theta)}\left(\frac{\nabla_\theta f(X,\theta)}{f(X,\theta)}\right)^T\right] \\
&= \mathbb{E}_\theta\left[(\nabla_\theta \log f(X,\theta))(\nabla_\theta \log f(X,\theta))^T\right] \\
&= I(\theta)
\end{aligned}$$

□

Remark: continuing the previous remark, in 1 dimension

$$\mathrm{Var}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X,\theta)\right] = I(\theta) = -\mathbb{E}_\theta\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X,\theta)\right]$$

this relates the variance of the score function and the curvature of l, both of which are relevant to describing the quality of the MLE $\hat{\theta}$ as an approximation to θ_0.

Suppose now $X = (X_1, \dots, X_n)$ is a vector of iid copies of a random variable. Let $I(\theta) = \mathbb{E}_\theta[(\nabla_\theta \log f(X_{i_1}, \theta))(\nabla_\theta \log f(X_{i_1}, \theta))^T]$ be the Fisher information of one copy of the random variable, and let

$$I_n(\theta) = \mathbb{E}_\theta\left[(\nabla_\theta \log f(X_1, \dots, X_n, \theta))(\nabla_\theta \log f(X_1, \dots, X_n, \theta))^T\right]$$

denotes the Fisher information of the random vector X.

Theorem 1.4. *In the setting described above, the Fisher information "tensorizes"*

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

Proof. By independence, $f(X_1, \dots, X_n, \theta) = \prod_{i=1}^n f(X_i, \theta)$. Then $\log f(X_1, \dots, X_n, \theta) = \sum_{i=1}^n \log f(X_i, \theta)$. We write

$$\begin{aligned} I_n(\theta) &= \mathbb{E}_\theta\left[(\nabla_\theta \log f(X_1, \dots, X_n, \theta))(\nabla_\theta \log f(X_1, \dots, X_n, \theta))^T\right] \\ &= \mathbb{E}_\theta\left[\left(\sum_{i=1}^n \nabla_\theta \log f(X_i, \theta)\right)\left(\sum_{i=1}^n \nabla_\theta \log f(X_i, \theta)\right)^T\right] \end{aligned}$$

Recall that $\mathbb{E}_\theta\left[\nabla_\theta \log f(X_i, \theta)\right] = 0$. Thus, by independence, all but the "diagonal" terms of the product remain, so

$$I_n(\theta) = \sum_{i=1}^n \mathbb{E}_\theta\left[(\nabla_\theta \log f(X_i, \theta))(\nabla_\theta \log f(X_i, \theta))^T\right] = nI(\theta)$$

□

Cramer-Rao bound

Theorem 1.5 (Cramer-Rao bound)**.** *Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a "regular" statistical model with $\Theta \subseteq \mathbb{R}$. Let $\tilde{\theta} = \tilde{\theta}(X_1, \dots, X_n)$ be an unbiased estimator of θ based on n iid observations from the model. For all $\theta \in \operatorname{int}(\theta)$, we have*

$$\operatorname{Var}_\theta(\tilde{\theta}) = \mathbb{E}_\theta\left[(\tilde{\theta} - \theta)^2\right] \geq \frac{1}{nI(\theta)}$$

Proof. Recall the Cauchy-Schwarz inequality:

$$(\mathbb{E}[YZ])^2 \leq \mathbb{E}[Y]^2\mathbb{E}[Z]^2$$

for random variables Y, Z. In particular, we will take $Y = \tilde{\theta} - \theta$ and $Z = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_1, \dots, X_n, \theta)$.

Note that $\mathbb{E}_\theta[Y^2] = \mathbb{E}_\theta\left[(\tilde{\theta} - \theta)^2\right]$. Also, by the previous theorem,

$$\mathbb{E}_\theta[Z^2] = I_n(\theta) = nI_n(\theta)$$

Furthermore,

$$\mathbb{E}_\theta[YZ] = \mathbb{E}_\theta\left[\tilde{\theta}\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1,\dots,X_n,\theta)\right] - \theta\,\underbrace{\mathbb{E}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1,\dots,X_n,\theta)\right]}_{=0}$$

$$= \int_{\mathcal{X}} \tilde{\theta}(X_1,\dots,X_n)\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}f(X_1,\dots,X_n,\theta)}{f(X_1,\dots,X_n,\theta)}f(X_1,\dots,X_n)\mathrm{d}x_1\dots\mathrm{d}x_n$$
$$= \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathcal{X}} \tilde{\theta}(X_1,\dots,X_n)f(X_1,\dots,X_n,\theta)\mathrm{d}x_1\dots\mathrm{d}x_n = \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[\tilde{\theta}] = 1$$

and the result follows from Cauchy-Schwarz. □

Remark: if $\tilde{\theta}$ is not unbiased, the same proof shows that

$$\mathrm{Var}_\theta(\tilde{\theta}) \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[\tilde{\theta}]\right)^2}{nI(\theta)}$$

The Cramer-Rao bound is about a variance of an estimate, hence is univariate in nature. Here is one multivariate generalisation. Suppose $\Theta \subseteq \mathbb{R}^p$ and $\Phi : \Theta \to \mathbb{R}$ is differentiable. Suppose $\tilde{\Phi}$ is an unbiased estimator of $\Phi(\theta)$ based on iid observations $(X_1,\dots,X_n)$ from a model $\{f(\cdot,\theta) : \theta \in \Theta\}$.

Theorem 1.6. *For all $\theta \in \mathrm{int}(\Theta)$, we have*

$$\mathrm{Var}_\theta(\tilde{\Phi}) \geq \frac{1}{n}\nabla_\theta\Phi(\theta)^T\left(I^{-1}(\theta)\right)\nabla_\theta\Phi(\theta)$$

Proof. Omitted. Can be derived using Cauchy-Schwarz. □

Example. Suppose $\Phi(\theta) = \alpha^T\theta$. Then $\nabla_\theta\Phi(\theta) = \alpha$ so the lower bound is

$$\mathrm{Var}_\theta(\tilde{\Phi}) \geq \frac{1}{n}\alpha^T I^{-1}(\theta)\alpha$$

In the example sheet, we will consider the special case of $\begin{pmatrix}X_1\\X_2\end{pmatrix} \sim \mathcal{N}(\theta,\Sigma)$ where $\theta = \begin{pmatrix}\theta_1\\\theta_2\end{pmatrix} \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2\times 2}$ is a known matrix. Let the sample size be $n = 1$.

Case 1: consider estimating θ_1 when θ_2 is known. This is a one-dimensional estimation problem, and we denote the Fisher information $I_1(\theta)$.

Case 2: consider estimating θ_1 when θ_2 is unknown. We can take $\Phi(\theta) = \theta_1 = \begin{pmatrix}1\\0\end{pmatrix}\theta$ in the theorem above to obtain a lower bound

$$I_\Phi(\theta) = \nabla_\theta\Phi(\theta)^T I(\theta)^{-1}\nabla_\theta\Phi(\theta)$$

of the variance of an unbiased estimator.

