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 $X_1, \dots, X_n$.

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

(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,

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

$$= \int_{\mathcal{X}} \tilde{\theta}(X_1,\ldots,X_n)\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}f(X_1,\ldots,X_n,\theta)}{f(X_1,\ldots,X_n,\theta)}f(X_1,\ldots,X_n)\mathrm{d}x_1\ldots\mathrm{d}x_n$$

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

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 $(X_1,\ldots,X_n)$ 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 $n = 1$.

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

<u>Case 2</u>: 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 not 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

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

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

$$\mathbb{P}(\omega \in \Omega : ||X_n(\omega) - X(\omega)||_2 \to 0 \text{ as } n \to \infty)$$
$$= \mathbb{P}(||X_n - X||_2 \to 0 \text{ as } n \to \infty) = 1$$

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. □

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. □

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. □

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 $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ 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 $\mathrm{Var}(X) < \infty$. □

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 $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ be iid copies of $X \sim P$ on $\mathbb{R}^k$, such that* $\mathrm{Cov}(X) = \Sigma$ *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 $X_1, \ldots, X_n$ from the parametric model $\{P_\theta : \theta \in \Theta\}$. An estimator $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, X_n)$ 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 $|f_n(x)| \leq g(x)$ 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 $\Theta_\varepsilon = \{\theta \in \Theta : ||\theta - \theta_0||_2 \geq \varepsilon\}$, 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

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

Consider the sequence of events

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

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).