We will show that $I_1(\theta)^{-1} < I_\Phi(\theta)$, unless X_1 and X_2 are independent (i.e unless Σ is diagonal).

Asymptotic theory of the MLE

Cramer-Rao is concerned with unbiased estimators, but not all estimators, even MLE's are unbiased.

On the other hand, a reasonable property to expect is *asymptotic unbiasedness*: $\mathbb{E}_\theta[\tilde{\theta}_n] \to \theta$ as $n \to \infty$, when $\tilde{\theta}_n$ is computed from n iid samples from P_θ.

A stronger but related concept is *consistency*: $\tilde{\theta}_n \to \theta$ as $n \to \infty$ (where convergence is defined in a precise way to be discussed later).

For consistent estimators, a reasonable optimality criterion is *asymptotic efficiency*: $n \operatorname{Var}_\theta(\tilde{\theta}_n) \to I(\theta)^{-1}$ as $n \to \infty$, when $\tilde{\theta}_n$ is computed from n iid samples from P_θ (and $p = 1$).

Note that Cramer-Rao does <u>not</u> imply that $\liminf_{n\to\infty} n\,\mathrm{Var}_\theta(\tilde{\theta}_n) \geq I(\theta)^{-1}$ for any consistent estimator. However, this is true under appropriate regularity conditions.

Now, we will show that the MLE is always (under regularity conditions) asymptotically efficient. In fact

$$\hat{\theta}_{\text{mle}} \approx \mathcal{N}\left(\theta, \frac{I(\theta)^{-1}}{n}\right), \text{ for any } \theta \in \text{int}(\Theta) \text{ and } n \text{ sufficiently large}$$

Stochastic Convergence

We now introduce several basic definitions/results that will be used without proof.

Definition. Let $\{X_n\}_{n\geq 0}$ and X be random vectors in $\mathbb{R}^k$, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. So $X : \Omega \to \mathbb{R}^k$, $\mathcal{A}$ is the set of measurable sets ("events").

1. We say X_n converges to X *almost surely*, or $X_n \xrightarrow{\text{a.s}} X$ as $n \to \infty$, if
$$\begin{aligned}&\mathbb{P}(\omega \in \Omega : \|X_n(\omega) - X(\omega)\|_2 \to 0 \text{ as } n \to \infty)\\ &= \mathbb{P}(\|X_n - X\|_2 \to 0 \text{ as } n \to \infty) = 1\end{aligned}$$
2. We say that X_n converges to X *in probability*, or $X_n \xrightarrow{P} X$ as $n \to \infty$, if for all $\varepsilon > 0$,
$$\mathbb{P}(\|X_n - X\|_2 > \varepsilon) \to 0$$
3. We say that X_n converges to X *in distribution*, or $X_n \xrightarrow{d} X$ as $n \to \infty$, if
$$\mathbb{P}(X_n \prec t) \to \mathbb{P}(X \prec t), \ \forall t \text{ where } t \mapsto \mathbb{P}(X \prec t) \text{ is continuous}$$
we write $\{X \prec t\}$ as a shorthand for $\{X_{(1)} \leq t_1, \ldots, X_{(k)} \leq t_k\}$. For $k = 1$, this simply means
$$\mathbb{P}(X_n \leq t) \to \mathbb{P}(X \leq t)$$
i.e convergence of the usual cdf.

Theorem 1.7. *Almost sure convergence implies convergence in probability, which implies convergence in distribution. i.e*

$$X_n \xrightarrow{a.s} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X$$

Proof. See Probability & Measure. □

Theorem 1.8 (Continuous mapping theorem)**.** *If $\{X_n\}$ and X take values in $\mathcal{X} \subseteq \mathbb{R}^d$ and $g : \mathcal{X} \to \mathbb{R}$ is continuous, then*

$$X_n \xrightarrow{a.s/P/d} X \implies g(X_n) \xrightarrow{a.s/P/d} g(X)$$

Proof. See Probability & Measure. $\square$

Theorem 1.9 (Slutsky's lemma)**.** *Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, where c is deterministic (i.e non-stochastic). As $n \to \infty$, we have*

1. $Y_n \xrightarrow{P} c$
2. $X_n + Y_n \xrightarrow{d} X + c$
3. *When Y_n is one-dimensional, $X_n Y_n \xrightarrow{d} cX$, and if $c \neq 0$, $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$*
4. *If $\{A_n\}_{n\geq 0}$ are random matrices such that $\{A_n\}_{ij} \xrightarrow{P} A_{ij}$ for all (i,j), where A is deterministic, then $A_n X_n \xrightarrow{d} AX$*

Proof. See Probability & Measure. $\square$

Theorem 1.10. *If $X_n \xrightarrow{d} X$ as $n \to \infty$, then $\{X_n\}_{n\geq 0}$ is bounded in probability, or $X_n = 0_p(1)$: for all $\varepsilon > 0$, there exists $M(\varepsilon) < \infty$ such that for all $n \geq 0$*

$$\mathbb{P}(||X_n||_2 > M(\varepsilon)) < \varepsilon$$

Proof. See Probability & Measure. $\square$

Law of Large Numbers (LLN)

Many results in statistics are based on convergence of averages of iid random variables.

Theorem 1.11 (Weak LLN)**.** *Let $X_1, \dots, X_n$ be iid copies of X with $\mathrm{Var}(X) < \infty$. As $n \to \infty$, we have $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}(X)$.*

Theorem 1.12 (Strong LLN)**.** *Let $X_1, \dots, X_n$ be iid copies of $X \sim P$ on $\mathbb{R}^k$, such that $\mathbb{E}[||X||_2] < \infty$. Then as $n \to \infty$ we have*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{a.s} \mathbb{E}[X]$$

We only prove the weak law or large numbers:

Proof. We will apply Chebyshev's inequality:

$$\mathbb{P}(|Z - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(Z)}{\varepsilon^2}$$

where $\mu = \mathbb{E}[Z]$. Take $Z_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}(X))$ for a fixed $\varepsilon > 0$. Then

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X)| \geq \varepsilon) = \mathbb{P}(|Z_n| \geq \varepsilon) \leq \frac{\mathrm{Var}(Z_n)}{\varepsilon^2}$$

So if suffices to show $\mathrm{Var}(Z_n) \to 0$. By independence of the X_i's, we have

$$\mathrm{Var}(Z_n) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\mathrm{Var}(X)}{n} \to 0$$

since $\mathrm{Var}(X) < \infty$. $\square$

Central Limit Theorem (CLT)

We now present a finer-grained characterisation of the behaviour of $\bar{X}_n$. The stochastic fluctuations of $\bar{X}_n$ around $\mathbb{E}(X)$ are of the order $\frac{1}{\sqrt{n}}$ and look normally distributed.

Theorem 1.13 (CLT)**.** *Let $X_1, \dots, X_n$ be iid copies of $X \sim P$ on $\mathbb{R}$, such that* $\mathrm{Var}(X) = \sigma^2 < \infty$*. As $n \to \infty$, we have*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Proof. Omitted. □

Remark: the CLT is useful for constructing confidence intervals. Suppose $X_1, \dots, X_n$ is a sequence of iid copies of a random variable with mean μ_0 and variance σ^2, and let $\alpha \in (0,1)$. Define the confidence region

$$\mathcal{C}_n = \left\{\mu \in \mathbb{R} : |\mu - \bar{X}_n| \leq \frac{\sigma z_\alpha}{\sqrt{n}}\right\}$$

where z_α is defined such that $\mathbb{P}(|Z| \leq z_\alpha) = 1 - \alpha$, for $Z \in \mathcal{N}(0,1)$. Then we can compute

$$\begin{aligned}
\mathbb{P}(\mu_0 \in \mathcal{C}_n) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \mu_0}{\sigma}\right| \leq \frac{z_\alpha}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i - \mathbb{E}(\tilde{X})\right| \leq z_\alpha\right) \\
&\to \mathbb{P}(|Z| \leq z_\alpha) = 1 - \alpha
\end{aligned}$$

where $\tilde{X}_i = \frac{X_i - \mu_0}{\sigma}$, is a zero mean, variance 1 random variable. So $\mathcal{C}_n$ is an asymptotic level $(1 - \alpha)$ confidence interval.

Theorem 1.14 (Multivariate CLT)**.** *Let $X_1, \dots, X_n$ be iid copies of $X \sim P$ on $\mathbb{R}^k$, such that* $\text{Cov}(X) = \Sigma$ *is positive definite. As $n \to \infty$ we have*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

Remark: recall that a random vector $X \in \mathbb{R}^k$ has a normal distribution with mean $\mu \in \mathbb{R}^k$ and covariance $\Sigma \in \mathbb{R}^{k \times k}$, denoted by $X \sim \mathcal{N}(\mu, \Sigma)$, if the pdf is

$$f(x) = \frac{1}{(2\pi)^{k/2}} \frac{1}{|\det(\Sigma)|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Remark: as a consequence of one of the theorems above, we also have

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X) = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right)$$

Consistency of the MLE

Definition. Consider iid draws $X_1, \dots, X_n$ from the parametric model $\{P_\theta : \theta \in \Theta\}$. An estimator $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \dots, X_n)$ is *consistent* if $\tilde{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$, whenever the X_i's are drawn from P_θ. We also write $\tilde{\theta}_n \xrightarrow{P_\theta} \theta$.

We will show that the MLE is unique and consistent under the following regularity assumptions:

Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a statistical model of pdf's/pmf's on $\mathcal{X} \subseteq \mathbb{R}^d$ such that

1. $f(x, \theta) > 0$ for all $x \in \mathcal{X}$, $\theta \in \Theta$
2. The function $f(x, \cdot) : \theta \mapsto f(x, \theta)$ is continuous for all $x \in \mathcal{X}$.
3. The set $\Theta \subseteq \mathbb{R}^p$ is compact.
4. For any $\theta, \theta' \in \Theta$, $f(\cdot, \theta) = f(\cdot, \theta')$ if and only if $\theta = \theta'$ (strict identifiability)
5. $\mathbb{E}_\theta\left[\sup_{\theta'} |\log f(X, \theta')|\right] < \infty$ for all $\theta \in \Theta$.

These will be referred to as "the usual regularity conditions" in this course and its Examples sheets/Exams.

Remarks:

- Assumptions 1 and 4 are required to apply the strict version of Jensen's inequality to deduce that θ_0 is the unique maximum of $l(\theta) = \mathbb{E}_{\theta_0}[\log f(X, \theta)]$.
- Assumption 5 implies that continuity of the function $\theta \mapsto \log f(x, \theta)$ carries over to continuity of $\theta \mapsto \mathbb{E}_\theta[\log f(X, \theta)] = l(\theta)$, according to the Dominated Convergence Theorem.

Theorem 1.15 (*Dominated Convergence Theorem*). *If a sequence of (measurable) functions $\{f_n\}$ converges pointwise to a function $f : \mathcal{X} \to \mathbb{R}$ such that $|f_n(x)| \le g(x)$ for all $x \in \mathcal{X}$, for some function $g : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}[|g(X)|] < \infty$, where X is a random variable taking values in $\mathcal{X}$, then*

$$\mathbb{E}|f_n(X) - f(X)| \to 0 \text{ as } n \to \infty$$

In particular, for any sequence $\theta_n \to \theta$ in Θ, we can define $f_n(x) = \log(f(x, \theta_n))$ and $g(x) = \sup_{\theta'} |\log f(x, \theta')|$ and conclude that $l(\theta_n) \to l(\theta)$.

Theorem 1.16. *Let $X_1, \ldots, X_n$ be iid samples of a model $\{f(\cdot, \theta) : \theta \in \Theta\}$ satisfying the above assumptions. Then an MLE exists, and any MLE is consistent.*

Proof. Note that the mapping $\theta \mapsto \bar{l}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i, \theta)$ is continuous on the compact set Θ. Thus a maximiser exists, so the MLE is well-defined.

To prove consistency, let θ_0 denote the true parameter. We use (without proof) the fact that under the regularity assumptions, we have the uniform convergence

$$\sup_{\theta \in \Theta} |\bar{l}_n(\theta) - l(\theta)| \xrightarrow{P_{\theta_0}} 0$$

(This is somewhat stronger than the LLN, which concerns convergence just at fixed θ)

Now define $\Theta_\varepsilon = \{\theta \in \Theta : ||\theta - \theta_0||_2 \ge \varepsilon\}$, for arbitrary $\varepsilon > 0$. We will show that for any sequence of MLE's $\{\hat{\theta}_n\}$, we have $\mathbb{P}(\hat{\theta}_n \in \Theta_\varepsilon) \to 0$ as $n \to \infty$.

Note that since Θ_ε is the intersection of Θ with a closed set, it is also compact. Thus, there exists $\theta_\varepsilon \in \Theta_\varepsilon$ such that $l(\theta_\varepsilon) = \sup_{\theta \in \Theta_\varepsilon} l(\theta) := c(\varepsilon) < l(\theta_0)$, since θ_0 is the unique maximiser of l.

Let $\delta(\varepsilon) > 0$ be such that $\delta(\varepsilon) < \frac{l(\theta_0) - c(\varepsilon)}{2}$. We now write

$$\begin{aligned}
\sup_{\theta \in \Theta_\varepsilon} \bar{l}_n(\theta) &\le \sup_{\theta \in \Theta_\varepsilon} l(\theta) + \sup_{\theta \in \Theta_\varepsilon} \left(\bar{l}_n(\theta) - l(\theta)\right) \\
&\le \sup_{\theta \in \Theta_\varepsilon} l(\theta) + \sup_{\theta \in \Theta} \left|\bar{l}_n(\theta) - l(\theta)\right|
\end{aligned}$$

Consider the sequence of events

$$A_n(\varepsilon) = \left\{ \sup_{\theta \in \Theta} \left|\bar{l}_n(\theta) - l(\theta)\right| \le \delta(\varepsilon) \right\}$$

By the assumed uniform convergence statement, we have $\mathbb{P}(A_n(\varepsilon)) \to 1$ as $n \to \infty$.

We now argue that $A_n(\varepsilon) \subseteq \{\hat{\theta}_n \notin \Theta_\varepsilon\}$, which then implies the desired result.

Indeed, on the events $\{A_n(\varepsilon)\}$, we have

$$\sup_{\theta \in \Theta_\varepsilon} \bar{l}_n(\theta) \leq c(\varepsilon) + \delta(\varepsilon) < l(\theta_0) - \delta(\varepsilon) \leq \bar{l}_n(\theta_0)$$

Thus, the MLE cannot lie in Θ_ε, completing the proof. $\square$

Remark: the proof can be simplfied under additional properties of the likelihood function, such as differentiability and/or uniqueness of zeros. This can be useful in situations where Θ is not compact (see Example sheet).

Uniform Law of Large Numbers

In the proof of consistency of the MLE, we assumed

$$\sup_{\theta\in\Theta} |\bar{l}_n(\theta) - l(\theta)| \xrightarrow{P_{\theta_0}} 0$$

Theorem 1.17 (ULLN)**.** *Let Θ be a compact set in $\mathbb{R}^p$ and let $q : \mathcal{X} \times \Theta \to \mathbb{R}$ be continuous in θ for all x. Suppose $\mathbb{E}\left[\sup_{\theta\in\Theta} |q(X,\theta)|\right] < \infty$, where X is a random variable defined over $\mathcal{X}$. Suppose $X_1, \ldots, X_n$ are drawn iid according to the distribution of X. Then as $n \to \infty$*

$$\sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{i=1}^{n} q(X_i, \theta) - \mathbb{E}[q(X,\theta)] \right| \xrightarrow{a.s} 0$$

We now discuss the proof of the theorem (*Non-examinable*). The main idea is that since Θ is compact, it can be covered by a finite subcover up to a fixed precision (Heine-Borel Theorem).

Beginning of non-examinable section

**Proof*.* It is relatively easy to show that for a finite set $\{\theta_1, \ldots, \theta_M\} \subseteq \Theta$, we have

$$\max_{1\leq j\leq M} \left| \frac{1}{n} \sum_{i=1}^{n} q(X_i, \theta_j) - \mathbb{E}[q(X,\theta_j)] \right| \xrightarrow{\text{a.s}} 0 \tag{$*$}$$

Let $h_j(\cdot) = q(\cdot, \theta_j)$. Let A_j be the event that $\frac{1}{n}\sum_{i=1}^{n} h_j(X_i) - \mathbb{E}[h_j(X)] \to 0$. Then $\mathbb{P}(A_j) = 1$ by the Strong LLN. Letting $A = \bigcap_{i=1}^{M} A_j$, we have

$$\mathbb{P}(A^c) = \mathbb{P}\left(\bigcup_{j=1}^{M} A_j^c\right) \leq \sum_{j=1}^{M} \mathbb{P}(A_j^c) = 0$$

For a general class $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}$, we say that a family of *brackets* $\{[\underline{h}_j, \bar{h}_j]\}_{j=1}^{N}$ *covers* $\mathcal{H}$ if for all $h \in \mathcal{H}$, there exists j such that

$$\underline{h}_j(x) \leq h(x) \leq \bar{h}_j(x),\ \forall x \in \mathcal{X}$$

$\square$

Theorem 1.18. *Suppose $\mathcal{H}$ is a class of functions such that for all $\varepsilon > 0$, there exist finitely many brackets $\{[\underline{h}_j, \bar{h}_j]\}_{i=1}^{N(\varepsilon)}$ which cover $\mathcal{H}$, and such that for all $1 \leq j \leq N(\varepsilon)$*

1. $\mathbb{E}|\underline{h}_j(X)| < \infty$, $\mathbb{E}|\bar{h}_j(X)| < \infty$
2. $\mathbb{E}|\bar{h}_j(X) - \underline{h}_j(X)| < \varepsilon$

If $X_1, \ldots, X_n$ are iid copies of X, then

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}[h(X)] \right| \xrightarrow{a.s} 0$$

Proof. For a given $\varepsilon > 0$, consider the set of $N := N(\varepsilon/3)$ brackets guaranteed to exist by the hypothesis of the theorem. By the convergence result $(*)$, we know that almost surely we have

$$\max_{1 \le j \le N} \left| \frac{1}{n} \sum_{i=1}^{n} \bar{h}_j(X) - \mathbb{E}[\bar{h}_j(X)] \right| < \frac{\varepsilon}{3}$$

$$\max_{1 \le j \le N} \left| \frac{1}{n} \sum_{i=1}^{n} \underline{h}_j(X) - \mathbb{E}[\underline{h}_j(X)] \right| < \frac{\varepsilon}{3}$$

For $n \ge n_0(\varepsilon)$. Now take $h \in \mathcal{H}$ arbitrarily. From the above inequalities, we have (for some j)

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}[h(X)] &\le \frac{1}{n} \sum_{i=1}^{n} \bar{h}_j(X) - \mathbb{E}[\bar{h}_j(X)] + \big(\mathbb{E}[\bar{h}_j(X)] - \mathbb{E}[h(X)]\big) \\
&\le \frac{1}{n} \sum_{i=1}^{n} \bar{h}_j(X) - \mathbb{E}[\bar{h}_j(X)] + \big(\mathbb{E}[\bar{h}_j(X)] - \mathbb{E}[\underline{h}_j(X)]\big) \\
&< \frac{\varepsilon}{3} + \mathbb{E}\left|\bar{h}_j(X) - \underline{h}_j(X)\right| \\
&< \frac{2\varepsilon}{3}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}[h(X)] &\ge \frac{1}{n} \sum_{i=1}^{n} \underline{h}_j(X) - \mathbb{E}[\underline{h}_j(X)] + \big(\mathbb{E}[\underline{h}_j(X)] - \mathbb{E}[h(X)]\big) \\
&\ge -\frac{\varepsilon}{3} - \mathbb{E}\left[|\bar{h}_j(X) - \underline{h}_j(X)|\right] > -\frac{2\varepsilon}{3}
\end{aligned}$$

So almost surely,

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} h(x_i) - \mathbb{E}[h(X)] \right| < \frac{2\varepsilon}{3} < \varepsilon, \quad \text{for } n \ge n_0(\varepsilon)$$

Since ε was arbitrary, this implies the result. □

To move from the preceding theorem to the proof of the ULLN, we need to find an appropriate bracketing cover for the set of functions $\mathcal{H} = \{q(\cdot, \theta) : \theta \in \Theta\}$.

We define the open balls

$$B_\eta(\theta) = \{\theta' \in \Theta : ||\theta - \theta'||_2 < \eta\}$$

Now define the functions

$$u_\eta(x, \theta) = \sup_{\theta' \in B_\eta(\theta)} q(x, \theta')$$

$$l_\eta(x, \theta) = \inf_{\theta' \in B_\eta(\theta)} q(x, \theta')$$

By assumption, $\mathbb{E}[|u_\eta(X, \theta)|] < \infty$ and $\mathbb{E}[|l_\eta(X, \theta)|] < \infty$ for each θ, η. Furthermore, by continuity of $q(X, \cdot)$, together with the Dominated Convergence Theorem, we can choose a radius $\eta_\varepsilon(\theta)$ for each θ such that the (expected) width of the corresponding brackets is bounded by ε.

Then by compactness of Θ, we can define a finite set $\{\theta_1, \ldots, \theta_N\} \subseteq \Theta$ constituting a subcover of Θ. Applying the preceding theorem completes the proof.

End of non-examinable section

Asymptotic normality of the MLE

Assumptions: Let $\{f(\cdot,\theta) : \theta \in \Theta\}$ be a statistical model of pdfs/pmfs on $\mathcal{X} \subseteq \mathbb{R}^d$ such that, in addition to the assumptions stated for consistency of the MLE, we have

1. The true θ_0 belongs to $\text{int}(\Theta)$.
2. There exists an open set $U \subseteq \Theta$ containing θ_0 such that $\theta \mapsto f(x,\theta)$ is twice continuously differentiable with respect to $\theta \in U$, for each $x \in \mathcal{X}$.
3. The Fisher information matrix $I(\theta_0) \in \mathbb{R}^{p\times p}$ is non-singular, and
$$\mathbb{E}_{\theta_0}\left[||(\nabla_\theta \log f(X,\theta))|||_{\theta=\theta_0}\right] < \infty$$
4. There exists a compact ball $K \subseteq U$ with $\text{int}(K) \neq \emptyset$ centred at θ_0, such that
$$\mathbb{E}_{\theta_0}\left[\sup_{\theta\in K} ||\nabla^2_\theta \log f(X,\theta)||_2\right] < \infty$$
$$\int_{\mathcal{X}} \sup_{\theta\in K} ||\nabla_\theta \log f(X,\theta)||_2 \mathrm{d}x < \infty$$
$$\int_{\mathcal{X}} \sup_{\theta\in K} ||\nabla^2_\theta \log f(X,\theta)||_2 \mathrm{d}x < \infty$$

These assumptions are stated only for rigor, and are *non-examinable*.

Theorem 1.19. *Suppose the statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ satisfies the above regularity conditions, and let $\hat{\theta}_n$ be an MLE based on n iid observations $X_1, \ldots, X_n$ with distribution P_{θ_0}. As $n \to \infty$, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$.*

Proof. Idea: "Mean Value Theorem + Central Limit Theorem".

Define $\varepsilon > 0$ such that the ball of radius ε around θ_0 is contained in K. Let $E_n = \{||\hat{\theta}_n - \theta_0||_2 \leq \varepsilon\}$. Then $\mathbb{P}(E_n) \to 1$ since $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$.

We can focus on these events $\{E_n\}$ for the rest of the proof, since we are trying to show something about convergence of cdf's. On these events, the regularity assumptions imply $\nabla_\theta \bar{l}_n(\hat{\theta}_n) = 0$, by the first-order optimality condition. Applying the Mean Value Theorem coordinate-wise between θ_0 and $\hat{\theta}_n$, we have

$$0 = \nabla_\theta \bar{l}_n(\hat{\theta}_n) = \nabla_\theta \bar{l}_n(\theta_0) + \bar{A}_n\left(\hat{\theta}_n - \theta_0\right)$$

where $\bar{A}_n$ is defined coordinate-wise as

$$(\bar{A}_n)_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \bar{l}_n(\theta^{(i)}), \text{ for some } \theta^{(i)} \in [\theta_0, \hat{\theta}_n]$$

Rearranging gives

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \left(-\bar{A}_n^{-1}\right)\sqrt{n}\nabla_\theta \bar{l}_n(\theta_0)$$

Note that

$$\sqrt{n}\nabla_\theta \bar{l}_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\nabla_\theta \log f(X_i,\theta) - \underbrace{\mathbb{E}_{\theta_0}\left[\nabla_\theta \log f(X,\theta_0)\right]}_{=0}\right)$$

Thus, by the multivariate CLT

$$\sqrt{n}\nabla_\theta \bar{l}_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}_{\theta_0}(\nabla_\theta \log f(X,\theta_0))) = \mathcal{N}(0, I(\theta_0))$$

Now it suffices to show

$$\bar{A}_n \xrightarrow{P} \mathbb{E}_{\theta_0}[\nabla^2_\theta \bar{l}_n(\theta_0)] = -I(\theta_0)$$

Since then by the continuous mapping theorem $(\bar{A}_n)^{-1} \xrightarrow{P} -I(\theta_0)^{-1}$. Then by Slutsky's Lemma

$$\sqrt{n}\nabla_\theta \bar{l}_n(\theta_0) \xrightarrow{d} I(\theta_0)^{-1}(0, I(\theta_0)) = \mathcal{N}(0, I(\theta_0)^{-1})$$

The rest of this proof is *non-examinable*. It suffices to prove the convergence $\bar{A}_n \xrightarrow{P}= I(\theta_0)$ for each entry of $\bar{A}_n$.

For each entry, we write

$$(\bar{A}_n)_{jk} = \frac{1}{n}\sum_{i=1}^n\left(\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X_i,\theta^{(j)}) - \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X,\theta^{(j)})\right]\right)$$

$$+\mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X,\theta^{(j)})\right] - \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X,\theta_0)\right] + (-I(\theta_0))_{jk}$$

Denoting $q(X,\theta) = \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(x,\theta)$, the regularity assumptions imply continuity of $q(x,\theta)$ and $\mathbb{E}_{\theta_0}[q(x,\theta)]$ for all $x \in \mathcal{X}$. We can then conclude by the ULLN that

$$\frac{1}{n}\sum_{i=1}^n\left(\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X_i,\theta^{(j)}) - \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X,\theta^{(j)})\right]\right) \xrightarrow{\text{a.s}} 0$$

(and so also converges in probability). Now note

$$|\mathbb{E}_{\theta_0}[q(X,\theta^{(j)})] - \mathbb{E}_{\theta_0}[q(X,\theta_0)]| \xrightarrow{P} 0$$

using the fact that $\theta^{(j)} \xrightarrow{P} \theta_0$ (by consistency of $\hat{\theta}_n$ and the continuous mapping theorem). $\square$

By the theorem, we conclude that the MLE is both asymptotically normal and asymptotically efficient.

Definition. In a parametric model $\{f(\cdot,\theta) : \theta \in \Theta\}$, a consistent estimator $\tilde{\theta}_n$ is *asymptotically efficient* if $n\,\mathrm{Var}_\theta(\tilde{\theta}_n) \to I(\theta)^{-1}$ for all $\theta \in \mathrm{int}(\Theta)$ (if $p = 1$) or when $p > 1$ $n\,\mathrm{Cov}_\theta(\tilde{\theta}_n) \to I(\theta)^{-1}$ fo all $\theta \in \mathrm{int}(\Theta)$.

Remarks:

1. At the expense of more complicated proofs, can reduce the regularity conditions required for the function $\theta \mapsto f(x,\theta)$. In particular, this allows us to consider Laplace distributions, where the log-likelihood is not everywhere differentiable, since the pdf is proportional to $\exp(-|x-\theta|)$.

2. Some notion of regularity is required: the uniform distribution on $[0,\theta]$ with density $f(x,\theta) = \frac{1}{\theta}1_{[0,\theta]}$, the likelihood is discontinuous. The asymptotic theory breaks down in this case (see Example sheet).

3. For θ_0 at the boundary of Θ, asymptotics also might not be normal. In the case of the model $\mathcal{N}(\theta,1)$ for $\theta \in [0,\infty)$, when $\theta_0 = 0$ (see Example sheet).

4. Although it can be shown that the optimal asymptotic variance for a "regular" estimator is indeed $\frac{1}{n}I^{-1}(\theta_0)$, estimators outside this class could have smaller variances. For example, Hodge's estimator: let $\hat{\theta}_n$ be the MLE. Consider
$$\tilde{\theta}_n = \begin{cases} \hat{\theta}_n & |\hat{\theta}_n| > n^{-1/4} \\ 0 & \text{otherwise} \end{cases}$$
This is not a very sensible estimator and does not satisfy the regularity assumptions. However it is "superefficient", i.e has smaller asymptotic variance than the MLE under the above $\mathcal{N}(\theta,1)$ model for certain values in Θ (see Example sheet).

   However the Hodges' estimator does worse than the MLE in terms of another criterion, the <u>minimax risk</u>.

Plug-in MLE & Delta Method

Now consider the following estimation problem: for a parametric model $\{f(\cdot,\theta) : \theta \in \Theta\}$, we wish to estimate $\Phi(\theta)$, where $\Phi : \Theta \to \mathbb{R}^k$ and $\Theta \subseteq \mathbb{R}^p$. First consider a special case, and introduce the following definition:

Definition. For $\Theta = \Theta_1 \times \Theta_2$, and $\theta = \begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}$, define the *profile likelihood*, for $\Phi(\theta) = \theta_1$ by $L^{(p)}(\theta_1) = \sup_{\theta_2 \in \Theta_2} L(\theta_1, \theta_2)$.

Note that maximising $L^{(p)}$ is equivalent to maximising L and taking the first argument of the maximiser.

Example. $\mathcal{N}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. Recall (from the Example sheet) $\hat{\mu}_{\text{MLE}} = \bar{X}_n$ and $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

For general function Φ, can show that if Φ is injective, an MLE in the new parameterisation in ϕ, given by $\{f(\cdot, \phi) : \phi = \Phi(\theta), \theta \in \Theta\}$, is obtained by taking $\Phi(\hat{\theta}_{\text{MLE}})$ (see Example sheet). For the above example, could take $\Phi : (\mu, \sigma^2) \mapsto (\mu, \sigma)$.

In fact when Φ is not injective, can also define an MLE for ϕ and state a similar result: $\Phi(\hat{\theta}_{\text{MLE}})$ is always an MLE for the *induced likelihood function* $L^*(\phi) = \sup_{\theta : \Phi(\theta) = \phi} L(\theta)$ - e.g see the profile likelihood above.

Definition. For a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$ and a function $\Phi : \Theta \to \mathbb{R}^k$, the *plug-in MLE* of Φ is the estimator $\Phi(\hat{\theta}_{\text{MLE}})$.

Theorem 1.20 (Delta Method)**.** *Let $\Phi : \Theta \to \mathbb{R}$ be continuously differentiable at θ_0, with gradient satisfying $\nabla_\theta \Phi(\theta_0) \neq 0$. Let $\{\hat{\theta}_n\}$ be a sequence of estimators such that $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} Z$, where Z is a random vector in $\mathbb{R}^p$. Then*

$$\sqrt{n}\left(\Phi(\hat{\theta}_n) - \Phi(\theta_0)\right) \xrightarrow{d} \nabla_\theta \Phi(\theta_0)^T Z$$

Proof. By the MVT, for some $\tilde{\theta}_n$ in the line segment $[\theta_0, \hat{\theta}_n]$, we have

$$\sqrt{n}\left(\Phi(\hat{\theta}_n) - \Phi(\theta_0)\right) = \nabla_\theta \Phi(\tilde{\theta}_n)^T \sqrt{n}(\hat{\theta}_n - \theta_0)$$

Since $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} Z$, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathcal{O}_p(1)$. Thus, for any $\varepsilon, \delta > 0$ there exists $M(\delta)$ such that

$$\mathbb{P}\left(\sqrt{n}\|\hat{\theta}_n - \theta_0\|_2 > M(\delta)\right) < \delta, \ \forall n \in \mathbb{N}$$

So if we choose n large enough such that $\frac{M(\delta)}{\sqrt{n}} < \varepsilon$, we have

$$\mathbb{P}\left(\|\tilde{\theta}_n - \theta_0\|_2 > \varepsilon\right) \leq \mathbb{P}\left(\|\hat{\theta}_n - \theta_0\|_2 > \varepsilon\right) \leq \mathbb{P}\left(\|\hat{\theta}_n - \theta_0\|_2 > \frac{M(\delta)}{\sqrt{n}}\right)$$

Implying that $\tilde{\theta}_n \xrightarrow{P} \theta_0$. By the Continuous Mapping Theorem, we have $\nabla_\theta \Phi(\tilde{\theta}_n) \xrightarrow{P} \nabla_\theta \Phi(\theta_0)$. The result then follows by Slutsky's Lemma. $\square$

Remarks:

1. The Delta Method can be generalised to other estimators, taking a sequence $r_n \to \infty$ instead of $\sqrt{n}$ (obvious from the proof).

2. In the case of the MLE, combining with the theorem about asymptotic normality, we have

$$\sqrt{n}\left(\Phi(\hat{\theta}_n) - \Phi(\theta_0)\right) \xrightarrow{d} \mathcal{N}(0, \nabla_\theta \Phi(\theta_0)^T I^{-1}(\theta_0) \nabla_\theta \Phi(\theta_0))$$

   So when θ is one-dimensional

$$\sqrt{n}\left(\Phi(\hat{\theta}_n) - \Phi(\theta_0)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\Phi'(\theta_0)^2}{I(\theta_0)}\right)$$

3. The previous calculation also shows that the plug-in MLE is asymptotically efficient. Recall the "multivariate" Cramer-Rao lower bound from before:

$$\mathrm{Var}_{\theta_0}\left(\Phi(\hat{\theta}_n)\right) \geq \frac{1}{n} \nabla_\theta \Phi(\theta_0)^T I^{-1}(\theta_0) \nabla_\theta \Phi(\theta_0)$$

4. We don't really use the fact that $\nabla_\theta \Phi(\theta_0) \neq 0$ in the proof. But in that case, we have convergence in distribution to 0, so this is not very informative (so should rescale by something bigger than $\sqrt{n}$)

5. The proof of the Delta Method extends to multivariate functions in a straightforward manner.

Example. Consider the model $\mathcal{N}(\theta, 1)$, $\Theta = \mathbb{R}$. Then $\hat{\theta}_{\mathrm{MLE}} = \bar{X}_n$. SO $\sqrt{n}(\hat{\theta}_{\mathrm{MLE}} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$. Suppose $\Phi(\theta) = \theta^2$. Then the Delta Method implies $\sqrt{n}(\hat{\theta}_{\mathrm{MLE}} - \theta_0^2) \xrightarrow{d} 2\theta_0\mathcal{N}(0,1) = \mathcal{N}(0, 4\theta_0^2)$. Consider the distribution of $n(\bar{X}_n^2 - 0)$. Then $n\bar{X}_n^2 = \left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right)^2 \xrightarrow{d} \chi_1^2$, by the CLT and continuous mapping theorem.

Asymptotic Inference

For the MLE $\hat{\theta}_n$, under regularity assumptions,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$$

Let e_j denote the jth canonical basis vector. Then

$$\sqrt{n}(\hat{\theta}_j - (\theta_0)_j) = e_j^T \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, e_j^T I(\theta_0)^{-1} e_j) = \mathcal{N}(0, (I(\theta_0)^{-1})_{jj})$$

Let

$$\mathcal{C}_n = \left\{ \nu \in \mathbb{R} : |\nu - \hat{\theta}_{n,j}| \leq \frac{(I(\theta_0)^{-1})_{jj}^{1/2}}{\sqrt{n}} z_\alpha \right\}$$

Where z_α is such that $\mathbb{P}(Z \leq z_\alpha) = 1 - \alpha$. This is a valid asymptotic confidence interval since

$$\mathbb{P}_{\theta_0}(\theta_{0,j} \in \mathcal{C}_n) = \mathbb{P}_{\theta_0}\left(\sqrt{n}(I(\theta_0)^{-1})_{jj}^{-1/2} |\hat{\theta}_{n,j} - \theta_{0,j}| \leq z_\alpha\right) \to \mathbb{P}(Z \leq z_\alpha) = 1 - \alpha$$

In order to construct this confidence interval, we need to evaluate $I(\theta)$ at θ_0. But we don't know θ_0! Instead we will estimate the required quantity by plugging in $\hat{\theta}_{\mathrm{MLE}}$.

Definition. The *observed Fisher information* is the $p \times p$ matrix

$$i_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left(\nabla_\theta \log f(X_i, \theta)\right) \left(\nabla_\theta \log f(X_i, \theta)\right)^T$$

It is common to use $\hat{i}_n = i_n(\hat{\theta}_{\mathrm{MLE}})$ as an estimator of $I(\theta_0)$.

Theorem 1.21. *Under the usual regularity conditions, we have* $\hat{i}_n \xrightarrow{P_{\theta_0}} I(\theta_0)$ *as* $n \to \infty$. *In particular, a confidence interval based on* $\hat{i}_n$ *will be asymptotically valid.*

Proof. Let $q(X, \theta) = \left(\nabla_\theta \log f(X, \theta)\right) \left(\nabla_\theta \log f(X, \theta)\right)^T$. For all $\theta \in \Theta$ we have

$$i_n(\theta) = \frac{1}{n} \sum_{i=1}^n q(X_i, \theta), \ I(\theta) = \mathbb{E}_{\theta_0}[q(X, \theta)]$$

Thus

$$\hat{i}_n - I(\theta_0) = \left(i_n(\hat{\theta}_{\mathrm{MLE}}) - I(\hat{\theta}_{\mathrm{MLE}})\right) + \left(I(\hat{\theta}_{\mathrm{MLE}}) - I(\theta_0)\right)$$

The first term is upper bounded by

$$\left|i_n(\hat{\theta}_{\text{MLE}}) - I(\hat{\theta}_{\text{MLE}})\right| \leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n q(X_i,\theta) - \mathbb{E}_{\theta_0}\left[q(X_i,\theta)\right]\right| \xrightarrow{P_{\theta_0}} 0$$

by the ULLN. The second term also converges in probability to 0, by consistency of the MLE combined with the continuous mapping theorem. ☐

Remark: it is also possible to use $\hat{j}_n(\theta) = j_n(\hat{\theta}_{\text{MLE}})$, where

$$j_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 \log f(X_i,\theta)$$

This is also a consistent estimator of $I(\theta_0)$ with a similar proof.

Definition. For all $\theta \in \Theta$, we define the *Wald statistic* as

$$W_n(\theta) = n\left(\hat{\theta}_{\text{MLE}} - \theta\right)^T \hat{i}_n \left(\hat{\theta}_{\text{MLE}} - \theta\right)$$

The Wald statistic is a quadratic form with positive semi-definite $\hat{i}_n$, so its level sets are ellipsoids that can be used to construct confidence sets for θ_0.

Theorem 1.22. *Let θ_0 be p-dimensional. Under the usual regularity conditions, the confidence region*

$$\mathcal{C}_n = \{\theta : W_n(\theta) \leq \xi_\alpha\}$$

where ξ_α is such that

$$\mathbb{P}(\chi_p^2 \leq \xi_\alpha) \leq 1-\alpha$$

is an asymptotically valid confidence region for θ_0.

Proof. Compute

$$\mathbb{P}(\theta_0 \in \mathcal{C}_n) = \mathbb{P}_{\theta_0}\left(W_n(\theta) \leq \xi_\alpha\right)$$

Under the assumptions, we have $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$, and $\hat{i}_n \xrightarrow{P_{\theta_0}} I(\theta_0)$. We can decompose the Wald statistic as

$$W_n(\theta_0) = \sqrt{n}(\hat{\theta}_n - \theta_0)^T I(\theta_0)\sqrt{n}(\hat{\theta}_n-\theta_0) + \sqrt{n}(\hat{\theta}_n-\theta_0)^T\left(\hat{i}_n - I(\theta_0)\right)\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$$

By the continuous mapping theorem, the first term converges in distribution to $U^TU = U_1^2 + \ldots + U_p^2$, with $U \sim \mathcal{N}(0, I_p)$, hence has a χ_p^2 distribution.

The second term is a product of $\sqrt{n}(\hat{\theta}_n - \theta_0)^T(\hat{i}_n - I(\theta_0))$ which converges in distribution (hence also in probability) to 0 by Slutsky's lemma, and applying Slutsky's lemma again the whole of the second term goes to 0. So $W_n(\theta_0) \xrightarrow{d} \chi_p^2$. ☐

Remark: the Wald statistic can also be used to design a test for

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \in \Theta \setminus \{\theta_0\}$$

where H_0 is rejected when $W_n(\theta_0) > \xi_\alpha$, since $\mathbb{P}_{\theta_0}(W_n(\theta_0) > \xi_\alpha) \to \alpha$.

Likelihood Ratio Test

Now consider a general "nested" hypthesis testing problem:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta \setminus \Theta_0$$

Where $\Theta_0 \subseteq \Theta \subseteq \mathbb{R}^p$. We want to find a decision rule Ψ_n which is a function of the observed data, mapping into $\{0, 1\}$, which outputs 0 with high probability under H_0 and 1 with high probability under H_1.

Definition. We have the following types of error:

- *Type I error*: (false positive) $\mathbb{P}_\theta(\Psi_n = 1) = \mathbb{E}_\theta[\Psi_n]$, for $\theta \in \Theta_0$
- *Type II error*: (false negative) $\mathbb{P}_\theta(\Psi_n = 0) = \mathbb{E}[1 - \Psi_n]$, for $\theta \in \Theta \setminus \Theta_0$.

Definition. The *likelihood ratio statistic* is defined as

$$\Lambda_n(\Theta, \Theta_0) = 2 \log \left(\frac{\sup_{\theta \in \Theta} \prod_{i=1}^n f(X_i, \theta)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n f(X_i, \theta)} \right) = 2 \log \left(\frac{\prod_{i=1}^n f(X_i, \hat{\theta}_{\text{MLE}})}{\prod_{i=1}^n f(X_i, \hat{\theta}_{\text{MLE},0})} \right)$$

Where $\hat{\theta}_{\text{MLE},0}$ is the MLE over the subset Θ_0.

Note that $\Lambda_n(\Theta, \Theta_n) \geq 0$, and we should reject H_0 when Λ_n is large.

Theorem 1.23 (Wilks' Theorem)**.** *Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a statistical model satisfying the usual regularity conditions, where $\Theta \subseteq \mathbb{R}^p$, and consider a hypothesis testing problem where $\Theta_0 = \{\theta_0\}$, for some fixed $\theta_0 \in \text{int}(\Theta)$. As $n \to \infty$ we have*

$$\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi_p^2$$

Proof. Let $\varepsilon > 0$ be such that the ball of radius ε around θ_0 is contained in Θ. Define $E_n = \{\|\hat{\theta}_n - \theta_0\|_2 \leq \varepsilon\}$, where $\hat{\theta}_n$ is the MLE. Then $\mathbb{P}(E_n) \to 1$ since $\hat{\theta}_n \xrightarrow{P} \theta_0$. Thus, on the events $\{E_n\}$ we have $\hat{\theta}_n \in \text{int}(\Theta)$. It suffices to restrict our attention to these events since we are talking about convergence in distribution.

By definition of the likelihood ratio

$$\Lambda_n(\Theta, \Theta_0) = 2l_n(\hat{\theta}_n) - 2l_n(\theta_0) = 2\nabla_\theta l_n(\hat{\theta}_n)^T(\hat{\theta}_n - \theta_0) - (\hat{\theta}_n - \theta_0)^T \bar{B}_n(\hat{\theta}_n - \theta_0)$$

Where $\bar{B}_n$ is defined using a Taylor approximation with remainder:

$$\bar{B}_n = -\nabla_\theta^2 l_n(\bar{\theta}_n), \text{ where } \bar{\theta}_n \in [\theta_0, \hat{\theta}_n]$$

The first term in the Taylor approximation is 0 since $\hat{\theta}_n$ is an MLE. Note that the second term can be written as

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0\right)^T j_n(\bar{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0)$$

where j_n is as defined previously. Using the same proof technique as before, we can show (via a ULLN) that $j_n(\bar{\theta}n) \xrightarrow{P_{\theta_0}} I(\theta_0)$:

$$j_n(\bar{\theta}_n) - I(\theta_0) = \left(\underbrace{j_n(\bar{\theta}_n) - I(\bar{\theta}_n)}_{\text{ULLN}}\right) + \left(\underbrace{I(\bar{\theta}_n) - I(\theta_0)}_{\text{CMT}}\right)$$

Applying Slutsky's lemma implies $\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi^2_p$. $\square$

Remarks:

1. As a result, $\Psi_n = 1\{\Lambda_n(\Theta, \Theta_0) \geq \xi_\alpha\}$ is a valid hypothesis test at level α.
2. Wilks' Theorem can be generalised to certain settings with composite null hyptheses, in which case the limiting distribution is a $\chi^2 p - p_0$ random variable where $p_0 \leq p$ is the "degrees of freedom" in Θ_0. For example, Θ_0 might be the hypothesis that fixes k values of the coordinates of θ, in which case $p_0 = p - k$.

Bayesian inference

For a given parametric model $\{f(\cdot, \theta) : \theta \in \Theta\}$, there are situations where it is convenient to consider θ as a random variable with distribution π on Θ.

For example, consider a finite parameter space $\Theta = \{\theta_1, \ldots, \theta_k\}$, and possible hypotheses $H_i : \theta = \theta_i$ for $1 \leq i \leq k$, with prior beliefs $\pi_i = \mathbb{P}(H_i)$. If the true hypothesis is H_i, then the distribution of an observation is

$$\mathbb{P}(X = x | H_i) = f_i(x)$$

By Bayes' rule, when observing $X = x$ we have

$$\mathbb{P}(H_i | X = x) = \frac{\mathbb{P}(X = x \text{ and } H_i)}{\mathbb{P}(X = x)} = \frac{\pi_i f_i(x)}{\sum_{j=1}^k \pi_j f_j(x)}$$

Thus, we "should" prefer H_i over H_j, given the observation $X = x$ if

$$\frac{\mathbb{P}(H_i | x = X)}{\mathbb{P}(H_j | X = x)} = \frac{f_i(x)}{f_j(x)} \frac{\pi_i}{\pi_j} \geq 1$$

If all the π_i's were equal, this would be a likelihood ratio test based on $\frac{f_i(x)}{f_j(x)}$. In the more general case, we have a weighted ratio and may want to update the π_i's.

Definition. For a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$, we will say that the *law of X given θ* is given by $X|\theta \sim f(x, \theta)$. The *posterior distribution* is defined as the law of $\theta|X$.